THESIS

SUPER-RESOLUTION GENERATIVE ADVERSARIAL NETWORK FOR WEATHER RADAR APPLICATIONS

Submitted by

Jacob T. Lesher-Garcia

Department of Electrical and Computer Engineering

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2023

Master's Committee:

Advisor: V. Chandrasekar

Margaret Cheney
Ketul Popat

ABSTRACT

SUPER-RESOLUTION GENERATIVE ADVERSARIAL NETWORK FOR WEATHER RADAR APPLICATIONS

Weather radars are vital to ensuring the safety of society by providing timely, accurate products used to forecast the development of weather phenomena. To this end, high spatiotemporal resolution data is paramount. Collecting high-resolution polarimetric observation data necessitates scan strategies with a slow scan rate. This thesis proposes the use of an established deep learning model in order to augment the current weather radar operational paradigm. Specifically, this thesis focuses on evaluating the efficacy of the super-resolution generative adversarial network (SRGAN) in generating physically realistic, pseudo-high-resolution radar scans – referred to as super-resolution (SR) scans – from low-resolution (LR) weather radar scans. With this, weather radar systems would be able to collect LR scans at faster scan rates while maintaining the quality of high-resolution (HR) scans by using the SRGAN to generate SR scans. This thesis aims to assess the generating capabilities of the SRGAN within the scope of generating SR scans from a pseudo-LR scan, processed from an actual HR scan. In order to accomplish this task, multiple experiments are setup, designed to test the SRGAN's capabilities in conducting SR for different architectural configurations, scan types, resolution scaling factors and downsampling methods, one of which simulates the characteristics of actual LR weather radar scans. The experimental SRGANs' performances are assessed both quantitatively and qualitatively, comparing between the SR scans and the baseline interpolation methods. The results of this thesis have found that the SRGAN model can outperform the baseline methods, specifically for the higher resolution scale factors and especially for the RHI radar scan type. Furthermore, the SRGAN is able to generate a physically representative SR image that reflects the natural features of a HR image. This is significant as it suggests that the SRGAN model is more effective when applied to practical applications.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Chandra for all of the opportunities and guidance he has given me. I would like to thank my committee, Dr. Cheney and Dr. Popat for the opportunity to defend my thesis and to present my research work. I would like to thank all of my colleagues in the CSU Radar and Communications Laboratory for their help along the way. I would also like to thank the National Science Foundation for supporting this research.

DEDICATION

To my awesome wife and to everyone in our wonderful family. For all of your support, encouragement and advice during my Master Quest, thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Weather radars are crucial in assuring emergency preparedness for severe storms and natural disasters, helping to minimize property damage and save lives. Data collected by weather radar systems contains a wealth of information that helps to characterize the volumetric targets of interest and predict their behavior. With weather radar data, critical information such as the weather event's starting time, place of origin, precipitation type, movement, developmental progression and more can be predicted and tracked throughout the storm's life cycle. These insights allow the populace and emergency response agencies to be better prepared in dealing with cataclysmic weather events such as natural disasters as well as their side effects such as fallen debris, floods, land slides and wild fires. This information is also useful in other applications such as assisting in decision making for aeronautical and agricultural industries as well as recreational activities. Altogether, weather radar systems help to secure the safety of individuals and society at large while enhancing the general quality of life.

Weather radars also play a pivotal role in progressing various fields of study such as atmospheric science, meteorology, remote sensing, geoscience and weather radar engineering. By providing such useful information, weather radars enable researchers and scientists to better observe and more thoroughly understand the world and its inner workings. This promotes the development of more accurate science models and more physically realistic simulations that are used to forecast weather event patterns and understand their behavior. Ultimately, weather radar data is an invaluable asset that advances the scientific study of atmospheric processes in order to help inform daily routine functions of people and industries while protecting society as a whole.

Because of the utility and applicability of weather radar data, the demand for high-quality weather radar data is ever-increasing. Typically, the quality of weather radar data often refers to its spatiotemporal resolution. The resolution of weather radar data is defined in both space and time. The spatial resolution of a weather radar system is defined by the gate width, the number of range

gates, and the beamwidth of the antenna that are used to represent an area of observation. Higher spatial resolution indicates that there are an increased number of data points that are representing the area, resulting in more information being observed and collected at a time in order to produce a high-quality weather radar scan. Temporal resolution is a measurement of the frequency of data collection and, from an operational standpoint, the reliability that the data will be collected and accessible at regular intervals. Higher temporal resolution indicates that the data is being collected at a faster rate, resulting in more information being observed and collected in a shorter time-scale. Increased data collection rates allow for rapidly developing storm systems, e.g., tornadoes and hurricanes, to be observed as their rapidly-changing characteristics can be captured with faster weather radar scans.

1.1 Problem Statement

The demand for high-quality weather radar data is only increasing, especially with the frequency, intensity, and affects of severe weather events and natural disasters being exacerbated due to climate change. While scan strategies and post-processing techniques can help, the most ubiquitous techniques implemented for collecting high-quality weather radar data begin in the radar system design itself. Because of this, working to improve the quality of data collection after the radar's installation is a difficult task. In addition, high spatial resolution weather radar data requires slower scan rates. This is due to the increased number of range gates being collected within a single scan. This can result in a single high-resolution (HR) weather radar scan taking minutes to be collected. Faster scan rates increase the data collection frequency; however, this results in lower spatial resolution weather radar scans being collected. Thus, the spatial and temporal resolutions are in opposition with one another. In order to provide a solution to this inherent issue, this thesis proposes the utilization of deep learning (DL) super-resolution (SR) models for super-resolving low-resolution (LR) weather radar scans.

Specifically, this thesis focuses on the effectiveness of the super-resolution generative adversarial network (SRGAN) model for this research. There are many advantages to this proposition.

Firstly, the weather radar systems and networks that are already installed would not have to be redesigned in order to improve their data quality. Utilizing a SRGAN model in software during the processing of the weather radar data would allow for existing weather radar systems to quickly implement this change and benefit from its higher quality data outputs without committing significant alterations to the system itself. Secondly, increasing the accessibility of HR weather radar data by applying SRGAN models to existing networks promotes further development of science models and physics simulations. Researchers and scientists would be able to use high-quality data all throughout the developmental process resulting in better informed, more accurate models, simulations and equations. Lastly, having a SRGAN model embedded within the operational function of a weather radar system would allow for faster scan rates to be used to collect the data, as done in LR weather radar scans, which would increase the temporal resolution of the data. Radar systems would be able to collect more data in a shorter amount of time for every weather event being observed. This is especially important when observing and collecting data for particularly turbulent, rapidly developing weather events such as tornadoes and hurricanes. Meanwhile, the spatial resolution of the data would also be maintained. The SRGAN model is used to push the LR weather radar scan collected into the HR regime. This is done by training the SRGAN model on HR and LR weather radar scan input pairs. Through this training process, the SRGAN model learns a mapping between the resolution regimes, allowing it to generate outputs of the same resolution as the HR input. These are referred to as SR outputs. This thesis will evaluate the efficacy of using the SRGAN model in super-resolving LR weather radar scans. If found to be effective, the proposal could substantially enhance the weather radar functional paradigm. Weather radar systems would be able to collect more data at a faster rate with high spatiotemporal resolution data quality when utilizing the SRGAN model to conduct SR on LR weather radar scans.

1.2 Research Objectives

This thesis aims to develop and assess the efficacy of utilizing the SRGAN model within the context of super-resolving LR weather radar scans. In order to do so, numerous experiments are

conducted and evaluated against baseline techniques considered to be standard in the literature. Architectural parameter experiments are performed in which different parameter values are tested that affect the model configuration. The goal of these experiments is to analyze the performance of different architectural configurations when used to train the SRGAN model and propose a set of optimal architectural parameters based on their evaluation results and visual perceptibility. Additionally, three dataset-type experiments are used to evaluate the performance of the SRGAN model in super-resolving different radar scan types, at different resolution scale factors and for different LR downsampling methods. One of the LR downsampling methods simulates the characteristics of an actual LR weather radar scan to determine how effective the SRGAN model can be when applied to real-world applications, a primary goal of this thesis' research. This thesis also aims to lay the groundwork and encourage further exploration in the field of applying machine learning (ML) networks in order to conduct SR on weather radar scans with the overall goal of enhancing radar systems' abilities to scan faster and collect more data while maintaining the data quality of HR scans.

1.3 Thesis Overview

Chapter 2 reviews the foundational aspects of the concepts behind SR techniques. For this, fundamental SR theory is explained and prominent SR works prevalent in the literature are presented that will serve as references for developing this thesis' research experimentation. The specifications of the SRGAN model – that is the primary focus of study in this thesis – are detailed including the driving theory and architectural makeup. The context, within which this thesis study's scope is defined, is examined by presenting other works in literature that investigate the use of DL models for conducting SR on weather radar scans.

Chapter 3 describes the principles of weather radar observation and data collection. Different weather radar scan strategies are described, including how the information that they provide differs and why they are evident choices for being research objectives studied throughout this thesis. Relevant weather radar products are expressed in terms of their algorithmic development from

data collect to product output. These serve as the primary training, validation and testing inputs to the ML model as it is developed. Current methods for employing SR into weather radar operational systems is explored and the significance of improving upon the current paradigm by using DL SR techniques is articulated.

Chapter 4 discusses the weather radar system used and the motive behind examining and utilizing one of its specific field campaign datasets throughout this thesis. The processing techniques used prior to training that generate the LR datasets and prepare the data for use in developing the models are thoroughly explained. SR datasets used throughout the literature and the composition and formulation of the dataset used to develop the SRGAN model in this thesis is presented. Contained therein are descriptions of the subsets utilized and the roles they play at different points in the SRGAN model's development.

Chapter 5 details the experiments conducted and provides the reasoning behind the experimental setup and procedure. Specifications of the coding software and processing environment are discussed. The methods employed as the baseline comparisons, against which the experimental models are assessed, are presented. The optimization process that fine tunes the SRGAN model's hyperparameters as well as its function and importance in the experimental models' development is explained. The sets of experimental variables that serve as the focal points of investigation for the experiments performed are described. The methodology used to train the experimental models is discussed. Metrics utilized to quantitatively evaluate the performance of the experimental models are defined.

Chapter 6 presents the evaluation results for each experimental model tested. Quantitative analyses compare the performances between the different experiments conducted. Example output products from the experimental models are illustrated and qualitative assessments are given that discuss their visual quality. A comprehensive survey of the experimental models' results examines cross-experimental behavioral patterns. The experimental models with the highest ranking performances – based on both quantitative and qualitative results – are compared against the baseline methods. Strengths of the experimental models' abilities are emphasized within the context of

real-world applications. These supply evidence and rationale for the conclusions drawn at the end of this thesis.

Chapter 7 recapitulates the study and research conducted within this thesis. Conclusions are drawn from the experimental results and analyses. These are used to determine the extent to which the thesis met the research objectives. Finally, comments on how the study could be further improved and expanded upon are provided to encourage future research and exploration in this field of study.

Chapter 2

Background on Super-Resolution

SR has been a growing field of research with new techniques being developed for many different applications. SR refers to the process in which one or more LR images are used in order to generate one or more HR images. Achieving a higher resolution is a constantly pursued goal whether working with images or data. Having HR data/images allows for sharper, more distinct samples to be collected and finer details to be observed and analyzed. Originally, the main way to achieve HR observations was through hardware-based solutions. These primarily comprised of either effectively increasing the amount of pixels per unit area by decreasing the pixel size or increasing the image sensor size [1, 2]. These solutions are nearing their limits due to the physical restrictions of obtaining an optimal pixel size as well as the high cost of manufacturing optics and image sensor technology with high precision quality. However, ever since the emergence of image processing techniques, software-based solutions have proved to be more cost-effective and accessible, as existing LR images and imaging systems could still be utilized, while producing noteworthy results [1, 2].

Many early image processing techniques focused on problems related to SR with two of the most prominent being image restoration and image interpolation. Image restoration algorithms focused on enhancing the perceptibility of images by mitigating any blur or noise affecting the image. Nevertheless, this did not increase the size of the image. Image interpolation focused on increasing the size of an image although the high-frequency components of the interpolated images were often degraded through the process [1]. Both of these can be considered as progenitors to SR techniques. The concept of SR has been researched thoroughly due to its ability to surpass the imaging systems' inherent limitations in resolution quality while improving upon other image processing methods. For these reasons, many different SR processes have been developed. Some of the prominent SR algorithms that were developed early on include the frequency domain approach [3], iterative back propagation [4], interpolation of non-uniformly spaced samples [5],

reconstruction-based techniques [6–8], and adaptive filtering [9]. However, developments of SR techniques as a solution to computer vision problems within the field of ML, especially when utilizing DL models, have enabled researchers, scientists, and engineers to further overcome these limitations and continue the endeavor for reaching higher resolution regimes. This thesis focuses on a specific, recently-developed DL SR model called a SRGAN.

Chapter 2 aims to thoroughly review the background into the realm of SR. First, Chapter 2.1 gives insight into the theory behind SR techniques. Chapter 2.2 then presents the prominent DL SR works in literature in which different DL SR models are developed and explained. A summary of the literature is presented in Table 2.1. Chapter 2.3 specifically presents details behind the SRGAN model, the primary focus for this thesis study, including the theory and the model architecture. Finally, Chapter 2.4 presents various SR implementation within the weather radar field of study with a focus on the use of the SRGAN model for super-resolving weather radar scan images.

2.1 Super-Resolution Theory

The fundamental concept of SR theory is to generate a pseudo-HR image, also referred to as a SR image, from an input LR image. To begin, however, initial complications that arise regarding the concept of using a LR image and, from it, generating a pseudo-HR image must be discussed and resolved. Information theory presents the data processing inequality (DPI) which initially appears to oppose SR theory. The DPI states that if three random variables $X \rightarrow Y \rightarrow Z$ form a Markov chain in this order, then the mutual information between them is satisfied as in Equation 2.1:

$$I(X;Z) \leq I(X;Y) \tag{2.1}$$

This implies that the conditional distribution of the random variable Z depends solely on the random variable Y while being conditionally independent of the random variable X. Effectively, the DPI dictates that, no matter how data is processed, information cannot be added that is not

already present. Another way to interpret this is that missing data cannot be recovered. So, in regards to SR theory, the DPI suggests that a single LR image cannot alone be processed to recover the high-frequency components within a corresponding HR image as the LR image is missing these components originally. Early SR algorithms resolved this issue by utilizing multiple LR images of the same scene from different angles and times or contextually similar scenes [3, 7–9]. This way, information from the other LR samples could be utilized to reconstruct details that are missing from an individual LR image and generate a SR image. Similarly, DL SR techniques train their models on large datasets containing a multitude of images. Most of these training datasets consist of contextually similar images but some DL SR techniques utilize a wide variety of non-similar images in order to develop a general SR model. This enables the DL SR models to conduct single image super-resolution (SISR), i.e., generating a pseudo-HR image from an individual LR image.

In addition, the concept behind SR theory is considered to be an ill-posed problem. There are many possible HR images that could be derived from a single LR image input, since the high-frequency components must be estimated from a LR sample, which means that the SR image solution does not satisfy the uniqueness requirement [1, 10]. Many conventional SR techniques start by defining the LR image in terms of the HR image in order to better characterize the SR problem. This is formulated in Equation 2.2 as follows:

$$I_{LR} = (I_{HR} * K)_{\downarrow s} + N \tag{2.2}$$

where the HR image is denoted as $I_{HR} = \{i_{HR} : i_{HR} \in \mathbb{R}\}^{h \times w \times c}$. The HR image matrix's size is defined by the height h, width w and channel c variables. Each pixel i_{HR} within the HR image matrix is a real number. The HR image is spatially convolved with a 2D downsampling kernel $K = \{k : k \in \mathbb{R}\}^{k_h \times k_w}$ where s is the resolution scaling factor that controls the amount of downsampling performed on I_{HR}. The size of the kernel is defined by the kernel height k_h and kernel width k_w. After the spatial convolution, additive noise $N = \{n : n \in \mathbb{R}\}^{h \times w \times c}$ is applied pixel-wise to the resulting downsampled HR image. These operations result in the downsized and degraded LR image which is denoted as $I_{LR} = \{i_{LR} : i_{LR} \in \mathbb{R}\}^{\frac{h}{s} \times \frac{w}{s} \times c}$. The LR

image matrix's size is defined as the resulting quotient of the HR image matrix's height and width and the resolution scaling factor. Throughout this process, the sizes $h \times w$, $k_h \times k_w$ and $\frac{h}{s} \times \frac{w}{s}$ define spatial resolutions. Meanwhile, c defines the number of color channels within the channel dimension where $c = 1$ for grayscale images and $c = 3$ for red, green and blue (RGB) images.

Equation 2.2 suggests that a single LR image can be mapped from multiple different HR images. With this definition of the LR image set, most of the foundational optimization-based SR algorithms will then formulate the generation of the SR image as a function of the LR image and a given input image to be processed [10]. Example-based SR algorithms were then developed that utilized training image pairs in the processing for super-resolution. The term "training image pairs" refers to a LR image and its corresponding HR image related to each other as shown in Equation 2.2. Patches of the SR image were then defined by dictionary elements of weights applied the the HR image. The weights were estimated based upon a dictionary made up of patches of the corresponding LR image. The emergence of learning-based SR algorithms built upon both these concepts. The learning-based SR methods utilized the concept of training image pairs from the example-based SR techniques in conjunction with the minimization function from the optimization-based SR algorithms. In order to do so, the learning-based SR models trained a SR model to minimize the objective loss function directly over the training image pairs [10]. This objective function is defined in Equation 2.3 as follows:

$$L(\theta) = \mathbb{E}\left[\|f(I_{LR};\theta) - I_{HR}\|_2^2\right] + R(\theta) \tag{2.3}$$

where $f(I_{LR};\theta) = I_{SR}$ denotes the resulting super-resolved image, $R(\theta)$ is a regularization term, $\mathbb{E}$ is the error and $L(\theta)$ is the resulting loss between the SR image and the HR image [10]. This equation formulates the foundation for most learning-based SR algorithms, including many DL SR models, which will be discussed further in Chapter 2.2, as well as the SRGAN ML SR model, which will be discussed further in Chapter 2.3, which is the primary focus for this thesis.

2.2 Deep Learning Super-Resolution Literature

DL describes a ML algorithm that is made up of complex, artificial neural networks whose architecture was originally inspired by how the brain operates. DL models are renowned for their ability to be trained on large datasets of unstructured data while learning the features automatically. When DL models were first being investigated for use in SR tasks, Dong et al. is accredited in literature with pioneering this field of study when proposing the Super-Resolution Convolutional Neural Network (SRCNN) in 2014 [11] and again in 2016 [12] which improved upon the first by exploring deeper structures, processing multi-channel color images, and furthering the results analysis. These papers presented the use of the convolutional neural network (CNN) model architecture for conducting SISR by learning a mapping from the LR to the HR regime. [13] presented a very deep CNN architecture as a SR model – called the Very Deep Super-Resolution (VDSR) model – that handles SR tasks at multiple resolution scales while using a single model. Kim et al. found that increasing the network depth greatly improved the performance of the model and also presented residual-learning CNN, a technique that employs skip connections and recursive convolution layers which has continued to be prominent in current SR developments [13].

The original SRGAN model was proposed by Ledig et al. in 2017 alongside the Super-Resolution Residual Network (SRResNet) [14]. SRResNet utilizes the architecture as presented in [15] to solve SR problems. The SRGAN model utilizes the foundational architecture of the Generative Adversarial Network (GAN) as presented in [16] in the context of SR as well. The GAN was a revolutionary DL model as it involved training two ML networks, the discriminator and the generator, simultaneously that work together in order to accomplish the task at hand. The discriminator is similar to a DL CNN model while the generator is a novel architecture that employs skip connections, as in [13]. Ledig et al. presented the first instance in literature in which the GAN architecture was utilized in order to conduct SR tasks. [14] also developed a novel "Perceptual Loss Function" used to drive the training of the SRGAN. This research produced promising results as the SRGAN model outperformed all other SR techniques of the time, including the DL CNN-based SR models. A significant result presented in [14] was that the SRGAN models was noted

as recovering the high-frequency features of the HR images, shown in the visual quality of the SR images. The SRGAN's output SR images were observed as having increased perceptibility and appearing to be more photo-realistic, even when compared to the DL CNN SR techniques. These results were part of the drive for making the SRGAN model the primary focus for this thesis.

Recent developments in DL SR models have continued past the inception of the original SRGAN model. [17] developed both a single resolution scale model called the Enhanced Deep Super-Resolution (EDSR) network as well as a multi-resolution-scale model called the Multi-Scale Deep Super-Resolution (MDSR) network that both built upon the residual-learning concepts from [13]. Some of the primary contributions that [17] presented were removing the batch normalization layers within the conventional network architecture as they were found to produce degrading artifacts in the output SR image, increasing the number of output features of each layer within the model, and employing a pre-training strategy for accelerating the training time and enhancing the performance of the MDSR model. [18] built upon the foundational SRGAN architecture from [14] with the Super-Resolution Perceptual Generative Adversarial Network (SRPGAN) by utilizing a perceptual loss function based on the discriminator of the SRPGAN model, combining the Charbonnier-based content loss function with the perceptual and adversarial losses. Zhang et al. [19] developed the Residual Dense Network (RDN) for conducting image SR. The RDN model utilizes residual dense blocks that are densely connected in order to extract features, in a residual learning manner, across the entire architectural block. Wang et al. [20] built further upon the original SRGAN architecture from [14] as well as the residual dense blocks from [19]. [20] developed the novel residual-in-residual dense blocks for use within the SRGAN architecture – while removing the batch normalization layers as in [17] – that joins the multi-level residual learning network architecture with dense connections in order to deepen the network and enhance its performance. The ESRGAN model also employed a relativistic GAN architecture from [21]. The results from [20] found that the ESRGAN outperformed most all of the other SR techniques for each evaluation metric tested. The literature on DL SR models is summarized and detailed further in Table 2.1.

Table 2.1: DL SR Literature Review: All VGG Training Losses are Perceptual Losses

Model	Comparison Methods	Dataset	Downsample Methods	Experiments	Training Loss	Evaluation Metrics	Results
SRCNN 2014 [11]	Bicubic SC K-SVD NE+NNLS NE+LLE ANR	Set5 Set14 ImageNet	Gaussian and Bicubic Interpolation Kernel x2 x3 x4	Dataset, Filter Number, Filter Size, and Resolution Scale	MSE	PSNR	SRCNN outperforms other methods for all datasets and all resolution scales
SRCNN 2016 [12]	Bicubic SC KK NE+LLE A+ ANR	Set5 Set14 BSD200 ImageNet	Gaussian and Bicubic Interpolation Kernel x2 x3 x4	Filter Number, Filter Size, Number of Layers, Color Channels, and Resolution Scale	MSE	PSNR SSIM IFC NQM WPSNR MSSSIM	SRCNN outperforms most other methods for all datasets/resolution scales in PSNR, SSIM, MPSNR, and MSSIM
VDSR 2016 [13]	Bicubic A+ RFL SelfEx SRCNN	Set5 Set14 B100 Urban100	Bicubic Inerpolation Kernel x2 x3 x4	Network Architecture and Resolution Scale	MSE	PSNR SSIM Time	VDSR outperforms all other methods for all evaluation metrics
SRGAN and SRResNet 2016 [14]	Bicubic NN SRCNN SelfExSR DRCN ESPCN	Training: ImageNet Evaluation: Set 5 Set 14 BSD100	Bicubic Interpolation Kernel x4	New Architecture Development and Loss Function	"New Perceptual Loss Function" VGG 19-22 VGG 19-54 Adversarial Loss	PSNR SSIM MOS	SRGAN exceeds all in MOS and SRResNet exceeds all in PSNR and SSIM
EDSR/+ and MDSR/+ 2017 [17]	Bicubic A+ SRCNN VDSR SRResNet	Set 5 Set 14 DIV2K B100 Urban100	Bicubic Interpolation Kernel x2 x3 x4	Dataset, Loss Function, Geometric Self-ensemble, and Resolution Scale	L1	PSNR SSIM	EDSR+ exceeds in most all cases and MDSR+ is second best in most cases
SRPGAN 2017 [18]	Bicubic A+ SRCNN FSRCNN SelfExSR RFL SCN VDSR DRCN LapSRN	Training: T91 BSDS200 General100 Evaluation: Set 5 Set 14 BSDS100 Urban100 Manga109	Bicubic Interpolation Kernel x2 x4 x8	Loss Function and Resolution Scale	"Charbonnier Loss" "Discriminator Based Loss" L1 L2	PSNR SSIM	For the x2 resolution scaling, only exceeds in 1/5 cases. For the x4 resolution scaling, only exceeds in 3/5 cases. For the x8 resolution scaling, exceeds in 4/5 cases
RDN/+ 2018 [19]	Bicubic SRCNN LapSRN DRRN SRDenseNet MemNet MDSR SPMSR FSRCNN IRCNN_C	Training: DIV2K Evaluation: Set 5 Set 14 BSD100 Urban100 Manga109	Bicubic Interpolation x2 x3 x4 Gaussian Kernel x3 Bicubic with added Gaussian Noise x3	Network Architecture, Low-Resolution Simulation Technique and Resolution Scale	L1	PSNR SSIM	RDN+ outperforms all other methods in terms of PSNR and SSIM
ESRGAN 2019 [20]	Bicubic SRCNN EDSR RCAN EnhanceNet SRGAN	Training: DIV2K Flickr2K OST Evaluation: Set 5 Set 14 BSD100 Urban100 PIRM-SR	Bicubic Interpolation Kernel x4	Network Architecture and Loss Function	MINC (Perceptual) VGG 19-22 VGG 19-54	PSNR Ma's Score NIQE	ESRGAN outperforms most other methods in terms of perceptual index (Ma's score and NIQE) while outperforming some in PSNR

2.3 SRGAN

The SRGAN model, since its inception in 2017 by Ledig et al. [14], has proven to be an effective SR technique. It is based on the novel GAN model from [16], which is recognized as the first ML model in literature that employed the use of two ML networks training and working, in tandem, with and against one another. The results for the SRGAN model from [14] revealed that the SRGAN model exceeded all other SR techniques, especially in terms of a person's perceptibility shown by the Mean-Opinion-Score (MOS) test. Some papers in literature argue that the VGG-based perceptual loss is ineffective in recovering the desired high-frequency information [18]. Many others, however, promote the SRGAN model's promising results and usefulness saying that its generative abilities produce detailed SR images even over diverse test samples [10], that the SRGAN model is effective in generating realistic textures [20], and that the visual quality of the SR images generally improve when utilizing the SRGAN compared to other SR techniques [22]. These acclamations further supported the decision for utilizing the SRGAN model as the primary focus for this thesis study. The rest of Chapter 2.3 will detail the theory that drives the SRGAN model as well as the ML network components that comprise the SRGAN model architecture.

2.3.1 Theory

The SRGAN model is comprised of two neural networks, the generator G and the discriminator D, that are trained in an adversarial manner in order to conduct SISR tasks. The goal of SISR is to generate a pseudo-high-resolution, also called super-resolved, image I^{SR} from an input LR image I^{LR}. In order to achieve this, the SRGAN model utilizes input image training pairs where I^{LR} is the corresponding LR image derived from the ground-truth HR image I^{HR}. I^{LR} is created by applying a downsampling bicubic interpolation kernel to I^{HR} with a resolution scale factor of r. Thus, I^{HR} and I^{SR} can be expressed as $rw \times rh \times c$ while I^{LR} is expressed as $w \times h \times c$, with c denoting the number of color channels [14]. Typically, c will equate to either 1 or 3 with $c = 1$ for grayscale images and $c = 3$ for colored images (e.g., RGB colored images, luma, blue-difference chroma, and red-difference chroma (YCbCR) colored images, hue, saturation, and intensity (HSI)

colored images, etc.). All of the images are conisdered to be real-valued tensors. The HR images are used during the training and validation phases of developing a SRGAN model but are absent during the testing phase.

[14] adheres to the same definition of the adversarial min-max problem as defined in [16] in which the discriminator D_{θ_D} and the generator G_{θ_G} are optimized to solve the objective function $O(D_{\theta_D}, G_{\theta_G})$, alternating between which network is being trained at a time. This method of training allows D to settle around its optimal solution, mitigating misclassifications, while ensuring that G is provided with proper feedback from D. The ultimate goal is to train G to generate SR images, with their generation expressed as $G_{\theta_G}(I^{LR})$, that are so similar in their composition to the ground-truth HR images that the classifier network D cannot distinguish between the generated SR images and the ground-truth HR images. The classification process for D is expressed as $D_{\theta_D}(x)$ that defines the probability of x either being real image, i.e. an HR image from the training dataset, or a fake image, i.e. a SR image generated by the generator. D is trained to accurately classify the input image as being an HR image from the training dataset or a SR image generated by G. Therefore, D is being trained to maximize $logD_{\theta_D}(I^{HR})$. Meanwhile, G is being trained to generate pseudo-HR images from the LR images with similar features as in the corresponding HR image pairs. Every iteration of training G is enhanced with feedback from D so that G can produce better and better SR images, with every training loop, until D is unable to tell the difference between a ground-truth HR image and a generated SR image. In other words, G is being trained to minimize $log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})))$. This describes the two-player, minimax game that D and G play during training in order to solve the objective function $O(D_{\theta_D}, G_{\theta_G})$ that is expressly formulated in Equation 2.4 [14, 16]:

$$min_{\theta_G}\, max_{\theta_D} O(D_{\theta_D}, G_{\theta_G}) = \mathbb{E}_{I^{HR}\sim p_{train}(I^{HR})}\left[logD_{\theta_D}(I^{HR})\right] + \mathbb{E}_{I^{LR}\sim p_G(I^{LR})}\left[log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})))\right] \tag{2.4}$$

Here, D and G are deep neural networks that are driven by parameters θ_D and θ_G, respectively. Both θ_D and θ_G can be expressed as $\{W_{1:L}, B_{1:L}\}$ with W denoting the weights and B denoting the biases within a L-layer deep network. The parameters for G are obtained through the SRGAN training process by optimizing l^{SR} over a training dataset of size $n = 1, ..., N$ containing image pairs (I_n^{LR}, I_n^{HR}). This is expressly written as in Equation 2.5 [14]:

$$\theta_G = arg\, min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l^{SR}(G_{\theta_G}(I_n^{LR}), I_n^{HR}) \tag{2.5}$$

In the case of the SRGAN model, l^{SR} is a novel perceptual loss function developed by [14]. This loss function drives the generator network training, enabling G to generate SR images with features of a higher perceptibility. It is composed of a weighted sum of an adversarial loss l_{Gen}^{SR} from the generative network G as well as a content loss l_X^{SR} as shown in Equation 2.6 [14]:

$$l^{SR} = l_X^{SR} + 10^{-3} l_{Gen}^{SR} \tag{2.6}$$

The inclusion of the generative loss allows the network to favor SR images that can fool D during training. The generative loss l_{Gen}^{SR} depends on $D_{\theta_D}(G_{\theta_G}(I^{LR}))$ which represents the probabilities classified from D. In Equation 2.6, G was being described as minimizing $log(1 - D_{\theta_D}(G_{\theta_G}(I^{LR})))$ through training. However, when this definition of the generative loss function is used for l_{Gen}^{SR}, insufficient gradients are given to G during training. During the first set of training loops, the generative model G is unable to produce well-representative SR images and D is able to classify them easily as they do not look similar to the HR images from the training dataset, without much training to D. Thus, in practice, it is preferred to train G to minimize $-log(D_{\theta_D}(G_{\theta_G}(I^{LR})))$. This gives G gradients from which it can learn more effectively during the first iterations of training. This expression is formalized in Equation 2.7 [14]:

$$l_{Gen}^{SR} = \sum_{n=1}^{N} -logD_{\theta_D}(G_{\theta_G}(I^{LR})) \tag{2.7}$$

One of the most prominent content losses l_X^{SR} that had been used in early DL SR techniques is a pixel-wise mean squared error (MSE) loss or L2 loss [11–13]. Using the MSE loss as the optimization target allowed the DL SR models to achieve higher evaluations on the peak signal-to-noise ratio (PSNR) metric [11]. This was a benefit as PSNR is one of the most widely used metrics in literature for evaluating SR techniques, and still is as shown in Table 2.1. The pixel-wise MSE content loss is expressed in Equation 2.8 [14]:

$$l_{MSE}^{SR} = \frac{1}{r^2 wh} \sum_{x=1}^{rw} \sum_{y=1}^{rh} (I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2 \tag{2.8}$$

Ledig et al., however, argued that MSE-driven optimization problems, in terms of SR tasks, resulted in less distinct high-frequency details being portrayed within the generated SR image as opposed to the sharper details seen in the corresponding target HR image. To this end, Ledig et al. defined the VGG loss to drive more perceptually-relevant SR solutions. This method involves using the feature maps of a pre-trained CNN in a frozen state, i.e. the layers are not updated during trainning allowing the pre-trained CNN to act as a fixed feature extractor. Both generated SR and ground-truth HR images are pushed through the first few layers of the CNN, which outputs feature maps. Feature maps are intermediate representations of specific features within the input image. The difference between the SR image's resulting feature map and the HR image's resulting feature map can then be employed as a loss function and minimized in order to train the model. [14] based their perceptual loss on the VGG-19 model, a convolutional network with 19 weight layers, developed by [23] that was trained on the ImageNet dataset [24]. In order to formulate this, [14] defined $\phi_{j,k}$ as the feature map produced by the j-th max pooling layer preceded by the k-th convolution layer. The VGG loss is then expressed as the pixel-wise sum of squares of the difference between the feature maps of the ground-truth HR images and the generated SR image as shown in Equation 2.9 [14]:

$$l^{SR}_{VGG/j,k} = \frac{1}{w_{j,k}h_{j,k}} \sum_{x=1}^{w_{j,k}} \sum_{y=1}^{h_{j,k}} (\phi_{j,k}(I^{HR})_{x,y} - \phi_{j,k}(G_{\theta_G}(I^{LR}))_{x,y})^2 \tag{2.9}$$

2.3.2 Architecture

The SRGAN is a complex, deep-learning model with an adversarial network architecture comprised of two neural networks connected together, designated as the generator G and the discriminator D. G is modeled as a feed-forward CNN that trains to develop a feature map of the LR image inputs in order to generate SR images that are representative of the corresponding HR images from the training pairs. D is modeled as an image classifier CNN that trains to classify the input image as either a real/HR image or a fake/SR image. The discriminator's output is provided as feedback to G during the training process, enabling G to improve the visual quality of its generated SR images. After thoroughly training both networks, the goal is that G will produce SR images that are highly representative of the HR target image, so much so that D will not be able to distinctly classify the input SR image as either real or fake. This section details the architectural composition of both G and D in terms of their computational layers.

The main layers that make up G and D are the input, 2D convolution (2DConv), activation, leaky rectified linear unit (LeakyReLU), batch normalization (BN), add, 2D upsampling (2DUp), and dense layers. This thesis utilizes the Tensorflow Keras Application Programming Interface (API) for writing the programming code to develop the SRGAN model. Thus, all of the layer descriptions that follow are based on the Tensorflow Keras layer definitions. The input layer instantiates the input as a tensor and defines its tensor shape. This formats the input, preparing it for further computation. For this research, the input tensor for G is an input LR image while the input tensor for D is an input HR or SR image.

The 2DConv layer creates a convolution kernel that is used to conduct spatial convolution over the input tensor. Convolution is defined as the element-wise multiplication sum of two matrices. In the case of the SRGAN model, the two matrices are the convolution kernel K and the input image

I convolved together to create the output tensor T. The convolution computation can be expressed as an element-wise sum of products as shown in Equation 2.10:

$$\begin{aligned} T_{w \times h \times f} &= K_{k_w \times k_h} * I_{w \times h \times c}[x, y] \\ &= \sum_{dx=-a}^{a} \sum_{dy=-b}^{b} (K[dx, dy] \cdot I[x + dx, y + dy]) \end{aligned} \tag{2.10}$$

Every element of the convolution kernel is defined by $-a \leq dx \leq a$ and $-b \leq dy \leq b$. The convolution kernel matrix's height k_h and width k_w is specified in the 2DConv layer by the $kernel\_size$ variable. For the SRGAN model, both spatial dimensions of the 2D convolution window are equivalent in all 2DConv layers. The kernel's values are also called weights. The weights that make up a convolution kernel drive how the convolution functions. Some common convolution kernels and their defined weights are the identity kernel $[[0, 0, 0], [0, 1, 0], [0, 0, 0]]$, the sharpening kernel $[[0, -1, 0], [-1, 5, -1], [0, -1, 0]]$ and the gaussian blur kernel $\frac{1}{16}[[1, 2, 1], [2, 4, 2], [1, 2, 1]]$. The initial set of weights used in the 2DConv layer are determined by a kernel initializer. But the weights change and are updated through the training process. For this thesis, all 2DConv layers used the default Tensorflow Keras kernel initializer known as the Glorot uniform intializer or the Xavier uniform initializer. Each of the initial weights w_i are defined as shown in Equation 2.11:

$$\begin{aligned} w_i &= \{x : x \in U(-a, a)\}, \\ \text{where } a &= \sqrt{\frac{6}{units_{in} + units_{out}}} \end{aligned} \tag{2.11}$$

where $units_{in}$ is the number of input units in the weight tensor and $units_{out}$ is the number of output units. Before convolving, the input tensor is padded by setting the $padding$ variable to "$same$". Padding the tensor means that additional rows and columns containing zeros are appended above and below as well as to the left and right of the input tensor, respectively. The

process of convolving the kernel matrix with the input tensor is controlled by the $strides$ variable of the 2DConv layer. This variable defines how the kernel moves along the input tensor's height and width, shown in Equation 2.10 as dy and dx, respectively. For example, a stride of (1, 3) specifies that the kernel will move only one space in the height dimension but will move 3 spaces in the width dimension during the convolution process. For the SRGAN model, the number of strides is the same value for both spatial dimensions. The $filters$ variable of the 2DConv layer is an integer that defines number of filters in the output convolution. This is expressed in Equation 2.10 as f and it defines the channel/depth dimension of the output tensor.

Sometimes, the $activation$ variable in a 2DConv layer will be set, in which case the specified activation function will be applied to the output of the convolution function. For the SRGAN model, the $activation$ variable is only applied for the final 2DConv layer of G. The activation function that is applied is the "tanh" activation function. The "tanh" activation function $a(x)$ expressed as in Equation 2.12. All of the other 2DConv layers in the SRGAN model do not apply an activation function.

$$\text{Tanh:} \quad a(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.12}$$

The activation and LeakyReLU layers are similar in nature. Both of these layers apply an activation function to the output of a preceding layer. Activation functions are an essential component to the learning process of ML models. They determine whether a neuron should be activated or not based on the importance of the neuron's input to the network's prediction. The main difference between the two Tensorflow Keras layers is that the activation layer allows for the use of any built-in activation function where as the LeakyReLU layer is specifically for the use of the LeakyReLU activation function. All of the activation layers used in this thesis' SRGAN models uses the rectified linear unit (ReLU) activation function as defined in Equation 2.13. The LeakyReLU function is similar except that when $x \leq 0$ (i.e., the unit is not active), a constant a is applied to x, as shown in Equation 2.13, typically allowing for a small gradient in the negative values of x.

$$\text{ReLU}(x) = \begin{cases} x & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases} \quad \text{and} \quad \text{LeakyReLU}(x) = \begin{cases} x & \text{if } x > 0, \\ ax & \text{otherwise.} \end{cases} \tag{2.13}$$

The BN layer works to normalize the input, maintaining the output's mean close to 0 and its standard deviation close to 1. However, it should be noted that this layer functions differently during training versus when making a prediction with the trained model. During training, the BN layer uses the computation for normalizing the input as expressed in Equation 2.14:

$$\frac{\gamma * (batch - \text{mean}(batch))}{\beta + \sqrt{\text{var}(batch) + \epsilon}} \tag{2.14}$$

where γ is a learned scaling factor that is initialized to 1 and updated during training, $batch$ refers to the batch of inputs (i.e., the number of training samples used during one iteration of the training loops), β is a learned offset variable initialized to 0 and updated during training, ϵ is a configurable constant, and mean() and var() are the average and variance functions, respectively. During predictions, after the ML model has been trained, the layer will normalize using a moving average of the mean and standard deviation learned from the training batches. This is expressed as in Equation 2.15:

$$\frac{\gamma * (batch - moving_{mean})}{\beta + \sqrt{moving_var + \epsilon}};$$

$$\text{where } moving_{mean} = moving_{mean} * momentum + \text{mean}(batch) * (1 - momentum); \quad (2.15)$$

$$\text{and } moving_{var} = moving_{var} * momentum + \text{var}(batch) * (1 - momentum)$$

where $moving_{mean}$ and $moving_{var}$ are non-trainable variables that are updated via the training process. Thus, the model must be trained on data with similar statistics as the prediction/testing data first before the BN layers will normalize the input when making predictions. The $momentum$ variable is used to compute the moving average $moving_{mean}$ and moving variance $moving_{var}$ variables. The $momentum$ is typically a float value that follows $0 < momentum < 1$. Therefore, the $momentum$ can be thought of as a weight factor that favors the $moving_{mean}$ and $moving_{var}$ variables (i.e., the mean and variance based on all previous batches of training samples) when $0.5 < momentum < 1$ and favors the mean$(batch)$ and var$(batch)$ variables (i.e., the mean and variance based on the batch of training samples for the current iteration of training) when $0 < momentum < 0.5$.

The description of each of the remaining three layers is deceptively simple but they still play crucial roles in the makeup of the SRGAN architecture. The add layer takes the inputs and outputs a tensor that is the sum of the inputs with the same size as the inputs. This element-wise addition also acts as a skip connection. Skip connections allow for inputs of earlier layers to influence the output of later layers. This layer helps to preserve gradients and carry information derived from earlier feature maps and input data throughout the network. By carrying information throughout the network, deeper and more complex abstractions of the input data can be learned during training, helping the ML model infer more intricate details when making predictions. The 2DUp layer

upsamples the input by repeating the rows and columns of the input. The number of repetitions is controlled by the $size$ variable. For this thesis' implementation, the x2 resolution scale generator contains one 2DUp layer while the x4 resolution scale generator contains two 2DUp layers. This layer transforms the input tensor size into the size of the HR regime in terms of the spatial dimensions. The feature map filters are still present after passing through the upsampling layer. The dense layer is a densely-connected layer that is primarily used in D. This is a fundamental component of most neural networks. The Tensorflow Keras dense layer computes the dot product between the input and a kernel and adds to it a bias vector. Then, an activation function is applied to the resulting sum. This is expressed as in Equation 2.16:

$$Output = Dense(input) = activation((input \cdot kernel) + bias) \tag{2.16}$$

where the kernel and bias vectors are initialized by the dense layer. The dense layer is configured by the $units$ and $activation$ variables. The $units$ variable is a positive integer that defines the dimensionality of the output space. The $activation$ variable is a string that defines which activation function is used in the layer. The $activation$ variables used within this SRGAN model architecture's dense layers are the "linear" and "sigmoid" activation functions. These activation functions $a(x)$ are expressed as in Equation 2.17:

$$\begin{aligned} \text{Linear:} \quad & a(x) = x \\ \text{Sigmoid:} \quad & a(x) = \frac{1}{1 + e^{-x}} \end{aligned} \tag{2.17}$$

With the layers defined, the architectures of the neural networks G and D can be detailed. The generator neural network G takes a LR image of the image training pair as an input and outputs a SR image that is an estimation of the corresponding HR image from the image training pair. When input into G, the LR image is first instantiated as a tensor via the input layer of shape $w \times h \times c$.

For this thesis, the LR size is defined as either $128 \times 128 \times 3$ or $64 \times 64 \times 3$ depending on whether the resolution scale factor used was x2 or x4, respectively. The LR image then passes through a 2DConv layer and a ReLU activation layer. Then, this initial output is passed through a series of residual blocks made up of a 2DConv, a ReLU layer and a BN layer followed by a 2DConv, a BN, and an add layer. The add layers create skip connections that connect the input of each residual block with the add layer at the end of each residual block. After the series of residual blocks, the input image is passed through a 2DConv, a BN and an add layer. The skip connection created by this add layer connects the initial output – the output of the first ReLU layer that preceded the series of residual blocks – with this final element-wise sum add layer. The output is then passed through upsample blocks that are made up of a 2DUp, a 2DConv and a ReLU layer. For the x2 resolution scale experimental SRGAN models, only one of these upsample blocks is used in the generator while two upsample blocks are used in the generator for the x4 resolution scale experimental SRGAN models. A final 2DConv layer is utilized to transform the filters of the output tensor's channel dimensionality into the same resolution space as the HR images and output the predicted SR image of the SRGAN model. This 2DConv layer also applies a *tanh* activation function to its output. A diagram of the generator architecture is provided in Figure 2.1. This

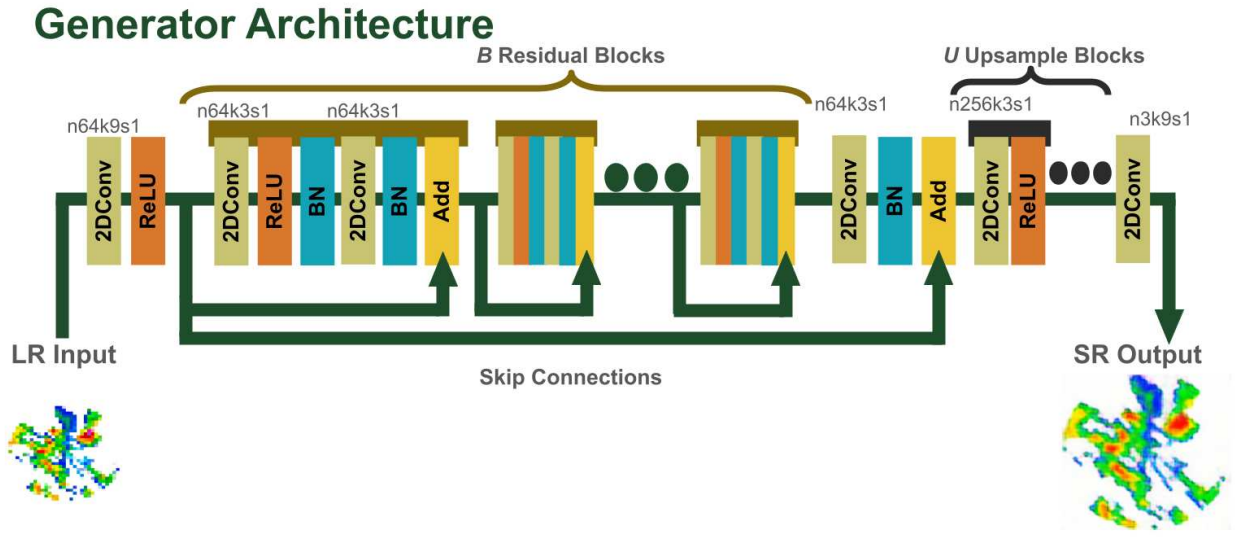

Figure 2.1: Generator Architecture diagram denoting the number of filters (n), kernel size (k) and stride (s) of the 2DConv layers as well as the number of residual blocks (B) and the number of upsample blocks (U).

thesis' experimental SRGAN models closely follow the network architecture defined as in [14].

A diagram of the discriminator architecture is provided in Figure 2.2. The discriminator neural network D takes either a HR image from the image training pair or a predicted SR image from G as an input and outputs a feature map that denotes whether or not the input image is a real image (indicating that it is more similar to the ground-truth, HR image) or a fake image (indicating that it is more similar to the generated, SR image). The input HR/SR image is first instantiated as a tensor via the input layer of shape $rw \times rh \times c$. For this thesis, the size of both the HR and SR images is defined as $256 \times 256 \times 3$. The HR/SR input image is then passed through a series of computational layers. First, the input image is passed through a 2DConv and a LeakyReLU layer. The rest of the layer blocks are made up of a 2DConv, a LeakyReLU, and a BN layer. Typically, there are four pairs of layer blocks within D, following the original SRGAN model presented by [14]. The $kernel\_size$ variable for each 2DConv layer is set to 3 while the $strides$ variable alternates between 1 and 2 for the first and second layer blocks in each pair, respectively. The $filters$ variable starts at 64 and then increases by a factor of 2 for each layer block pair. The output is then passed through a dense layer with the $activation$ variable set to the linear activation function and the $units$ variable set to the $filters$ variable of the previous layer block's 2DConv layer increased by a factor of 2. A LeakyReLU activation layer is then applied to the output followed by another pass through a dense layer. This dense layer has the $activation$ variable set to the sigmoid activation

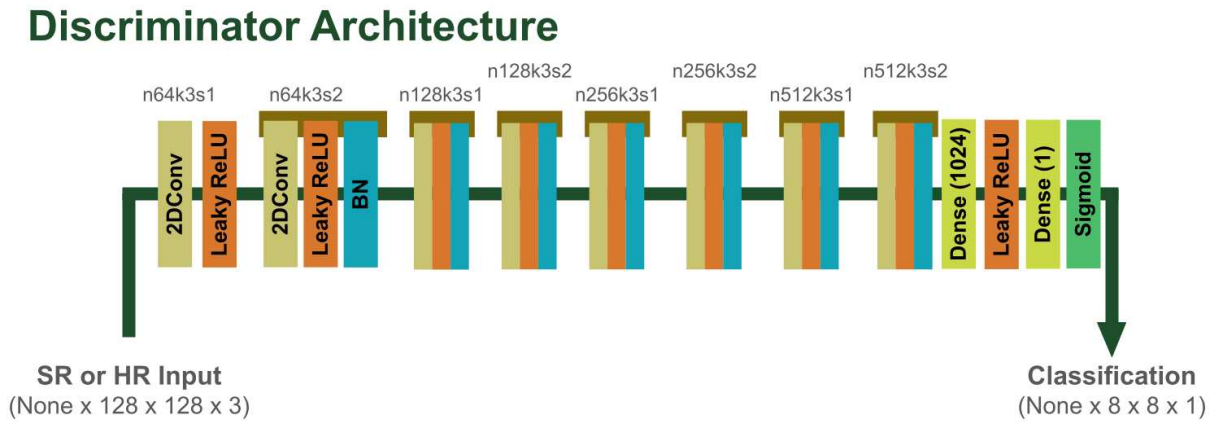

Figure 2.2: Discriminator Architecture diagram denoting the number of filters (n), kernel size (k) and stride (s) of the 2DConv layers.

function and the $units$ variable set to 1. The resulting output is the probability of classification for the input sample as either a real, corresponding to the HR target image, or a fake, corresponding to the SR generated image.

2.4 Literature on Deep Learning Super-Resolution in Weather Radar

DL SR techniques have proven to be quite useful and effective. Many of the DL SR models discussed thus far have been developed for super-resolving images within either the natural image manifold – such as animals, landscapes, buildings and people's faces – or for digital art [11–14, 17–20]. Beyond these general use cases, many different types of SR models, including the SRGAN model, have been implemented across many different fields, with applications including: medical diagnosis, fingerprinting, surveillance, microscopy, astronomy, and more. Comprehensive studies of SR applications can be found in [2, 25]. The scope of this thesis study focuses on the use of DL SR techniques, specifically the SRGAN model, for use within the weather radar regime. There exists an extensive amount of studies in literature in which DL models are utilized in remote sensing for various purposes including storm tracking [26], forecasting/nowcasting [27, 28] and classification [29], for example. A comprehensive study on the use of DL techniques within remote sensing and their applications can be found in [30]. The literature also contains many papers studying SR applications with weather radar data that do not use a DL model [31–33]. However, research in DL SR models has recently been increasing. Most DL SR research in remote sensing, and weather radar in particular, utilize a CNN model, as in [34–36]. There is sparse representation in the literature of the SRGAN model within the field of weather radar. A thorough search of the relevant literature yielded only one related article of the SRGAN model implemented for weather radar [37], showing that this implementation of DL SR is in need of further study.

Chen et al. followed the SRGAN architecture as developed in [14]. The training dataset for [37] comprised data from the S-band China New-Generation Weather Radar (CINRAD-SA) as well as the X-band dual-polarization radar (XPRAD) provided by the China Meteorological Administra-

tion. Chen et al. utilized bicubic interpolation, iterative back propagation (IBP), and nonlocal self-similarity sparse representation (NSSR) [31] techniques as their comparison methods. PSNR and the structural similarity index measure (SSIM) were used for the evaluation metrics. Interestingly, during post-processing, Chen et al. cropped the HR images within the dataset into many smaller images. This resulted in a larger dataset pool with smaller sizes for more efficient training. [37] found that the SRGAN model outperformed all other methods when performing SR on the weather radar scans. This demonstrates the effectiveness of the SRGAN model within the weather radar regime.

Table 2.2: DL SR in Weather Radar Literature Review: All VGG Training Losses are Perceptual Losses

Model	Comparison Methods	Dataset	Downsample Methods	Experiments	Training Loss	Evaluation Metrics	Results
SRGAN 2019 [37]	Bicubic IBP NSSR	CINRAD-SA and XPRAD	Gaussian and Bicubic Interpolation Kernel x2 x4	Data Products and Resolution Scale	L1	PSNR SSIM	SRGAN outperforms all other methods in terms of PSNR and SSIM

DL SR models are just recently starting to be applied to the weather radar field. Preliminary results are promising, showing that DL SR models are effective techniques in super-resolving weather radar data. However, the literature on the SRGAN DL SR model applied to weather radar data is found to be under-studied. Thus, in order to enhance the literature and promote further study within this field, this thesis aims to research the effectiveness of the SRGAN DL SR model in super-resolving weather radar data.

Chapter 3

Weather Radar and Observations

Weather radar systems are vital in ensuring the safety of society by providing insightful data that helps researchers and scientists better understand and predict the behavior of one of nature's most relentless and omnipresent forces affecting the world: weather. Atmospheric and geographic properties such as humidity, temperature, pressure, etc. and mountain ranges and oceans, for example, combine and interact in complex ways with one another to make up what is known as weather. Weather affects people's daily lives and its effects can vary greatly. Transportation and agricultural industries, for example, utilize weather forecasts to determine which routes to take to reach a destination or when to begin planting and harvesting crops. Entertainment industries use weather forecasts to schedule premiere events. The populace in general relies on weather forecasts everyday to help decide on what clothes to wear and when to go out. In more extreme situations, emergency response agencies depend on accurate forecasts to prepare and issue warnings for severe weather phenomena and natural disasters. Blizzards, hurricanes and tornadoes and their aftermath such as fallen debris, flooding, landslides, and wildfires necessitate timely, accurate preparations to mitigate their fallout. All of this changes with the weather, and weather radar systems collect the most useful data that contains a wealth of information about weather events of interest. Weather radars are renowned for producing the highest quality weather event data compared to other weather sensing instruments such as raingauges, disdrometers and wind profilers.

One of the most effective and predominant methods for weather data collection is weather radar, especially for forecasting. Weather radars are complex electromechanical systems that interface many different components with one another in order to scan and collect information about volumetric targets. This information comes in the form of voltages and currents of energy signatures. But, once processed, valuable and physically meaningful information can be determined. Weather radar data gives insight into the characterization and developmental process of the weather event being observed. The starting time, place of origin, movement, rate of development, and clas-

sification of weather events can all be discerned by analyzing weather radar data. Due to its usefulness, weather radar data has been driving force, continuously being used to expand and develop areas of research for not only meteorological and atmospheric science applications such as intense weather event detection, tracking and modeling but also, more recently, weather radars have been utilized for space observation as well. Since weather radars are a vital source of information for maintaining the safety of society and are widely used throughout various fields of scientific study, this thesis focuses its efforts in evaluating the efficacy of utilizing the SRGAN model within the context of weather radar SR on LR weather radar scans. The impacts of this research are substantial as, if found to be effective, the proposed model could significantly enhance the operational paradigm of weather radars as a whole.

Chapter 3 will explain in detail important concepts for weather radar implementation. Chapter 3.1 will discuss the principles behind weather radar data collection from the weather radar data components to the signal propagation. Chapter 3.2 describes the fundamental scan strategies used within weather radar field for data visualization and analysis. Chapter 3.3 explains the fundamental equations behind producing the radar reflectivity factor, a crucial radar output widely used in weather radar data analysis and visualization. Chapter 3.4 discusses the importance of HR data collected from weather radars and how DL SR techniques using digital signal/image processing can increase the availability of SR weather radar data while enhancing the performance of weather radar data collection.

3.1 Weather Radar Principles

Weather radars are complex systems that require many components operating in tandem simultaneously. First, the central control and communications processor initiates control signals telling the radar how and where to scan. These are sent to the motion controller and the signal processor. The motion controller sends drive signals to the positioner and the antenna to move the radar system in the desired scan strategy. The signal processor triggers the transmitter to start transmitting the electromagnetic (EM) pulse and triggers the polarization switch to begin alternating

the transmitted signal between the horizontal and vertical polarization channels. The transmitter propagates the EM pulses and samples their phase for use in the receiver. The EM pulses from the transmitter pass through the duplexer – which allows the transmitted and received signals to be isolated from one another while sharing a common antenna – and then through the polarization switch. From there, the EM transmit pulses are sent to the antenna via waveguides. The antenna then directs the EM pulse to the reflector which focuses and projects the pulse as a beam through the atmosphere. The angular resolution, i.e. the beamwidth, of the radar's transmitted beam depends upon the transmit frequency and the size of the antenna reflector. Here, the beamwidth is, more specifically, referring to the half power beamwidth (HPBW). The HPBW can be found using an antenna radiation pattern by creating a line that connects the radiation pattern origin to the half power points on each side of the antenna radiation pattern's major lobe. The angle between that separates these lines is the HPBW. The HPBW is the angle that defines the area in which magnitude of the radiation pattern is 50% of the main beam's peak power. This is equivalent to reducing the peak power/gain by 3 dB. The HPBW is calculated as the fraction between the wavelength and the diameter multiplied by a beam factor scalar as shown in Equation 3.1 [38]:

$$HPBW = \Delta\theta = \frac{kc}{Df} = \frac{k\lambda}{D} \tag{3.1}$$

in which k defines the beam factor that varies depending on the antenna configuration and design, c is the speed of light constant, f is the radar transmit frequency, D is the diameter of the radar antenna, and λ is the wavelength of the radar's transmitted signal with the relation $\lambda = \frac{c}{f}$ to the radar's transmit frequency. When the EM pulse collides with particles in the atmosphere, the signal is scattered in all directions. A portion of the energy is back-scattered, i.e. reflected back towards the radar; larger particles result in a greater amount back-scattering and the farther the transmitted signal travels, the more it weakens. Depending on the radar transmit frequency as well as the application, the wakening of the signal can be mitigated in post-processing using attenuation correction. This energy signature carries information about the particles back to the radar system. This is based on the phase and the magnitude of the received signal in comparison to the transmitted

signal and the time that the signal traveled from transmission to reception. For weather radar, the target particles are typically hydrometeors/precipitation particulates (e.g. rain, hail, graupel, snow) within storm systems; however, atmospheric particles from other natural phenomenon that do not include precipitation, such as airborne debris from tornados, insect swarms, and smoke from wild fires, can also be observed via weather radar. The antenna collects the back-scattered signal, now called the received signal, and sends it back through waveguides to the polarization swith, the duplexer and then to the receiver. The receiver takes the received signal and a phase sample of the transmitted signal from the transmitter and processes the voltages and currents from the received signal into its real and imaginary constituents [39]. The received signal can be expressed as a complex signal of the form shown in Equation 3.2:

$$r(t) = M(t)e^{jP(t)} \tag{3.2}$$

in which $M(t)$ represents the magnitude function and $P(t)$ represents the phase function. Based on Euler's Formula $e^{ix} = \text{cis}(x) = \cos(x) + i\sin(x)$, the signal can be written into its cosine and sine components as: $r(t) = M(t)\cos(P(t)) + jM(t)\sin(P(t))$. The components $M(t)\cos(P(t))$ and $M(t)\sin(P(t))$ are also referred to as the in-phase and quadrature-phase (I/Q) components, respectively, due to the orthogonal nature of cosine and sine waves. Thus, receiver processes the received signal $r(t)$ into I/Q data [40]. The I/Q data, alongside angle positioning data from the motion controller, are sent to the signal processor which generates the spectral moments of the radar data. This is then sent to the central control and communications processor for data dissemination and further radar product generation processing [39].

This thesis focuses on modern dual-polarization, Doppler weather radar. Dual-polarization describes radars that transmit the EM pulses and receive the back scattering of those pulses in both a horizontal and a vertical orientation. Having both polarization signals allows the radar system to collect additional information about the weather target of interest. From this, additional radar moments can be calculated that give further insight into the storm system's composition – including the size, shape, and homogeneity of the hydrometeors within – as well as help distinguish between

the received signal and any ground clutter that could be affecting the radar observation. Doppler radars are able to collect information as to the position and movement of the meteorological targets of interest by analyzing the phase shift (difference) between the transmitted signal and the back-scattered echo. This is conceptually similar to the Doppler shift commonly thought of when observing sound waves. It should also be noted that the transmitter and receiver are commonly combined in modern radar systems into a single device called a transceiver that both transmits and receives the radar signals.

The properties of the radar scan are determined by the physical characteristics of the radar build as well as how the radar scan strategy is configured. When researching the implementation of SR techniques in weather radar, an important consideration is the radar's antenna scan rate and the number of samples used while scanning. A radar's antenna scan rate describes how fast a radar is conducting a scan. This is formulated as in Equation 3.3 [40]:

$$\theta_s = \frac{\Delta\theta}{D_t} = \frac{\Delta\theta}{N_b\,(\mathrm{PRT})} \tag{3.3}$$

where the antenna scan rate θ_s, with its units in deg/s, is defined as the quotient between the antenna beamwidth $\Delta\theta$, with its units in degrees, deg, and the dwell time D_t, with its units in seconds, sec. The antenna beamwidth $\Delta\theta$ defines the resolution of the scanning angle in either azimuth or elevation depending on the type of radar scan strategy used. It can also be used to determine the signal strength of the radar transmission. The dwell time D_t is described as the data collection time, i.e. the amount of time in which the transmitted EM pulse is hitting a target. D_t can also be formulated as the product between the number of beam pulses N_b used in a radar scan and the pulse repetition time (PRT) with units in sec as $D_t = N_b(PRT)$. A radar scan typically consists of a train of EM pulses transmitted from the radar antenna. N_b defines the number of pulses within the transmitted train. PRT describes the cycle of transmitted pulses, i.e. the time period from the start of one pulse in the train to the start of the next pulse in the transmitted train. Rearranging Equation 3.3 allows for N_b to be written in terms of $\Delta\theta$, θ_s and the PRT as shown in Equation 3.4:

$$N_b = \left(\frac{\Delta\theta}{\theta_s\,(\text{PRT})}\right) = \left(\frac{\Delta\theta f_p}{\theta_s}\right) = \left(\frac{\Delta\theta f_p}{6w_m}\right) \tag{3.4}$$

where the PRT is converted into the pulse repetition frequency f_p, with units in hertz Hz, as the PRT is the inverse of f_p which is expressed as PRT $= \frac{1}{f_p}$. The antenna scan rate θ_s is also converted from deg/s into the antenna scan rate w_m in revolutions per minute rev/min. As there are 360 degrees within a single revolution and 60 seconds in a minute, this conversion is formulated as $w_m = \frac{\theta_s deg}{sec}\frac{rev}{360deg}\frac{60sec}{min} = \frac{\theta_s}{6}\frac{rev}{min}$ which can also be written as $\theta_s\frac{deg}{sec} = 6w_m\frac{rev}{min}$. Both $\Delta\theta$ and the PRT or f_p are considered to be constants determined by the radar antenna and system design. Thus, the only variable that can be manipulated within Equations 3.3 and 3.4 is the antenna scan rate θ_s or w_m. This relationship is important as it demonstrates that increasing the antenna scanning rate to double or quadruple its original value will correspondingly half or quarter the number of overall pulses, respectively, resulting in a LR radar scan. Therefore, in order to properly simulate radar data from a LR radar scan, with higher antenna scan rates, the number of pulses within the scan must be fractioned accordingly. This concept is carefully employed in this thesis when generating the LR dataset from the HR dataset for use when training the SRGAN model. This will be further explained in Chapter 4.

3.2 Scan Strategies

Weather radar scan strategies define how the radar moves and, thus, the area of interest where the radar is scanning. The radar system's movement is controlled by the motion controller's signals that are sent to the drive motors within the positioner. Radar motion is typically defined with two angles: azimuth and elevation. The azimuth angle of a radar changes the pointing direction of the radar around its vertical axis, allowing the radar to scan across a horizontal plane. When looking directly at a radar, this can be interpreted as a left-to-right motion. Most radars are capable of changing their azimuth angle by a full 360°. The elevation angle of a radar rotates the radar dish around its horizontal axis. This allows the radar to point higher and lower in the sky. When looking

directly at a radar, this can be thought of as an up-and-down motion. It is common practice for most radars to limit their elevation angle motion from 0 to 90° even if the positioner is capable of exceeding these limits. However, due to the possibility of the reflector dish colliding with the positioner, the change in elevation angle movement is typically restricted. Figure 3.1 illustrates the elevation and azimuth angles with respect to the radar scan.

Scanning the radar while changing the azimuth and elevation angles allows the radar to collect data within different sections of the weather event under observation. This thesis utilizes two of the primary radar scan types: Range-Height Indicator (RHI) and Plan-Position Indicator (PPI). RHI weather scan strategies are defined as scanning the radar along different elevation angles while maintaining a constant azimuth. The radar then collects data in radial sweeps of a vertical sector within the meteorological volume of interest. Once processed, visualization of the vertical sector from the RHI data can give information on the layers that makeup a storm system. For example, RHI scans are used to determine the melting layer within a storm. Hydrometeors that reside within higher altitudes of the atmosphere, in upper layers of a storm system, are typically in a frozen state as ice crystals as higher altitudes results in lower temperature. As the altitude decreases, the temperature increases. Thus, as a hydrometeor falls and approaches the earth, the frozen crystalline structure begins to melt and transform into other hydrometeors. The melting layer is defined as the altitude at which the temperature has increased enough so that the frozen hydrometeors melt into rain, snow, sleet or graupel. The melting layer is an important part of hydrometeor classification, which is possible due to RHI weather radar data. PPI weather scan strategies are defined as scanning the radar along different azimuth angles while maintaining a constant elevation. The radar then collects data in radial sweeps of a horizontal, cone-shaped sector within the natural phenomenon of interest. This gives us a pseudo-horizontal slice of the meteorological volume. Once processed, visualization of the horizontal sector of the storm system from the PPI data can give information such as the velocity, the positional history/tracking, the type and behavior of the storm, as well as the span of the storm itself. Figure 3.1 portrays the differences between the RHI and PPI scan types.

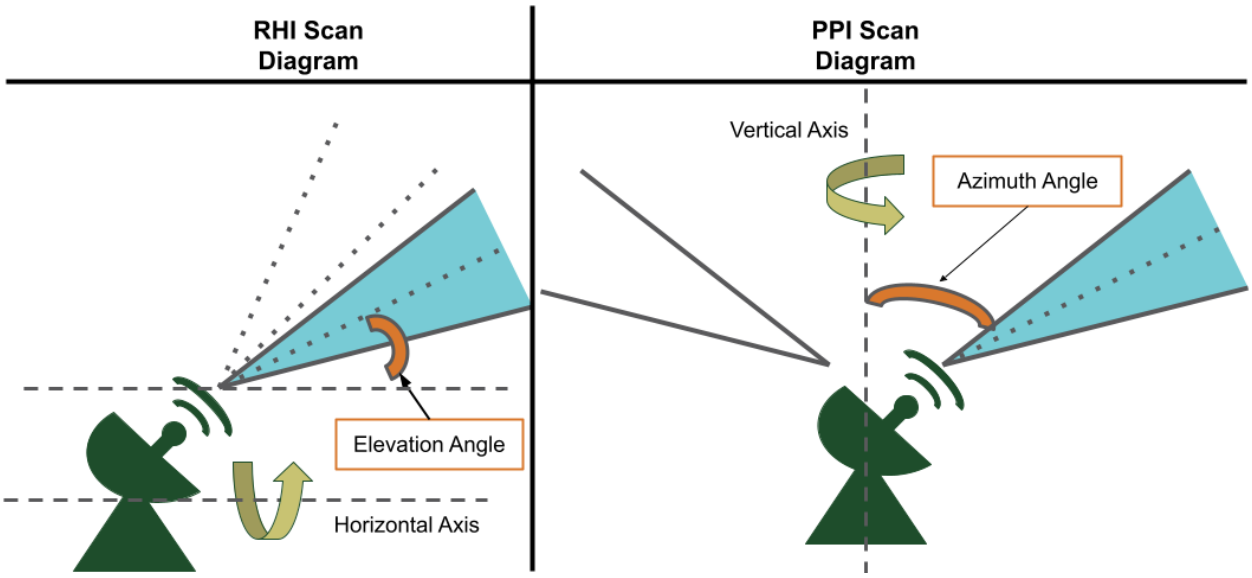

Figure 3.1: Diagram of Radar Scan Types: RHI and PPI

3.3 Weather Radar Moments

When the transmitted EM beam collides with particulates as it travels through the air, the energy is scattered in all directions. A portion of the energy being propagated is reflected back towards the signal's origin, the radar. This is called the back scatter. A measurement, referred to as the radar reflectivity η, can be made that describes the radar's efficiency in intercepting and receiving the returned energy signature. The radar reflectivity is defined as the back scattering cross section per unit volume. The size, shape, aspect and dielectric properties of the radar target all affect the radar reflectivity. For meteorological targets, this corresponds to the size, relative shape or shapes, physical states (e.g., ice or water), aspect and the number of hydrometeors within the volume of interest. This is formulated as in Equation 3.5:

$$\eta = \sum_{h} N_h \sigma_h \tag{3.5}$$

where the summation is over all of the hydrometeors within a unit volume, N_h is the number of hydrometeors per unit volume and σ_h is the back scattering cross section. For Rayleigh scattering due to a dielectric sphere, $\sigma_h = \frac{\pi^5}{\lambda^4}|K|^2 D^6$ well approximates the back scattering cross section of

a small, spherical water drop of diameter D and dielectric constant K [38, 40]. Substituting in the definition for σ_h into Equation 3.5, the radar reflectivity can be written as written in Equation 3.6:

$$\eta = \frac{\pi^5}{\lambda^4}|K|^2 \sum_h N_h D^6 = \frac{\pi^5}{\lambda^4}|K|^2 \int_D D^6 N(D) dD \tag{3.6}$$

in which the summation is converted into an integral of the drop-size distribution $N(D)$ as a continuous function of drop size D. This integral is also used to define one of the primary radar moments, the radar reflectivity factor Z. Moments typically describes radar variables that are directly output by the radar signal processor and are important for use in radar data analysis and visualization within the weather radar field. As the integral in Equation 3.6 sums the sixth-powers of the hydrometeors' diameters within a unit volume of space, Z is in units of $mm^6 m^{-3}$. However, different precipitation particles can vary greatly in terms of their diameter. In fact, there are several orders of magnitude difference between the Z from water vapor that composes a cloud at around $50\mu m$ versus the Z from raindrops of precipitating clouds at around $5mm$. Hail can produce Z values on the order of 10^7. This results in a wide dynamic range of the observed reflectivities. For this reason, Z is commonly expressed using a logarithmic transformation of $10log_{10}(Z)$. The units then becomes decibels relative to $1\ mm^6 m^{-3}$ which corresponds to $0\ dBZ$ [39, 40]. This simplification clarifies interpretation of visual and graphical depictions of Z. By rearranging Equation 3.6, Z can be solved for in terms of the reflectivity η. Substituting Equation 3.5 for η and then converting the summation into the integral of the drop-size distribution $N(D)$ and the back scattering cross section σ allows for Z to be expressed as in 3.7 [40]:

$$\begin{aligned} Z_{e_{h,v}}(dBZ) &= \frac{\lambda^4}{\pi^5 |K_w|^2}\eta \\ &= \frac{\lambda^4}{\pi^5 |K_w|^2}\sum N\sigma_{h,v} \\ &= \frac{\lambda^4}{\pi^5 |K_w|^2}\int \sigma_{h,v}(D)N(D)dD \end{aligned} \tag{3.7}$$

where σ denotes the back scattering radar cross section and the subscripts h and v denote the horizontal and vertical polarization, respectively, for a dual-polarization radar. Since the type of hydrometeors within the target resolution volume is generally an unknown variable, the K constant used is conventionally set to the dielectric constant for water K_w. Thus, in the field of radar meteorology, it is standard to consider the reflectivity factor Z to be the equivalent reflectivity factor Z_e that utilizes K_w. Z_e is a fundamental radar moment that is commonly plotted for visualizing the intensity of the weather event of interest. This is the radar moment used to make the radar plots utilized in the input LR and HR image datasets throughout this thesis.

3.4 Super-Resolution in Weather Radar

HR weather radar data is crucial for providing longer lead times by increasing the accuracy of forecast predictions and allowing for small scale storm features to be perceived and identified quicker and with increased reliability. This enhanced analysis better assures preparedness for intense storm systems and other severe weather events. Reservoir management, flood control, sewer-stormwater agencies and critical emergency response operations are able to respond to severe weather events and natural disasters faster and with more confidence. HR weather radar data can also provide insight into the dynamics of a weather event's development by enabling researchers to better observe and study the microphysical processes occurring within a storm system. This would help weather radar researchers better understand how different types of weather events progress in terms of their intensity, movement, speed, structure, and composition. Enhancing the understanding of weather events enables researchers to make more informed predictions and develop more suitable models and visualizations for assisting with storm warnings and tracking. Overall, HR weather radar data allows for the populace and emergency response agencies to be better prepared for severe weather events and natural disasters which, in turn, mitigates damage to property and, ultimately, saves lives.

Many of the considerations for collecting HR radar scans are typically approached through balancing and engineering the trade-offs between the characteristics of the radar design itself. For

example, Equation 3.1 shows that increasing the diameter of the radar antenna can effectively decrease the beamwidth, increasing the scan resolution. However, the extent of increasing the antenna diameter is restricted due to practical constraints such as cost considerations and physical functionality limitations. The radial range and azimuth sampling resolutions can both be enhanced by configuring the physical parameters of the radar. The Next Generation Weather Radar (NEXRAD) system is a radar network operated by the National Weather Service (NWS), the Federal Aviation Administration (FAA), and the U.S. Air Force. By employing techniques such as selective data windowing, radial recombination and range oversampling, the NEXRAD network's resolution was improved from a 1 km-by-1 deg polar grid to a 250 m-by-0.5 deg grid. The improved processing and physical configuration allowed the NEXRAD radars to collect what was termed as super-resolution data [41]. SR weather radar data has already been found useful in better characterizing the signatures from more dynamic, turbulent weather events such as tornado vortices [42, 43].

While radar design is a vital factor in collecting HR radar scans, there are many advantages to employing the use of DL SR models to generate SR radar scans from LR radar scans. Having a DL SR model embedded within the operational function of a weather radar system would enable the radar to conduct faster scans using LR radar scan strategies by increasing the antenna scan rate θ_s without sacrificing data quality. In turn, this would allow the radar to collect more data in a shorter amount of time for all weather events of interest. Since the LR radar scans collected would be post-processed into SR radar scans using the DL SR model, the overall data quality of a HR radar scan would be maintained while operating with the advantages of faster scan rates and increased data collection from LR radar scan strategies. Overall, weather radar systems would be able to collect more data at a faster rate with HR radar data quality when utilizing DL SR models through digital signal/image processing. Furthermore, the DL SR techniques could be quickly applied to existing radar networks. All that would be required is a dataset of HR radar scans, which should be readily available, in order to develop the DL SR model. This is much more accessible than redesigning, purchasing parts and upgrading the components of the existing radar system. Increasing the ac-

cessibility of HR weather radar data helps society as a whole. The importance of accurate, fast, and reliable weather information is only increasing as people become more susceptible to natural hazards whose impacts are being exacerbated due to climate change. Furthermore, more abundant use of DL SR models within weather radar field allows climatologists, meteorologists, atmospheric scientists and other researchers who develop weather pattern simulation models and atmospheric physics equations to have increased access to HR radar data. Which, in turn, would allow them to develop better informed and more accurate models/equations. In addition, the increased scan rate of the LR radar scan strategy allows for more data to be collected in a shorter amount of time. This is especially important for researching turbulent weather events – such as tornadoes and hurricanes – that progress and change rapidly. The faster scan rate would help give deeper insight into the nature of these highly dynamic natural phenomena including the conditions needed for them to begin, the progress of their development, distinguishing between types, and predicting their movement. All of this will help to enhance the accuracy and confidence of forecasts given to emergency response agencies and any people that may be affected by these natural disasters.

Chapter 4

Dataset

The dataset utilized serves as the backbone when developing a SRGAN model. Compiling the dataset is vital to the operation and end performance of any machine learning model, being a direct factor in the model's ability to perform the desired task. Datasets are the foundation from which the DL SR model learns how to develop the super-resolved images from multiple repetitions of training on a subset of the dataset itself. The entire developmental process of the SRGAN model; from the initial hyperparameter optimization period, to the experimental training and the final evaluation, these all depend on the images within the dataset.

Takano et al. conducted a study to determine how the SRGAN model learns and generates its SR images as well as how the training dataset used affects the performance of a SRGAN model [44]. The study carried out experiments in which the training dataset was altered to test how the SRGAN model learns and generates its SR images. It was found that a SRGAN model trained on colored images produced SR images in color even when supplied with grayscale image inputs. Another experiment tested the trained SRGAN models on edge outline images. It was found that the SRGAN models attempted to re-create and color the full target HR image even with only edge outline images as the input. From these, it was concluded that the SRGAN model fundamentally learns the input images' color, shape, and texture during training. Then, the SRGAN model attempts to generate or redraw the input LR image in the likeness of the target HR image instead of simply sharpening the image edges or features [44]. In order to determine how the training dataset affects the performance of the model, three datasets were used that comprised of different types of images, one containing people's faces, another containing dinging rooms, and another containing buildings. Each of these datasets were used to train a SRGAN model, with each model having the same architecture, and then each trained SRGAN model was evaluated on each of the three testing datasets. The evaluations showed that the SRGAN model had a higher performance during evaluation with the same dataset that it was trained on [44]. This study demonstrates

the importance of the training dataset utilized as it is vital to the end performance of the SRGAN model being tested.

Chapter 4 details how the dataset was formed for this research's experiments. Chapter 4.1 presents information on the weather radar that collected the data used throughout the rest of this thesis study. Important resolution parameters for the weather radar data are shown and defined. Chapter 4.2 discusses the campaign during which the weather radar data was collected in order to give further context behind the data being used throughout this research. Chapter 4.3 presents an in-depth explanation into the pre-processing techniques used to prepare the data for developing the experimental SRGAN models. The pre-processing used to generate the LR datasets, including the specifics on the physically representative downsampling method. Chapter 4.4 describes all aspects of the SR dataset used to develop the experimental SRGAN models, including how the dataset is subset for the SRGAN model development as well as the specifications for the RHI and PPI datasets.

4.1 The CSU-CHIVO Radar

This thesis utilizes data collected by the Colorado State University C-band Hydrological Instrument for Volumetric Observation (CSU-CHIVO) radar. The radar data contains both PPI and RHI radar scans. The CSU-CHIVO Radar is a dual-polarization, Doppler weather radar. It operates within the C-band frequency range of 4 - 8 GHz. The positioner that it uses to move the radar's antenna and reflector in the desired scan strategy is a semi-yoke elevation over azimuth type with a brushless alternating current servo motor. It's antenna is a center-fed parabolic reflector with a beam width of $0.95°$. For receiving and processing the return signal, the CSU-CHIVO radar utilizes a Sigmet Digital Receiver and a RVP900 Signal Processor [45]. With these, it computes radar moments such as the radar reflectivity factor Z as described in Chapter 3.3. Figure 4.1 depicts the CSU-CHIVO radar installation location site during the field campaign studied throughout this thesis.

Figure 4.1: CSU-CHIVO radar deployed near Alta Gracia - Argentina during the RELAMPAGO campaign [45]

The resolution of the radar data sampling volume for the CSU-CHIVO radar can be defined with the radial range, angle sampling (beamwidth), and the cross range resolution parameters. Table 4.1 outlines these radar beam resolution parameters for the CSU-CHIVO radar. The values within Table 4.1 are primarily the same except for the radial range resolution parameters, which are 150 m for the RHI scan strategy and 200 m for the PPI scan strategy.

The radial range parameter ΔR defines the distance between the beginning and end of a single range gate within the radar scan. It defines the radar's ability to distinguish between two or more targets along the same angle but at different ranges away from the radar. Typically, the range resolution can be calculated using the pulse width τ of the transmitted signal as $\Delta R = \frac{c\tau}{2}$. The angle sampling resolution $\Delta\Theta$ is the same as the beamwidth parameter of the radar. These are

Table 4.1: CSU-CHIVO Radar Resolution Specifications

Parameter	RHI Scan	PPI Scan
Radial Range Resolution	150 m	200 m
Angle Sampling Resolution (Beamwidth)	0.95°	0.95°
Cross Range Resolution	≈ 165 m at 10 km ≈ 2.3 km at 140 km	≈ 165 m at 10 km ≈ 2.3 km at 140 km

both used to describe the angle that defines the sampling volume of the radar beam. Spatially, this can be translated into the cross range resolution $R(\Theta)$ which is defined as the distance between each side of the radar's beam. This resolution parameter is both opposite to the beamwidth angle and orthogonal to the radial range. Similar to a flashlight's beam, the radar's beam spreads the further it travels. So, $R(\Theta)$ is a function of the range and beamwidth. This can be expressed as $R(\Theta) = r_{abs} * tan(\Delta\Theta)$ where r_{abs} is the absolute range away from the radar. Figure 4.2 portrays a diagram of these resolution parameters with specifications provided from the CSU-CHIVO radar.

4.2 The RELAMPAGO Campaign

The dataset utilized for this thesis is derived from data collected by the CSU-CHIVO radar during the three months – from November 2018 to January 2019 – that it conducted operations for the international Remote sensing of Electrication, Lightning, and Meso-scale/micro-scale Processes with Adaptive Ground Observations (RELAMPAGO) field campaign. RELAMPAGO's mission focused on observing and analyzing data in one of the most thunderstorm active regions of the world. Experimentation focused on five primary areas of study: convective initiation including studying differing intensities of convection events and pre-convective environmental conditions, severe weather such as supercells and hail events, upscale growth analysis with an emphasis on analyzing how the terrain and cold pools impact the development of new convective updrafts as well as the processes that make deep convection intensify, lightning analysis specifically determining their characteristic features and how they may differ within different convective systems,

and hydrometeorology efforts including studying the relationship between surface and atmospheric processes and data collection via an acoustic Doppler current profiler [46, 47].

The full campaign lasted from June 2018 until April 2019 in central Argentina's Cordoba region. During the time it was deployed, the CSU-CHIVO radar was able to collect crucial radar data during an intense observing period (IOP) which occurred from November 1, 2018 through December 15, 2018 [48]. The CSU-CHIVO radar was stationed near the Andes mountain region for weather observation and data collection [49]. The region of observation for the RELAMPAGO campaign was strategically selected. Supercells, mesoscale convective systems (MCS), intense rainfall and heavy hail storms are particularly prominent within the Sierras de Córdoba. The complex terrain environment promotes the formation of severe weather events making them more frequent and more intense. Some of the tallest storms studied have been observed within the RELAMPAGO campaign's domain. One of the largest, hail-producing storms documented during

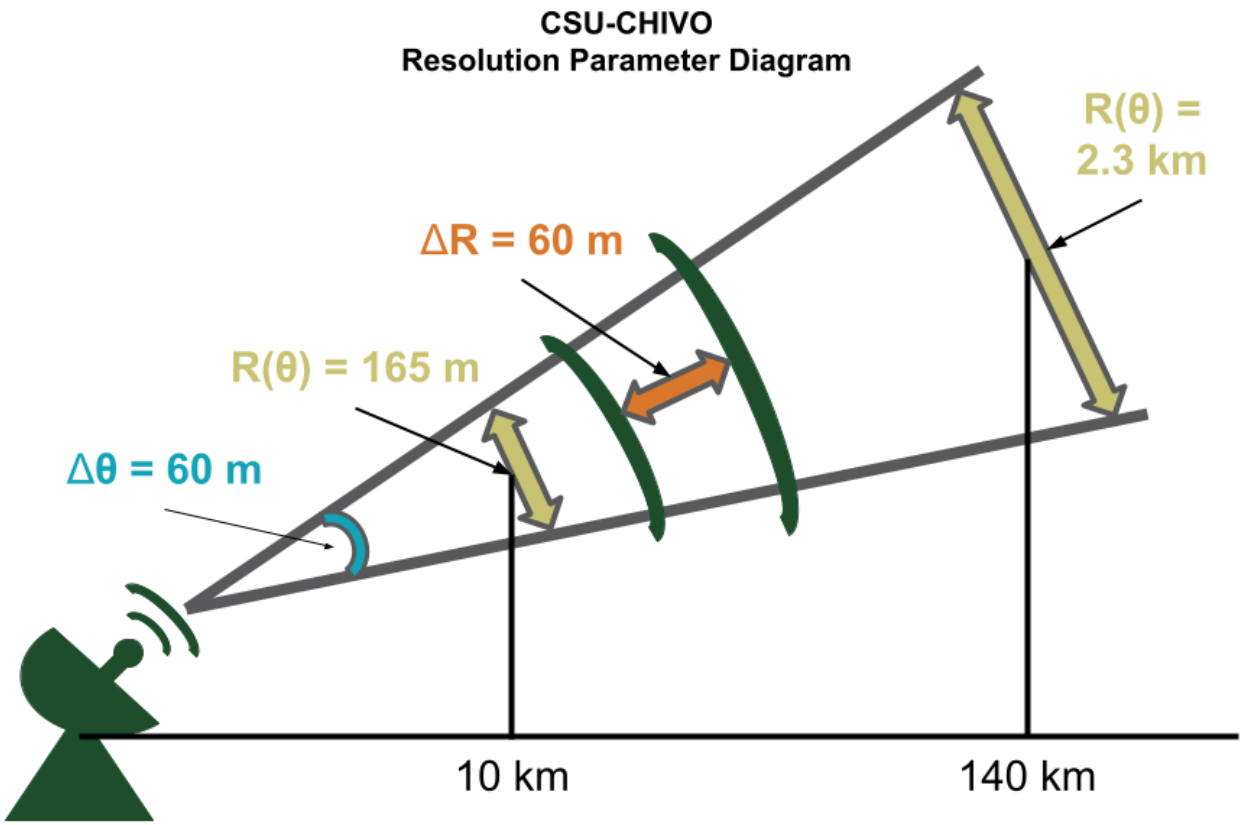

Figure 4.2: CSU-CHIVO Resolution Parameter Diagram

the RELAMPAGO campaign – occurring on January 25, 2019 within the CSU-CHIVO radar's range – reached up into the stratosphere with a cloud top of over 20 km [45, 47, 49].

This data was specifically chosen because both PPI and RHI data was collected within a similar time frame for multiple intense rainfall storm events. Having both horizontal and vertical profiles available within the dataset is quite valuable. This allowed for experimentation of the SRGAN model's ability to generate super-resolution images to be conducted on both of these prominent radar scan types and compare the performances. Furthermore, since the RELAMPAGO domain covered one of the most active storm regions of the world, the data itself is quite significant. The dataset includes supercell, MCS, stratiform storm systems, heavy rainfall, and hail storm radar data. Having many different types of storm events allows for more comprehensive research to be conducted to determine the efficacy of utilizing the SRGAN model to super-resolve weather radar scan images.

4.3 Data Pre-Processing

Weather radar data goes through multiple levels of processing in order to derive useful geophysical products and enhance the data quality. Initially, when the weather radar data is first collected by the radar, it is considered to be raw, binary data. This is converted into I/Q data, as described in Chapter 3.1. This is what is considered to be Level I data in the weather radar community. The next main step in the data processing is when the data is converted into the geophysical variables, i.e. radar moments. These are used for visualizing and analyzing the data as it transforms the raw data (with values in the form of voltages and currents) into physically interpretable information about the weather radar volume of interest. This is considered to be Level II data and it is typically the primary output of a weather radar system. However, before this data is utilized to train the experimental SRGAN models, it must be pre-processed. For this thesis, pre-processing involves all of the data transformations conducted before being input into the experimental SRGAN models for training. This involves quality correcting the data, mitigating any clutter present in the data,

converting the data into a more usable format by plotting it and then generating the dataset of LR radar scans.

As the EM pulse transmitted by the radar propagates through the air, its energy weakens as it collides with atmospheric particles, scatters and travels further away from the radar. Because of this, the energy signatures received by the radar from targets at farther ranges are attenuated. This means that, if the data was to be used in the same condition as it was output directly from the radar, the Z values, for example, would be underestimated at farther ranges for every radar scan. Thus, before being utilized, it is beneficial to pre-process the radar data with attenuation correction. This thesis utilized the attenuation correction algorithm called the Dual-Polarization Radar Operational Processing System 2.0 (DROPS2.0) developed by Chen et al. at Colorado State University [50]. DROPS2.0 conducts the data quality control by estimating the specific differential phase K_{dp}, another derived radar moment, in order to conduct attenuation correction and remove both ground clutter and other non-meteorological echoes.

After being attenuation corrected, the Z variable of the radar data is put through a threshold. Any values outside of a typical range for Z, from 0 dBZ to 100 dBZ with some extra padding on the top-end of the threshold, are excluded from the radar data matrix before plotting. This helps to ensure that any ground clutter echoes are mitigated as well as accounting for any invalid reflectivity values. With this, the radar data's Z variable is ready to be plotted. The weather radar data was plotted so that the study closely matched the experimental set up compared to the SRGAN source material from [14] instead of testing the SRGAN model on the data itself. Furthermore, the SRGAN model is designed for square images whose length and width parameters are equivalent, which is rarely the case for the polar-coordinate-oriented radar moment data. In order to plot the weather radar data, the Phyton API Pyart was utilized [51]. After plotting, the resulting images were then cropped to exclude the title, axes, and colorbar information so that the weather radar data visualization information was the only input from which the experimental SRGAN models learned. The resulting plots were then resized to be square images. The images within the HR dataset have an image size of 256x256x3. The images within the LR dataset have an image size of

either 128x128x3 or 64x64x3 depending on whether the resolution scale factor used was x2 or x4, respectively.

4.3.1 LR Dataset Pre-Processing

Training a GAN model typically uses a randomly generated, Gaussian noise image as the LR image input for the generator model [16]. However, for SR methods in particular, a more complex image information mapping is developed between the LR and HR domains through the use of LR and HR image pairs [14]. The paired LR and HR image datasets must be prepared before training can commence. For SRGAN models, the HR images are considered to be the ground-truth images from which the LR images are derived. In the case of this thesis, the HR images were created using the Level II data, the weather radar moments, processed by the weather radar. The LR dataset is then directly generated from the HR images, thus making the HR and LR image pairs that are input into the SRGAN during training.

Conventionally, DL SR studies in literature generate the LR dataset by utilizing a Gaussian or bicubic interpolation kernel, typically of a x2, x3, or x4 size, as shown in Table 2.1, and convolving it with the HR images in order to build the LR image dataset. The kernel filter is convolved with each image in the HR dataset. This means that each of the kernel's elements are multiplied with the elements of a subset of the radar data of a corresponding size to the kernel filter. These are then added together to form a sum of products that interpolates, i.e. numerically estimates a new data point, the corresponding pixel in the downsampled, LR image. The kernel filter then moves through the HR image until the new, downscaled LR image is completely generated. The amount by which the HR image is reduced can be controlled by the kernel size. Typically, each of the LR images are downscaled to a half, a third, or a quarter of the image size of the HR dataset. These fractions correspond to what is called the resolution scaling of the LR to the HR image datasets, which are x2, x3, or x4, respectively. The resolution scaling refers to the factor that is necessary to upscale from the LR image to the HR image size.

While interpolation is considered to be the standard downsampling method used for DL SR techniques, the LR image that it produces does not have the same characteristic features of a LR radar scan. An interpolated LR image appears to be more burred than its HR counterpart whereas a LR radar scan appears to have a different structural composition compared to its HR counterpart. When a LR radar scan is collected, the radar's antenna scan rate is increased. This increases an individual scan's sampling volume by decreasing the number of beams in the LR scan, as shown in Equation 3.4. When compared, a LR weather radar scan would appear to be more block-like and have larger beams than its corresponding HR weather radar scan. Since the interpolation kernel method of downsampling cannot simulate the structural features of a physical LR scan, a new method of downsampling was developed and utilized for this purpose. This thesis utilizes a radar data processing technique in order to generate a LR dataset that would better resemble the characteristic features of a LR weather radar scan in addition to the interpolated LR dataset. The technique that downsamples the HR weather radar data into a semi-realistic LR weather radar scan is referred to as the physically representative downsampling method.

4.3.2 Physically Representative Downsampling Method

Chapter 3 describes how a radar collects data by transmitting EM pulses while moving in either its elevation or azimuth angles. Chapter 3.4 provides further details about the difference between a HR and LR radar scan. Equation 3.4 shows that as the scanning rate of the antenna increases, the number of pulses decreases by the same factor. For example, doubling the scanning rate of the antenna will effectively half the number of pulses that are taken during the scan, resulting in LR weather radar data being collected. Because this thesis' scope is specified for super-resolving images of weather radar data, the idea of developing a new method for processing the LR dataset, in a similar way that reflects how LR weather radar scan data is collected, was explored. This will be referred to as the physically representative downsampling method. Using this method, the LR dataset will more closely resemble actual LR weather radar scans that would have been taken

during data collection. This is done by pre-processing the radar data to have fewer angles before plotting it.

The physically representative downsampling method was employed during the pre-processing stage of dataset creation. When the radar data is loaded into the pre-processing script, it is read as a two dimensional matrix. One dimension of the matrix represents the range of the radar while the other represents either the elevation or the azimuth angles, depending on whether the input data used was RHI or PPI, respectively. See Figure 4.3 below for an example of the input radar data matrix . In the figure, each column of the radar data matrix corresponds to a range gate of the radar scan and each row of the radar data matrix corresponds to an angle bin. A range gate defines a distance that signifies how far away a single sampling area is from the radar. An angle bin defines the scanning angle of the radar in either azimuth or elevation for PPI and RHI radar scans, respectively. Figure 4.3 depicts this interpretation of the radar data matrix.

An image of the HR radar scan is produced when the entire input radar data matrix is processed and plotted. At this point, a LR image can be created by using a bicubic interpolation kernel on the HR image. However, a LR radar image can be made without needing to make the HR image first, by pre-processing the HR weather radar data. Instead of processing and plotting the entire radar data matrix, a number of the angle rows, as depicted in Figure 4.3, can be removed from the matrix. This effectively decreases the number of radar beams in the data matrix. Once the radar data matrix is downsampled, processed and plotted, a LR image can be generated that not only pairs with the HR image but, also, exhibits the physical properties of an actual LR radar scan. This is the physically representative downsampling method that was exercised during this thesis. The level of downsampling can be controlled depending on how many angles are removed. If every other angle row is removed from the matrix, a LR image half the size of the HR image can be created. Likewise, if the last three rows of every four rows is removed, then, a LR image a quarter of the size of the HR image can be made. The diagram in Figure 4.3 illustrates how the radar data matrices are transformed in the physically representative downsampling method of pre-processing the Level II weather radar data to build the LR datasets.

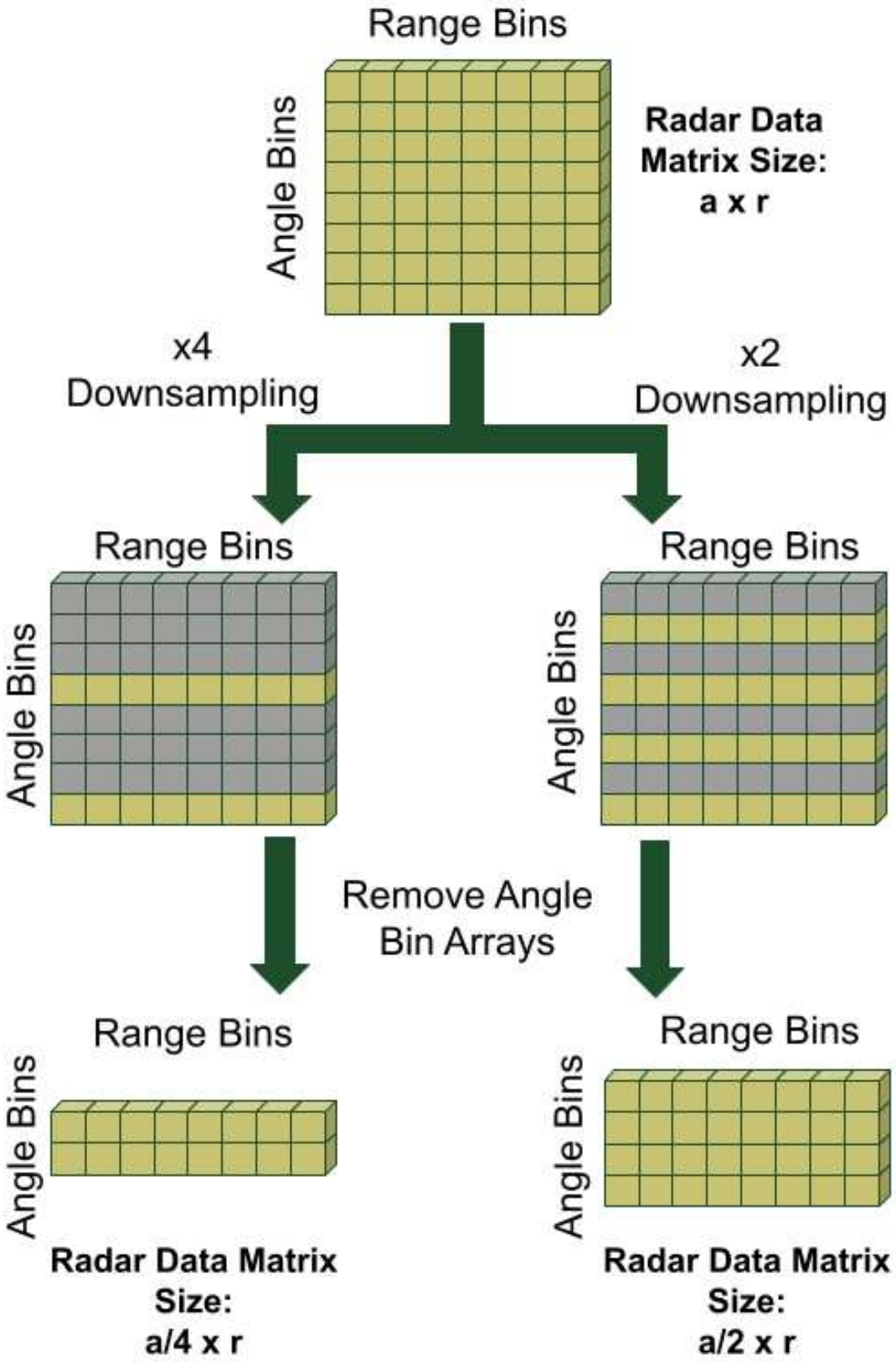

Figure 4.3: Radar Data Matrix Example

As described in Chapter 3.2, a RHI scan is conducted as the radar changes its elevation angle while a PPI scan is conducted as the radar changes its azimuth angle. In order to create a LR radar scan from a HR radar scan, the radar data matrix must reduce the number of angle rays that it contains. For example, the RHI x2 physically representative LR dataset was generated by removing every other elevation angle row from the data matrix. This reduces the radar data matrix by a factor of two along the elevation angle dimension and effectively simulates a weather radar scan that would have been taken at twice the antenna scanning rate, as according to Equation 3.4. Likewise, the RHI x4 physically representative LR dataset was generated by only keeping the first ray out of every four rays. These LR images are illustrated in Figure 4.4 below. It depicts side-by-side comparisons of the LR images that are generated by the bicubic interpolation and the physically representative downsampling methods for a given input HR RHI radar scan.

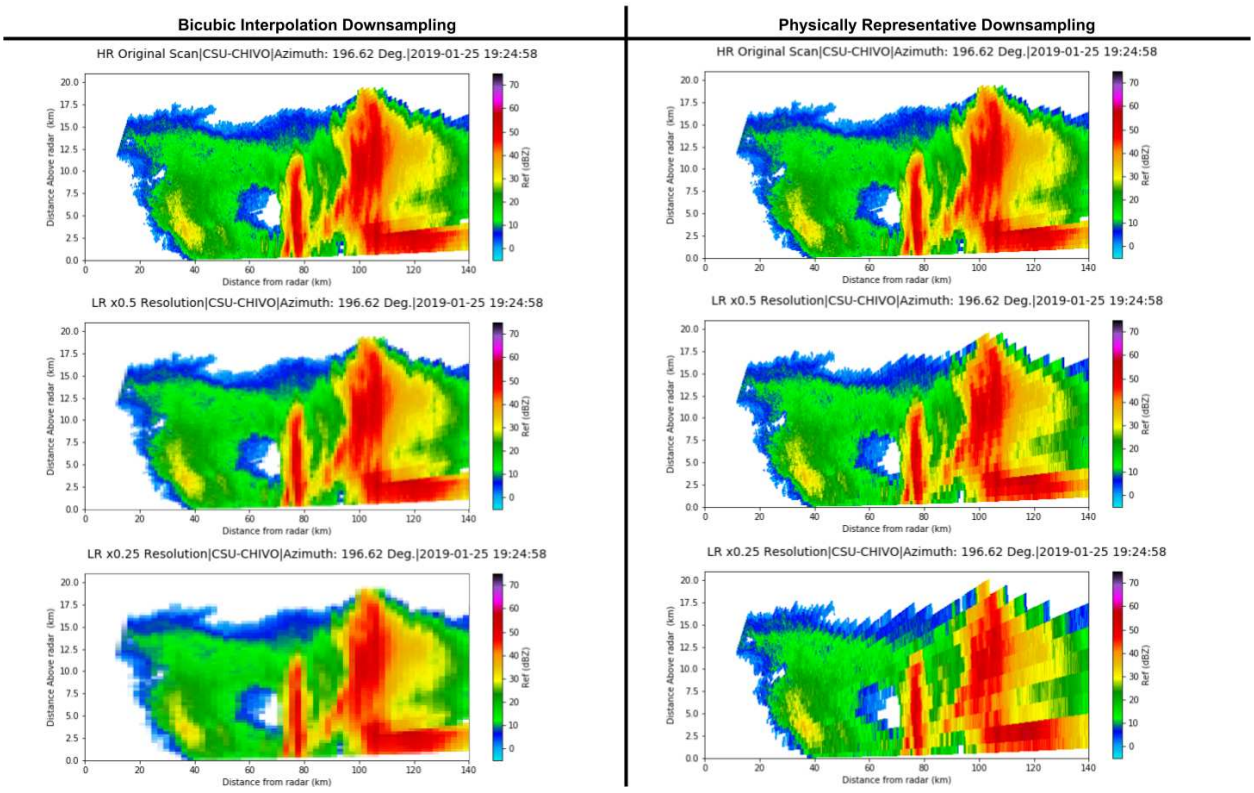

Figure 4.4: RHI Example for Downsampling Methods

As can be seen from Figure 4.4, the bicubic interpolation downscaling technique makes the image appear more blurry, distorting the high-frequency details within the radar scan plot. The characteristics of the radar scan image do not change much otherwise. On the other hand, in the physically representative downsampling examples, the physical properties of a LR radar scan are distinctly exhibited. The decreased number of beams results in wider, more box-like beams being plotted. This shows how the sampling volume has increased as each range gate increases in size, increasing its effective pixel area coverage, and, thus, decreasing its overall resolution.

The structural features portrayed in the physically representative downsampling LR images are much more indicative of actual LR radar scans. These characteristics are also seen in physically representative LR PPI scans as well. The PPI x2 physically representative LR dataset was generated by removing every other azimuth angle row from the weather radar data matrix. This reduces the radar data matrix by a factor of two along the azimuth angle dimension, which is essentially the same as the elevation angle for the RHI radar scan in terms of function, and effectively simulates a radar scan that would have been taken at twice the antenna scanning rate, as according to Equation 3.4. Likewise, the PPI x4 physically representative LR dataset was generated by only keeping the first ray out of every four rays. This is illustrated in Figure 4.5 below. It illustrates side-by-side comparisons of the LR images that are generated by both downsampling methods for an input HR PPI radar scan.

The PPI scans also portray properties similar to those of an actual LR radar scan, even though these characteristics may be more difficult to perceive in the PPI examples. This is because the full range of the scan for the RHI plots is 0 - 140 km whereas the full range of the scan for the PPI plots is -150 - 150 km for a total of 300 km. Which means that the detailed features in the PPI plots are almost twice as small as those in the RHI plots. Nevertheless, if closely observed, the PPI plots in Figure 4.5 also depict the wider beams indicative of actual LR radar scans in the physically representative downsampling example plots. This is particularly evident at farther ranges.

There are some considerations when using this physically representative pre-processing method that need to be accounted for. During an operational LR radar scan pattern, the number of beams

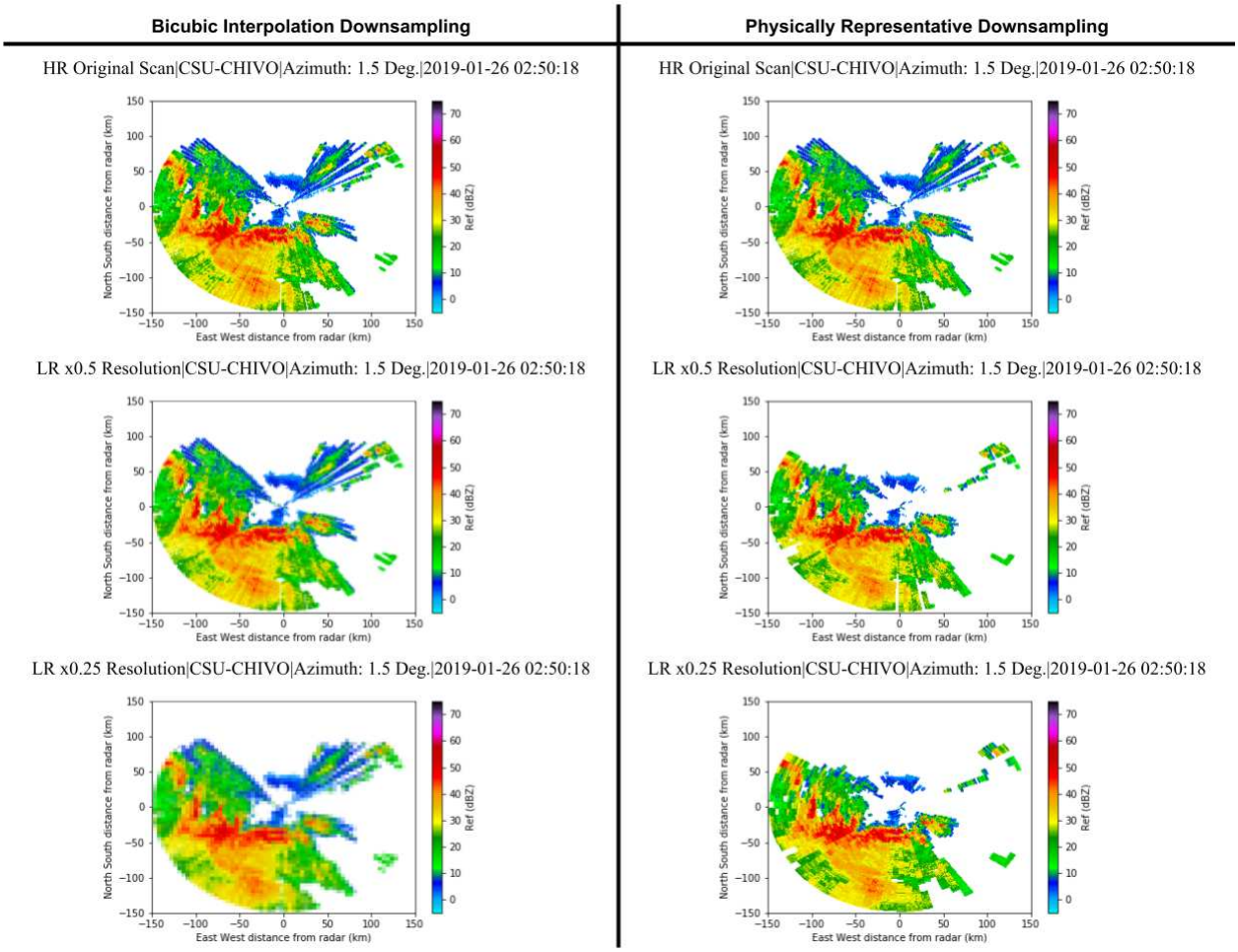

Figure 4.5: PPI Example for Downsampling Methods

is reduced as the scanning rate is increased; however, the observed range itself is unaffected. This method tries to follow the actual LR radar scan as closely as possible by only changing the data in one dimension, the angle dimension in the radar data matrix. The number of angles is changed by removing rays from the HR data in order to create the LR radar scan data and the range is unaffected. However, by the nature of SRGANs since their inception, the images that are input into a SRGAN during training must be square images, meaning that the images must have the same dimensions for the height and width. This would equate to not only reducing the number of angles within the radar data matrix, but reducing the number of range gates as well. Although this follows for creating a LR dataset for input into the SRGAN, it does not adhere to the physical sciences of radar data collection. That is why, instead of changing the number of range gates within the radar

data matrix to ensure that the LR images are square, it was deemed necessary that the physically representative LR dataset was also resized using a bicubic interpolation kernel. Doing so allows the physical properties of an operational LR radar scan to be maintained during the physically representative pre-processing of the HR images while also having the images in the correct square format for input into the SRGAN models during training. This decision creates further considerations as well. Resizing the radar scan images when creating the datasets changes the initial existent shape of the data. Also, the resizing was done with a bicubic interpolation kernel as well meaning that the end result of the physically representative LR dataset will not be independent of the other LR pre-processing method. The physically representative LR pre-processing method leads to bigger range gates and thus greater effective pixel area coverage. Meaning one range gate in the LR now accounts for more pixels within the image than in the HR image.

Both of these illustrations, Figures 4.4 and 4.5, serve to visually demonstrate support for why the physically representative downsampling method was developed and employed in this thesis study. The difference between these two downsampling methods is quite apparent. Removing the rays within the weather radar data during pre-processing also creates an inherent complication for, in doing so, it causes the problem to become more complex than a standard SR problem. This is primarily due to the LR images no longer portraying the same features as portrayed in the HR image pair. However, both of these downsampling methods are important to research as the bicubic interpolation downsampling technique is utilized in prominent SRGAN research papers while the physically representative downsampling approach better simulates LR inputs specifically for the scope of using SRGAN to super-resolve weather radar scans.

Exploring this downscaling method that is more applicable for radar scans in particular will be a valuable contribution of this thesis. Since the use of machine learning models within the radar community is still uncommon, the results of this experiment will show if SRGANs can be utilized with LR images that more closely resemble the physical properties of LR radar scan data. This thesis aims to show the efficacy of utilizing SRGAN with LR radar scan images to generate physically realistic SR radar scans in the same resolution as a HR radar scan. The results of

this method could encourage further research into utilizing the radar data itself for input into the SRGAN model and support the possibility of incorporating LR scan strategies as a viable method for collecting more data at a faster rate on any precipitation event going forward. This would increase the efficiency of radar data collection as a whole. A faster scanning rate would allow radars to collect more data in a smaller amount of time. This is crucial because in this day and age more data allows for access to a wider pool of knowledge which will expand the horizons of radar research endeavors.

4.4 The Super-Resolution Dataset

Many of the DL SR models in literature utilize standard, benchmark image datasets for training and testing. Most of these datasets comprises images from the natural image manifold including people, animals, landscape scenery, and buildings. These are commonly referred to as "natural images". Some of the most prominent SR datasets used in literature are non-specific and primarily consist of one or more types of natural images. The Set5 dataset contains five images commonly used in literature for portraying the SR models' results, namely: "baby", "bird", "butterfly", "head", and "woman" [52]. The Set14 dataset houses fourteen images that are commonly used throughout SR literature as they are larger and considered to be more diverse than Set5 [53]. The Urban100 dataset contains 100 images of urban environments depicting scenes with distinct structures with repetitive patterns [54]. The ImageNet dataset is a very large natural image dataset consisting of 3.2 million annotated images at the time of the citation's publication [24]. The Diverse 2K resolution (DIV2K) dataset with 1000 RGB color natural images collected from the internet [55]. Finally, the Manga109 dataset has been utilized with increasing frequency recently. It contains 109 manga – Japanese comic books – with 21,142 pages of images drawn by professional manga artists called "mangaka" [56].

The datasets that are considered standard in literature tend to be made up of a large number of images, typically on the order of tens to hundreds of thousands of images. These image datasets also consist of many different types of images from the natural image manifold. The size of the

datasets utilized in this thesis is notably smaller, on the order of thousands of images, due to the computational resources available. The types of images that comprise the datasets that are utilized throughout this thesis are also homogeneous, i.e. consisting of one type of image, namely, RHI or PPI weather radar scans, to focus the experimental SRGAN models' learning on solely the radar scans themselves. Although many SRGAN papers in literature focus on developing SRGAN models for general SR tasks, [44] found that SRGAN models have an increased performance when trained on a homogeneous dataset consisting of a single type of image when compared to being trained on an arbitrary dataset that contains multiple types of images. It should be considered that the original SRGAN model from [14] that this thesis basis its SRGAN models on was not designed to perform SR on weather radar scans, specifically. It's architecture was designed for super-resolving many different types of images within the natural image manifold from a heterogeneous dataset. Nevertheless, the SRGAN is a prevalent SR model used in literature that has been utilized for many applications that train on a homogeneous dataset.

Currently, the literature does not contain many studies examining the use of DL models for super-resolving weather radar scans in general. These factors, coupled with the limitations to the available computing resources, were taken into consideration when deciding how to build the dataset. It was decided that the dataset should be constrained to contain images solely within the weather radar regime. This way, the experimental SRGAN models would have a curated environment for enhanced SR performance. Once the CSU-CHIVO weather radar data from the RELAMPAGO campaign was obtained, the weather radar scans were split up by their radar scan type: RHI and PPI, as individual and separate datasets. RHI and PPI scans have distinct image structures from one another due to the physical means by which the radar collects the data, as described in Chapter 3.2. By narrowing down the scope of the types of images within the dataset, the experimental SRGAN models that are tested should have the best environment to train under. Thus, the experimental SRGAN models will have a better chance at achieving the ability to generate physically realistic, super-resolved weather radar scans. This is especially important since this thesis uses a smaller dataset than most other studies in literature.

4.4.1 The SRGAN Dataset Subsets

This thesis utilizes two primary datasets, one for each type of main radar scan conducted during the RELAMPAGO campaign, namely RHI and PPI. Both of these datasets contain radar scans for a variety of clear air, light rain, and heavy rain events. The HR datasets are processed directly from the radar moments output from the radar system. During the pre-processing stage, the LR datasets are created. The LR datasets are created with either x2 or x4 resolution scaling factors as these were the most prominently used in the existing literature. In addition, two different image processing downsampling methods were used to create the LR images in order to study the effectiveness of these downsampling methods. One used the standard bicubic interpolation kernel, consistent with the rest of the literature, and the other used the physically representative downsampling method. The physically representative downsampling method that this thesis proposes will generate LR images that more closely resemble the characteristics of the LR scans collected by a radar, as described in 4.3.2, in order to generate the LR dataset. In order to fully investigate each of these experiments, individual SRGAN models were trained for each combination of the experimental dataset variables described above. After compiling the dataset, pre-processing the radar moment data, and creating the plots, this resulted in a total of eight sets of HR and LR datasets being generated and eight experimental SRGAN models to be tested. These will be referred to throughout this thesis as: RHIx2_Interp, RHIx4_Interp, PPIx2_Interp, PPIx4_Interp, RHIx2_PhysRep, RHIx4_PhyRep, PPIx2_PhysRep, and PPIx4_PhysRep. The nomenclature for these defines the radar scan type, either "RHI" or "PPI", the resolution scaling factor used during the downsampling pre-processing, either "x2" or "x4", and the downsampling method used to generate the LR dataset during the pre-processing stage, either "Interp" or "PhysRep" for the bicubic interpolation and the physically representative downsampling method, respectively. The creation of all of these different datasets was necessary for thorough experimentation of the SRGAN model's capabilities. These will test if the SRGAN is able to perform well at generating super-resolved images for both radar scan types, RHI and PPI, for multiple levels of resolution scaling, x2 and x4, and, finally, for different methods of LR pre-processing techniques, bicubic interpolation and the physically

representative downsampling method. All of this experimentation will give further insight into the efficacy of utilizing the SRGAN model for conducting SR on weather radar scans.

Each dataset is then further broken up into the training, validation and testing datasets. Subsetting the original dataset into these three types of datasets is necessary during the development of the SRGAN. They perform three distinct roles and will be used at different points in the experimental SRGAN models' development. The validation dataset typically contains 10 - 20% of the initial dataset and is utilized during the hyperparameter optimization process. This helps optimize the performance of the SRGAN by testing different hyperparameters. More on this can be found in Chapter 5.3. The training dataset typically consists of 60 - 80% of the initial dataset. This dataset is the largest as it is used during the experimental SRGAN models' dynamic developmental learning phase. Multiple variations of the SRGAN models are trained for multiple iterations, i.e. epochs as referred to in the ML community, over the training dataset so that they can learn how to generate physically realistic super-resolved images. More on this can be found in Chapter 5.4. The testing dataset is used for the final evaluation when comparing the experimental SRGAN model variations. It typically consists of the last 10 - 20% of the initial dataset. These tests are compared against the baseline models and measure the level of success that the SRGAN models can achieve for super-resolving images. Separate datasets are used during each stage of the SRGAN's development so that they do not have influence on one another. This is especially important for decreasing the bias within the final evaluation. Each image that the SRGAN is tested on is completely unknown to the model during the final evaluation phase, ensuring that the quality of the performance measured is assured. This thesis split the training, validation and testing datasets into 60-20-20%, respectively, as described above. The images of the primary RHI and PPI datasets were shuffled and then randomly assigned to either the training, validation, or testing datasets. Table 4.2 lays out the size of each of the primary datasets that were used throughout this thesis.

Table 4.2: Dataset Sizes

Size of Dataset Type	RHI	PPI
Training Dataset (# of images)	3833	2217
Validation Dataset (# of images)	1278	739
Testing Dataset (# of images)	1278	739
Total Dataset (# of images)	6,389	3,695
Total Dataset Storage (MB)	242	61

4.4.2 The RHI Dataset

The RHI dataset consists of 6,389 images in total. Table 4.3 lists the dates and the number of scans collected during each date within the RELAMPAGO campaign for the RHI dataset. It specifies these for the training, validation and testing datasets individually. This information is consistent for both of the x2 and x4 resolution scaling factors as well as the interpolation and physically representative downsampling methods utilized during experimentation. The RHI training, validation and testing datasets were split in 60-20-20%, respectively, of the total scans within the RHI dataset. As shown in Table 4.2, the RHI training dataset consists of 3,833 RHI scans, the RHI validation dataset consists of 1,278 RHI scans, and the RHI testing dataset consists of 1,278 RHI scans. During this time, two RHI scan strategies were conducted. Both had a scan rate of $4^\circ/\text{sec}$. One RHI scan stratey consisted of scanning the radar between 0° and 45° in which either six RHI scans were collected with an azimuthal spacing of 6° for a total scan time of 1:28 in minutes:seconds, ten RHI scans were collected with an azimuthal spacing of 6° for a total scan time of 2:25, or six RHI scans were collected with an azimuthal spacing of 10° for a total scan time of 1:35. The other RHI scan strategy scanned the radar between 0° and 30° in which eighteen RHI scans were collected with an azimuthal spacing of 10° for a total scan time of 3:25.

Table 4.3: RHI Dataset Dates

Dataset	**RHI**			
	Dates	Number of Scans	Dates	Number of Scans
Training Dataset	11/10/2018	48	11/30/2018	482
	11/11/2018	22	12/01/2018	661
	11/12/2018	541	12/04/2018	194
	11/22/2018	312	12/14/2018	185
	11/26/2018	50	01/25/2019	377
	11/27/2018	432	01/26/2019	529
Validation Dataset	11/10/2018	13	11/30/2018	149
	11/11/2018	9	12/01/2018	221
	11/12/2018	208	12/04/2018	75
	11/22/2018	86	12/14/2018	69
	11/26/2018	13	01/25/2019	109
	11/27/2018	151	01/26/2019	175
Testing Dataset	11/10/2018	15	11/30/2018	143
	11/11/2018	8	12/01/2018	220
	11/12/2018	176	12/04/2018	81
	11/22/2018	94	12/14/2018	66
	11/26/2018	11	01/25/2019	128
	11/27/2018	153	01/26/2019	183

4.4.3 The PPI Dataset

The PPI dataset consists of 3,695 images in total. Table 4.4 lists the dates and the number of weather radar scans collected during each date within the RELAMPAGO campaign for the PPI dataset. It specifies these for the training, validation and testing datasets individually. This information is consistent for both the x2 and x4 resolution scaling factors as well as the interpolation and physically representative downsampling methods utilized during experimentation. The PPI

Table 4.4: PPI Dataset Dates

Dataset	**PPI x2 and x4**									
	Dates	Number of Scans	Dates	Number of Scans	Dates	Number of Scans	Dates	Number of Scans	Dates	Number of Scans
Training Dataset	11/10/2018	19	11/29/2018	55	12/06/2018	87	12/20/2018	110	01/07/2019	34
	11/11/2018	153	11/30/2018	106	12/12/2018	26	12/21/2018	30	01/09/2019	93
	11/12/2018	50	12/01/2018	139	12/13/2018	8	12/27/2018	68	01/13/2019	169
	11/22/2018	56	12/02/2018	123	12/14/2018	50	12/28/2018	92	01/24/2019	49
	11/26/2018	13	12/04/2018	51	12/18/2018	132	12/29/2018	7	01/25/2019	49
	11/27/2018	121	12/05/2018	108	12/19/2018	102	01/03/2019	71	01/26/2019	46
Validation Dataset	11/10/2018	4	11/29/2018	20	12/06/2018	29	12/20/2018	41	01/07/2019	13
	11/11/2018	53	11/30/2018	31	12/12/2018	7	12/21/2018	11	01/09/2019	28
	11/12/2018	9	12/01/2018	45	12/13/2018	3	12/27/2018	24	01/13/2019	60
	11/22/2018	14	12/02/2018	45	12/14/2018	15	12/28/2018	35	01/24/2019	20
	11/26/2018	8	12/04/2018	16	12/18/2018	41	12/29/2018	5	01/25/2019	7
	11/27/2018	33	12/05/2018	48	12/19/2018	34	01/03/2019	26	01/26/2019	14
Testing Dataset	11/10/2018	7	11/29/2018	19	12/06/2018	32	12/20/2018	43	01/07/2019	3
	11/11/2018	48	11/30/2018	35	12/12/2018	11	12/21/2018	15	01/09/2019	25
	11/12/2018	11	12/01/2018	62	12/13/2018	1	12/27/2018	20	01/13/2019	57
	11/22/2018	15	12/02/2018	36	12/14/2018	15	12/28/2018	21	01/24/2019	15
	11/26/2018	5	12/04/2018	15	12/18/2018	43	12/29/2018	2	01/25/2019	20
	11/27/2018	34	12/05/2018	47	12/19/2018	36	01/03/2019	26	01/26/2019	20

training, validation and testing datasets were split in 60-20-20%, respectively, of the total scans within the PPI dataset. As shown in Table 4.2, the PPI training dataset consists of 2,217 PPI scans, the PPI validation dataset consists of 739 PPI scans, and the PPI testing dataset consists of 739 PPI scans. During this time, low-level 360° surveillance scans were conducted. Two PPIs at 0.5° and 1.5° elevation were collected at a scan rate of $20^\circ/\text{sec}$. The total scan time for this scan strategy was 42 seconds.

Chapter 5

Research Methodology and Experiments

In order to thoroughly assess the potential of SRGANs within the weather radar regime, the details of the experimentation process were carefully planned. Chapter 5 serves to explain this research's methodologies and the reasoning behind its experiments. The coding software and processing environment used are discussed in Chapter 5.1 which also presents tables that outline the specifications for these. Chapter 5.2 presents the baseline interpolation techniques that were established as the "control variables" for experimentation. These will be used as the standards in super-resolution to compare the performance of the experimental SRGAN models against. However, before experimentation, the hyperparameters of the experimental SRGAN models were optimized. This ensured that the models were fine-tuned during experimentation. The process to optimize the hyperparameters is detailed in Chapter 5.3. Many experimental SRGAN models were set up for investigation throughout this study. Chapter 5.4 outlines the experimentation process including the dataset-based and architectural-based experiments carried out by training a multitude of experimental SRGAN models with different configurations; all in order to test for the best performing SRGAN model. Chapter 5.5 discusses the training methodology used for the experimental SRGAN models. After training, each experimental SRGAN model was evaluated using a set of evaluation metrics which consisted of image processing related metrics similar to those listed in Table 2.1. These evaluation metrics are defined in Chapter 5.6. All of this together is used to thoroughly explain the experimentation methodologies as they were developed and conducted in this thesis.

5.1 Environment and Software

A Linux environment was utilized to conduct the subsequent research tasks. Linux allows for efficient module creation and training/evaluating operations. Additional modules can be developed at any time after the initial scripts are written and can simply be imported in-line, ready-to-use,

while operational runs can be started with a single terminal command. All training processes were run using a NVIDIA Tesla P100 GPU with 12 GB of memory. A GPU is highly recommended and, often times, necessary when training deep learning models on larger data types, especially images, since it drastically shortens training times. Further details of the Linux environment are shown in Table 5.1 below.

Table 5.1: Processing Environment Specifications

System	Specification
Operating System	CentOS Linux 7 (Core)
CPU	1x Intel Xeon E5-2683 16-core (2.1 GHz) Memory: 128 GB RAM
GPU	1x NVIDIA Tesla P100 Memory: 12 GB

All of the programming scripts were written in Python. The code was designed using object-oriented programming (OOP) for the model architectures, hyperparameter optimization, training and evaluation processes. This makes the code more manageable as well as more adaptable to future research applications. For this thesis, the Tensorflow Keras API was selected as the foundational machine learning framework as its logical workflow and flexibility enables researchers to rapidly develop machine learning models. All of the primary software version requirements for this thesis are listed in Table 5.2 below.

Table 5.2: Coding Software

Software	Version	Description	Software	Version	Description
CUDA	11.5	CentOS Linux 7 (Core)	Keras	2.4.3	Tensorflow's High-Level API
Python	3.9.6	Programming Language Utilized	Imageio	2.9.0	Image Transformations
Tensorflow	2.4.1	Machine Learning Library	Matplotlib	3.4.3	Image Plotting/Generation

5.2 Baseline Models

Ever since imaging technology was created, there has been a persistent drive to develop clearer, higher resolution images. Digital image processing spurred a new revolution of reaching these high image resolution regimes with the concept of super-resolving LR images. Since its inception, SR image processing techniques have been continuously researched and developed, especially within the computer vision community. Chapter 2 comprehensively explained both the fundamental principles as well as more current innovations to SR methods. This thesis research focuses specifically on the use of the SRGAN model [14] and its effectiveness in super-resolving weather radar scans.

In order to assess the effectiveness of the SRGAN model in super-resolving weather radar scans, the trained SRGAN models' performances needs to be compared against standard techniques prevalently used in the rest of the literature. These standard techniques are referred to as baseline models. Table 2.1 shows that the baseline model most widely used for evaluating DL SR models is interpolation. Because of this, interpolation techniques were established as the control variables for experimentation against which the experimental SRGAN models will be compared. This will demonstrate how well the experimental SRGAN models perform compared to the baseline models that are already used commonly in the literature. The baseline models used were determined with the help of Table 2.1. As can be seen in the table, it is apparent that the most prominent comparison method utilized is bicubic interpolation. Each and every study evaluated their DL SR technique against bicubic interpolation making it an obvious choice for a baseline model. For this thesis study, bicubic interpolation as well as additional interpolation techniques are utilized as the evaluation baselines, namely the nearest neighbors and Lanczos interpolation techniques.

5.3 Hyperparameter Optimization

Before training can begin, the experimental SRGAN models undergo the hyperparamter optimization (HPO) process. Hyperparameters define the ML model's architecture and, thus, must be set before training the ML model. The HPO process guides the SRGAN model to be able to reach

its maximum potential. Throughout this process, the model becomes fine tuned by adjusting the primary characteristics of the model's architecture; in turn, this optimization enhances the overall performance of the model itself. These primary characteristics have been termed hyperparameters. Essentially, they work as the operative controllers of the model's functional efficiency. For SRGANs specifically, this means that the HPO process works to improve the SRGAN's ability to generate SR images that are closer to the target HR images from the foreign input LR images that were not used to either train or optimize the network, previously. A standard procedure for conducting manual HPO on a SRGAN model is to define the optimization space (determine hyperparameters to experiment with and which values to investigate for each of those hyperparameters), train individual models on the training dataset over the entire optimization space, evaluate the models on the validation dataset, and, finally, analyze the evaluations and compare between them for the highest performing combination of hyperparameters.

After reviewing the research literature, each paper's HPO methodology and how the HPO process was conducted is not always explicitly described in a detailed manner. Instead, most papers solely list the hyperparameters used. Be that as it may, this thesis will wholly describe the HPO process utilized. HPO is a crucial step in developing a DL model. By fine-tuning the hyperparameters of the ML model's architecture, the model's overall performance can be enhanced. HPO is necessary in order to develop a well-functioning, robust DL model that achieves its optimal performance. Although many HPO techniques exist, this thesis study uses the manual grid search HPO method based on other SRGAN models in literature found during research. This was utilized so that the model's operative function could be closely monitored by the designer and any unexpected deviations in function could be diagnosed quickly.

The manual grid search HPO process begins with defining the optimization space by choosing a set of hyperparameters to investigate and a range for each of their values to fine-tune. The selected hyperparameters and ranges are based on other optimization spaces utilized in literature, such as in the comprehensive study found in [57]. Once defined, multiple training runs of the experimental SRGAN models are conducted with each combination of the hyperparameters' values until all

possible combinations have been trained. Individual models are trained for each combination of the hyperparameters' sets of values. The models are then evaluated using the validation dataset described in Chapter 4.4. The validation dataset is only used for the HPO process. This ensures that the experimental SRGAN models' development is only based on the validation dataset during the HPO process and the training dataset during the training process, making it independent from the testing dataset that they will be evaluated on during the final evaluation.

The realm of the optimization space is characterized by a set of hyperparameters and the ranges of their values. After searching through this space, a subset of the hyperparameters are found to be the optimized set through the use of HPO techniques. These hyperparameters allow the SRGAN to reach its full potential as they are dialled in to optimize its performance. During training, they work to ensure that the SRGAN will produce super-resolution images of a higher quality. The rest of this chapter will detail the different hyperparameters and their range of values used throughout the HPO process.

The hyperparameters were selected by considering the architectural construct of the SRGAN while including standard hyperparameters that have been optimized for other SRGAN models in the literature. This thesis will use the optimizer learning rate, the optimizer beta_1 variable, the batch normalization momentum values for both the discriminator and the generator, as well as the batch size for its hyperparameters. These define the optimization space for HPO that will be manually searched through in order to find the optimized set of hyperparameters that will allow the experimental SRGANs to reach their full potential. For this thesis, HPO is performed on each individual dataset listed in 4.4.1. Reference values for these hyperparameters were set to 0.0002 for the learning rate, 0.5 for the beta_1 variable, 0.8 for the batch normalization momentum variables for both the discriminator and the generator, and 10 for the batch size. These helped guide the complete population of the optimization space by basing the ranges of each of the hyperparameters around these reference quantities listed. After preliminary test runs were conducted using these reference values as a basis, each hyperparameter's range of values was found. These ranges are detailed in the paragraphs following.

The first two hyperparameters, the optimizer's learning rate and beta_1 variable, define the optimizer of the ML model. Optimizers are used within most ML models. Their algorithmic function is to guide the experimental model during training, helping the model to produce better results. This thesis employs the Tensorflow Keras' Adam optimizer for the experimental SRGAN models. The Adam optimizer is a standard across most SRGANs in literature. This optimizer implements the Adam algorithm as described in [58]. The Adam algorithm adaptively estimates the low-order moments in order to optimize objective functions using stochastic gradient descent. For the SRGAN model, this means that the Adam optimizer assists in optimizing the objective function as defined in Equation 2.4. The Adam optimizer's primary parameters that control its functionality are its learning rate and beta_1 variable. These function, and are commonly utilized, as hyperparameters for the ML model itself during HPO.

The learning rate affects the level of correction the model undergoes from one epoch of training to the next. The beta_1 value is the first moment estimates' exponential decay rate. For the SRGAN, both the learning rate and the beta_1 variable of the optimizer is used to compile each of the neural networks that comprise the SRGAN model. This includes the generator, discriminator and the VGG19 model. The optimizer learning rate affects the extent to which these models are updated throughout the training process. Since the VGG19 model is pre-trained, as it determines the perceptual loss of the model's objective function, the optimizer's hyperparameters are solely used to compile the VGG19 model itself. It is defined as untrainable during the training process. The discriminator is trained and updated independently from the combined SRGAN model. This way, the discriminator does not over-train and outpace the generator. The generator model is updated with every epoch of training the combine SRGAN model. This occurs after the generator gets feedback on its SR image from the discriminator's classification of how close it is to the HR, ground-truth target image. The learning rate is defined as the rate of change of the error between each training epoch. If the learning rate is set to be too high a value, the rate of change will be drastically increased and the model will likely not settle into the minimum value of the objective loss function in Equation 2.4 as it will over-correct and miss the minimum/maximum

point. However, if the learning rate is too small, then the rate of change will be almost negligible, nullifying any changes between the training epochs so that the minimum/maximum point may not be reached within the training runs allotted. A small learning rate can also be time inefficient as it would require an increased number of training runs in order to reach the minimum/maximum point. Thus, a balance must be found for this hyperparameter to achieve maximum efficiency in reaching the minimum/maximum point while maintaining close attention to detail as to not overlook it. The set of values for the learning rate hyperparameter used within the optimization space was defined as: [0.0002, 0.0004, 0.0008].

The beta_1 variable of the optimizer controls the exponential decay rate of the Adam algorithm's exponential moving averages. These moving averages are estimates of the mean of the gradient, also called the first moment. The gradient is defined as the vector of partial derivatives of the objective function. A typical range for the beta_1 variable that covers a broad span of values is when $\beta_1 \in [0, 0.9]$ as described in [58]. Based on this range, the set of values for the beta_1 hyperparameter to be investigated was defined as: [0.1, 0.5, 0.9].

The batch normalization layers within the SRGAN model apply a transformation to the input by passing a kernel filter over the image. In doing so, the batch normalization layer acts to normalize the output, holding the output's mean close to 0 and its standard deviation close to 1. The momentum variable of the batch normalization controls the spatial movement of the filter and the amount of filter normalization coverage that passes over the image. It is important to note that the batch normalization layers behave differently during training and during inference. During training, the batch normalization layers use the mean and standard deviation of the current batch of inputs in order to normalize the output. For each channel being normalized, the batch normalization layer returns an output according to Equation 2.14 and described in Chapter 2.3.2. After training, when the trained SRGAN model is called to predict/generate a SR image, the batch normalization layers use a moving average of the mean and standard deviation of the batches that it saw during training in order to normalize the output according to Equation 2.15 and described in Chapter 2.3.2. The set of values for the discriminator's batch normalization momentum hyperparameter used within

the optimization space was defined as: [0.4, 0.6, 0.8, 0.99]. The set of values for the generator's batch normalization momentum hyperparameter used within the optimization space was defined as: [0.4, 0.6, 0.8, 0.99].

The batch size determines the number of HR, LR input image pairs that the SRGAN model trains on during a single epoch of the training session. Using a subset of the training dataset, i.e. a batch, speeds up training times. The size of the batch used during training can impact the SRGAN model's learning development. Typically, functional batch sizes for training GAN models reside within the small-batch regime. Having a larger batch size results in diminishing marginal returns as the batch size is increased, resulting in a degradation in the quality of the model's output. For example, a model being trained on a batch of size n will expend $O(n)$ of the available memory in order to conduct $O(n)$ calculations. However, the amount of uncertainty in the gradient is only reduced by $O(\sqrt{n})$ [59]. Larger batch sizes also requires greater computational resources to run each epoch of the training sessions, leading to slower training times. For this thesis, computational resources were a limiting factor and, as such, the set of values for the batch size hyperparameter used within the optimization space was defined as: [1, 5, 10]. Table 5.3 presents a comprehensive list of the investigated hyperparameters including their reference values and the set of experimental hyperparameter values tested during HPO, i.e. the optimization space.

The SRGAN HPO training runs were conducted in two sets of two, following a similar format to how the experimental runs are trained, as discussed in Chapter 4.4.1. One set consisted of two different training datasets corresponding to different radar scan types, PPI and RHI. These sets were then further divided by the x2 and x4 resolution scaling. These model characteristics were determined to directly affect the experimental parameters of the model. Since the structural layout of the PPI and RHI images are distinct from one another, HPO was conducted for individual SRGAN models for each of these scan types because SRGAN models are found to perform better when trained and evaluated on the separate datasets, even though comparable performances can be achieved with a more general, wide range of images used for training. This would allow the experimental SRGAN models the best environment to train under for enhancing the end-performance.

Table 5.3: HPO Optimization Space

Parameter	Reference Value	Set of Values
Learning Rate	0.0002	[0.0002, 0.0004, 0.0008]
Beta_1	0.5	[0.1, 0.5, 0.9]
Discriminator Batch Normalization Momentum	0.8	[0.4, 0.6, 0.8, 0.99]
Generator Batch Normalization Momentum	0.8	[0.4, 0.6, 0.8, 0.99]
Batch Size	10	[1, 5, 10]

HPO was conducted for individual SRGAN models for each of the resolution scaling factors utilized as well, due to the factors discussed prior. In addition, most studies in literature train individual SRGAN models based on the SR scaling being performed. These divisions were put in place to stay consistent with the procedural paradigm in literature while setting the thesis SRGAN model up for its optimum level of success. Therefore, four different SRGAN models are individually optimized via HPO, namely: RHIx2, RHIx4, PPIx2, and PPIx4 SRGAN models.

The SRGAN models were trained with each and every combination of the hyperparameters defined within the optimization space: the optimizer learning rate, the optimizer beta_1 value, the batch normalization momentum for the discriminator and the generator, and the batch size. Training was conducted on the training dataset while the HPO evaluation was conducted on the validation dataset. The primary purpose of the validation dataset is to be utilized for the hyperparameter tuning of the model before the final experimental evaluations are conducted. By doing this, the model's development is not influenced by the testing dataset during either training or optimization so that the final experimental evaluation results are not biased towards the testing dataset samples. Until the final evaluation of the experimental models is underway, the testing dataset should be completely foreign to the SRGAN models being tested. The same evaluation metrics

will be employed for both the HPO evaluation as well as the final experimental model evaluation: PSNR, MSE, and SSIM. For further details on the evaluation metrics, see Chapter 5.6. In addition, the HPO evaluation will also take into account the final validation loss. The final validation loss is calculated and tracked during training. This is the perceptual loss mentioned in Chapter 2.1. These values were all considered when determining which combination of hyperparameters would be used for the experimental RHIx2, RHIx4, PPIx2, and PPIx4 SRGAN models going forward during the training and evaluation phases. The optimized set of hyperparameters for each experimental model can be found at the bottom of Tables 5.4 - 5.7.

For the RHIx2 SRGAN model's HPO runs, the optimized set of hyperparameters was quite apparent. The evaluations for the RHIx2 HPO runs can be found in Table 5.4. When comparing the different values tested for the discriminator batch normalization momentum, the reference value, 0.8, was chosen as the optimized value since it performed quite well on all of the evaluation metrics. It had the second highest overall PSNR and the lowest overall MSE values across all of the RHIx2 HPO runs. In addition, it had the second to lowest final validation loss as well as the highest SSIM when compared with the other discriminator batch normalization momentum evaluations. Similar logic was used when determining that the reference value, 0.8, was chosen as the optimized value for the generator batch normalization momentum. Although the 0.6 value had the highest overall SSIM evaluation, the reference value, 0.8, still held the second highest overall PSNR, the lowest overall MSE, and the lowest final validation loss when compared with the other evaluations in the generator batch normalization momentum group. For the batch size hyperparameter, 10, the reference value, was found to be the optimized value. In addition to having the second highest overall PSNR and the lowest overall MSE, it had the lowest validation loss and the highest SSIM when compared with the other evaluations in the batch size group. The reference learning rate was determined to not be the optimized value; instead, 0.0004 was used as the optimized learning rate hyperparameter value for the RHIx2 SRGAN model. Several factors were considered when making this decision. The 0.0004 value HPO evaluation had the second lowest overall final validation loss and the highest overall PSNR making it the elected candidate for the optimized

Table 5.4: Hyperparameter Optimization SRGAN: RHI x2

Test Parameter	Dis. Batch Norm	Gen. Batch Norm	Batch Size	Learning Rate	Beta	Final Validation Loss	PSNR	MSE	SSIM
Reference	0.8	0.8	10	0.0002	0.5	8.97	23.00	**0.030**	0.861
Dis. Batch Norm	0.6	0.8	10	0.0002	0.5	9.09	22.86	**0.030**	0.860
	0.4	0.8	10	0.0002	0.5	9.31	22.70	0.033	0.855
	0.99	0.8	10	0.0002	0.5	8.95	22.79	0.032	0.857
Gen. Batch Norm	0.8	0.6	10	0.0002	0.5	9.09	22.75	0.031	**0.862**
	0.8	0.4	10	0.0002	0.5	9.16	22.67	0.031	0.858
	0.8	0.99	10	0.0002	0.5	12.92	20.95	0.042	0.831
Batch Size	0.8	0.8	1	0.0002	0.5	16.45	8.91	0.518	0.461
	0.8	0.8	5	0.0002	0.5	9.28	22.56	0.034	0.850
Learning Rate	0.8	0.8	10	0.0004	0.5	8.53	**23.07**	0.031	0.856
	0.8	0.8	10	0.0008	0.5	**8.33**	22.73	0.031	0.857
Beta	0.8	0.8	10	0.0002	0.1	9.46	22.16	0.034	0.851
	0.8	0.8	10	0.0002	0.9	9.56	22.02	0.036	0.842

Optimized Set of Hyperparameters Used

	0.8	0.8	10	0.0004	0.5				

learning rate hyperparameter value. The reference value of 0.5 was also used as the optimized beta_1 hyperparameter value for the same reasons mentioned above for the batch normalization momentum and the batch size hyperparameters. Thus, the optimized set of hyperparameters for the RHIx2 SRGAN model chosen proceeds as follows: [0.8, 0.8, 10, 0.0004, 0.5].

Table 5.5 displays the evaluations for the RHIx4 SRGAN model's HPO runs. Through these HPO runs, it was determined that the reference values for the generator batch normalization momentum and the batch size, 0.8 and 10 respectively, would be chosen as elements within the optimized set of hyperparameters. The reference values for both of these hyperparameters scored the best overall on each of the evaluation metrics within their respective optimization groups. The reference generator batch normalization momentum value also had the second lowest final validation loss within its group while the reference batch size value had the lowest final validation loss within its group. The optimized value for the discriminator batch normalization momentum was found to be 0.6 as it not only had the best overall performance on all of the evaluation metrics: PSNR, MSE, and SSIM; but, also, had the second lowest final validation loss when compared to the rest of the HPO runs. These assessments made it the clear choice to include in the optimized set of hyperparameters. Similar to the RHIx2 SRGAN model HPO runs, 0.0004 was concluded to be the optimized value for the learning rate. This HPO run had the highest overall PSNR within its optimization group; the second highest PSNR overall. It also had the lowest final validation loss overall while maintaining a SSIM and MSE comparable to the highest SSIM and the lowest MSE, obtained by the reference value, within its optimization group. Lastly, the optimized beta_1 value was found to be 0.9. Although this value had the second best final validation loss and PSNR in its optimization group, it had the best MSE and SSIM within its optimization group which were also the second best MSE and SSIM over all of the HPO runs for the RHIx4 SRGAN model. The final optimized set of hyperparameters for the RHIx4 SRGAN model was found to be: [0.6, 0.8, 10, 0.0004, 0.9].

The evaluations for the PPIx2 SRGAN model's HPO runs can be found in Table 5.6. After examining the table, the optimized set of hyperparameters becomes quite evident. It is interesting

Table 5.5: Hyperparameter Optimization SRGAN: RHI x4

Test Parameter	Dis. Batch Norm	Gen. Batch Norm	Batch Size	Learning Rate	Beta	Final Validation Loss	PSNR	MSE	SSIM
Reference	0.8	0.8	10	0.0002	0.5	15.29	20.73	0.051	0.822
Dis. Batch Norm	0.6	0.8	10	0.0002	0.5	14.87	**20.98**	**0.050**	**0.827**
	0.4	0.8	10	0.0002	0.5	15.34	20.47	0.055	0.814
	0.99	0.8	10	0.0002	0.5	15.91	19.52	0.075	0.800
Gen. Batch Norm	0.8	0.6	10	0.0002	0.5	15.48	20.40	0.051	0.820
	0.8	0.4	10	0.0002	0.5	15.21	20.41	0.054	0.817
	0.8	0.99	10	0.0002	0.5	18.68	19.11	0.078	0.792
Batch Size	0.8	0.8	1	0.0002	0.5	16.75	19.91	0.066	0.805
	0.8	0.8	5	0.0002	0.5	15.55	20.35	0.057	0.813
Learning Rate	0.8	0.8	10	0.0004	0.5	**14.71**	20.96	0.052	0.820
	0.8	0.8	10	0.0008	0.5	17.94	15.04	0.267	0.738
Beta	0.8	0.8	10	0.0002	0.1	15.91	20.23	0.058	0.804
	0.8	0.8	10	0.0002	0.9	15.37	20.72	0.051	0.824

Optimized Set of Hyperparameters Used

	0.6	0.8	10	0.0004	0.9				

to note that, for this set of HPO runs, only one of the reference values was evaluated to have a high enough performance to be included within the optimized set of hyperparameters. All of the rest of the hyperparameters' optimized values as other values from within the optimization space. This is a stark difference to the RHIx2 SRGAN model's HPO evaluation whose optimized set of hyperparameters consisted primarily of the reference values for the hyperparameters. The only reference value utilized in the optimized set of hyperparameters for the PPIx2 SRGAN model was the batch size with a batch size value of 10. The reference value for the batch size was found to be optimal since it had the lowest final validation loss, the highest PSNR, the lowest MSE, and the highest SSIM within its optimization group. For the discriminator batch normalization momentum hyperparameter, the optimized value was found to be 0.99. Inputting this value for the hyperparameter gave the model the best performance on all of the evaluation metrics for the HPO runs conducted within its optimization group, making it the clear choice for the optimized set. Similarly, the generator batch normalization momentum's optimized value was found to be 0.6. Once evaluated, this HPO run was determined to have the best performance across all of the evaluation metrics when compared to the rest of its optimization group as well. The optimized value for the learning rate, 0.0008, had the absolute best overall performance on the HPO evaluations for the PPIx2 SRGAN model HPO runs across all of the evaluation metrics. Therefore, it was decisively selected for inclusion in the optimized set of hyperparameters as the optimized learning rate value. Finally, the beta_1 hyperparameter's optimized value was determined to be 0.1. When evaluated, this hyperparameter's HPO evaluation had tied with the reference value, 0.5, for the lowest MSE evaluation within the optimization group. Furthermore, even through the reference value was evaluated as having a higher SSIM, the beta_1 HPO run of 0.1 was evaluated as having the lowest final validation loss as well as the highest PSNR out of its optimization group. When they are all grouped together, the final optimized set of hyperparameters for the PPIx2 SRGAN model that was utilized proceeds as follows: [0.99, 0.6, 10, 0.0008, 0.1].

Lastly, Table 5.7 contains all of the evaluations for the PPIx4 SRGAN model's HPO runs. It is interesting to note that the optimized set of hyperparameters for the PPIx4 SRGAN model's HPO

Table 5.6: Hyperparameter Optimization SRGAN: PPI x2

Test Parameter	Dis. Batch Norm	Gen. Batch Norm	Batch Size	Learning Rate	Beta	Final Validation Loss	PSNR	MSE	SSIM
Reference	0.8	0.8	10	0.0002	0.5	11.09	20.77	0.044	0.872
Dis. Batch Norm	0.6	0.8	10	0.0002	0.5	11.02	20.76	0.044	0.872
	0.4	0.8	10	0.0002	0.5	10.73	20.96	0.043	0.872
	0.99	0.8	10	0.0002	0.5	10.54	21.09	0.042	0.874
Gen. Batch Norm	0.8	0.6	10	0.0002	0.5	10.96	21.24	0.040	0.873
	0.8	0.4	10	0.0002	0.5	11.42	20.11	0.047	0.861
	0.8	0.99	10	0.0002	0.5	14.53	19.93	0.051	0.841
Batch Size	0.8	0.8	1	0.0002	0.5	18.43	8.13	0.616	0.346
	0.8	0.8	5	0.0002	0.5	11.89	20.52	0.047	0.863
Learning Rate	0.8	0.8	10	0.0004	0.5	10.07	21.58	**0.037**	0.884
	0.8	0.8	10	0.0008	0.5	**9.91**	**21.59**	**0.037**	**0.888**
Beta	0.8	0.8	10	0.0002	0.1	10.81	20.81	0.044	0.869
	0.8	0.8	10	0.0002	0.9	12.51	19.17	0.063	0.850

Optimized Set of Hyperparameters Used

	0.99	0.6	10	0.0008	0.1				

runs contained only the reference values chosen as the optimized values for all of the hyperparameters. The final validation loss evaluation for the set of reference values for the hyperparameters was determined as having the highest performance overall out of all of the HPO runs for each hyperparameter group. The only HPO run that was close in evaluation of its final validation loss to the corresponding reference value was the learning rate HPO run with a value of 0.0004 which tied for the lowest final validation loss of 18.46. This is similar to the findings of the SSIM evaluation metric. The HPO run evaluation results showed that the set of reference hyperparameter values had the highest SSIM overall out of all of the HPO runs for each hyperparameter group. The only HPO run that contested this result in the SSIM compared to the reference value was the discriminator batch normalization HPO run with a value of 0.6 which tied for the highest SSIM of 0.840. The PSNR evaluation for the hyperparameters' set of reference values was determined to be the highest performing overall out of all of the HPO runs for each hyperparameter group. The MSE evaluation for the hyperparameters' set of reference values was found to be tied with the 0.0004 learning rate HPO run as having the second lowest MSE overall. The discriminator batch normalization HPO run with a value of 0.4 had the lowest MSE overall out of all of the HPO runs for each hyperparameter group. However, since the set of reference values for the hyperparameters was found to have the highest performing final validation loss, PSNR and SSIM overall of the HPO runs, the set of reference values were utilized as the optimized set of hyperparameters for the PPIx4 SRGAN model, which was defined as: [0.8, 0.8, 10, 0.0002, 0.5].

From this HPO procedure, patterns can be discerned that give insight into the SRGAN model's performance with different hyperparameters in general. For instance, when the batch size was set to the lowest value within the optimization space to a batch size of 1, the RHIx2 and PPIx2 SRGAN models' HPO runs were evaluated to have the lowest performance out of any HPO run for every evaluation metric used. This suggests that low batch size values are undesirable for x2 resolution scale SRGAN models. The HPO run with a batch size of 1 had lower performances for the RHIx4 and PPIx4 SRGAN models' HPO runs as well; however, they were not the lowest performing sets of hyperparameter values used for these SRGAN models. This suggestion is further promoted by

Table 5.7: Hyperparameter Optimization SRGAN: PPI x4

Test Parameter	Dis. Batch Norm	Gen. Batch Norm	Batch Size	Learning Rate	Beta	Final Validation Loss	PSNR	MSE	SSIM
Reference	0.8	0.8	10	0.0002	0.5	**18.46**	**19.12**	0.063	**0.840**
Dis. Batch Norm	0.6	0.8	10	0.0002	0.5	18.65	18.83	0.068	**0.840**
	0.4	0.8	10	0.0002	0.5	18.76	19.09	**0.062**	0.838
	0.99	0.8	10	0.0002	0.5	18.57	19.10	0.065	0.838
Gen. Batch Norm	0.8	0.6	10	0.0002	0.5	18.49	18.95	0.064	0.838
	0.8	0.4	10	0.0002	0.5	18.90	18.81	0.064	0.818
	0.8	0.99	10	0.0002	0.5	24.55	17.80	0.081	0.806
Batch Size	0.8	0.8	1	0.0002	0.5	19.91	18.62	0.066	0.818
	0.8	0.8	5	0.0002	0.5	18.95	18.78	0.068	0.833
Learning Rate	0.8	0.8	10	0.0004	0.5	**18.46**	19.00	0.063	0.825
	0.8	0.8	10	0.0008	0.5	18.94	18.49	0.074	0.832
Beta	0.8	0.8	10	0.0002	0.1	19.49	17.61	0.091	0.819
	0.8	0.8	10	0.0002	0.9	22.82	14.69	0.192	0.780

Optimized Set of Hyperparameters Used

	0.8	0.8	10	0.0002	0.5				

the observation that every optimized set of hyperparameters for every SRGAN model investigated throughout the HPO process contained the largest value within the optimization space for the batch size hyperparameter with a batch size of 10.

5.4 Experiments

Once the HPO process was completed, the experiments for this thesis study were defined. The first set of experiments specifically pertains to the dataset used when developing the SRGAN model. As further detailed in Chapter 4, the dataset serves as the backbone when developing a SRGAN model. The datasets utilized are imperative to the overall function and performance of any SRGAN model. Because of this, the dataset was chosen as a central focus for experimentation. The radar scan type, the resolution scale factor, and the downsampling method used to create the LR input dataset were all used as experimental variables in this thesis. The other primary set of experiments revolves around adjusting parameters within the SRGAN architecture, namely the discriminator filter size, the generator filter size, and the number of residual blocks used in the generator. These parameters were found to affect the performance/behavior of the SRGAN model during initial testing. The rest of this chapter further describes the specifics of these experiments and how they were carried out.

5.4.1 Dataset Experiments

A SRGAN model's dataset is a foundational component that is used during all of aspects of developing the model. The dataset is split into three datasets that fulfill different roles during the developmental stages. These three datasets: the training, validation, and testing datasets, were utilized throughout the thesis to progress the SRGAN model and to allow it to generate the best possible physically realistic, SR weather radar images. The training and validation datasets were used during the HPO process to ensure that the best parameters controlling the behavior of the SRGAN were used throughout the experimentation process. The training dataset was used again during the experimental runs to educate the SRGAN on how to generate SR images that are as

close to the HR input images as possible. Finally, the testing dataset was used during the evaluation process to assess the performance of the SRGAN on taking input LR images, that are completely foreign to the SRGAN model, and generating SR weather radar images. These were compared against the HR weather radar images that paired with the input LR images. Since the dataset greatly influences the progress of the SRGAN's evolution, experiments were set up to determine the extent of the SRGAN model's capabilities in super-resolving weather radar image datasets with different characteristics. Three main features of the datasets used established the experimental variables. These were determined as follows: the radar scan type, the resolution scaling factor, and the method for processing the LR dataset.

The scope of this thesis specifies investigating the efficacy of utilizing SRGAN models for super-resolving weather radar images. In order to accomplish this, the study must examine the main types of weather radar images, namely: RHI and PPI. As discussed in Chapter 3, there are two primary types of weather radar scans that collect information on different sections of a weather event. RHI scans probe a vertical sector of the observed hydrometeor volume. RHI's give information about the precipitation present at different elevations of the radar. These types of radar scans show the layers of precipitation and can give insight into valuable weather event characteristics such as the melting layer. PPI scans probe a horizontal sector of the hydrometeor volume. PPI's give information about the precipitation present at different azimuths of the radar. PPI data is commonly employed as weather forecasting data. These scans clearly illustrate the location and severity of a weather event. PPI's also give significant insight into the shape of weather events, the progression of their development, as well as other characteristic behaviors that allow for better informed forecasting and safety decision making. Since both of these types of radar scans are important in the weather radar field, it was determined that both RHI and PPI data would be used as experiments to test the capabilities of the SRGAN models to generate super-resolved images for both of the primary types of radar scans. This allowed for any discrepancies between the SRGAN's performance in super-resolving RHI images versus PPI images to be thoroughly investigated. Because of this decision, the experimental SRGAN models were split into two groups. One used the

RHI images as inputs into the SRGAN models for training. The other input the PPI images into the SRGAN models during training. This resulted in two experimental SRGAN model groups being formed: RHI and PPI SRGAN models.

As discussed in Chapter 4 and shown in Table 2.1, the resolution scaling factor commonly serves as an experimental factor in prevalent super-resolution research papers. From Table 2.1, x2 and x4 are the most common resolution scaling factors used for experimentation. Since these resolution scaling factors are utilized throughout DL SR model literature, these were also selected as experimental variables to be explored in this thesis. In order to give the experimental SRGAN models the best environment for its developmental learning, the experimental SRGAN models were split further into two more groups so that individual models were trained for each resolution scaling factor tested. One used the x2 resolution scaling factor and input LR images with a size of 128x128x3 pixels. The other used the x4 resolution scaling factor and input LR images with a size of 64x64x3 pixels. For reference, the HR dataset contains images with a size of 256x256x3, which is the same size as the generated SR images as well. Thus, the experiments to be tested so far were determined to be four different experimental SRGAN model groups, listed as: RHIx2, RHIx4, PPIx2 and PPIx4 SRGAN models.

The final dataset-type experiment that was established pertained to the pre-processing, downsampling method used for creating the input LR image dataset from the HR image dataset. According to the literature, it is standard practice to use a bicubic or gaussian interpolation kernel in order to produce LR versions of the HR input images. This can be seen in the "Downsample Methods" column in Table 2.1. In order to stay consistent with the literature, a bicubic interpolation kernel was used to downsample the HR images. However, these LR images are not indicative of the physical characteristics present when producing a LR radar scan. Typically, the scanning rate of the radar is increased when collecting LR radar data. As can be seen in Equation 3.4, when a radar's scanning rate is increased, this results in the number of radar beams decreasing. This effect can be seen as in Figures 4.4 and 4.5 above when plotting the LR data with the physically representative LR downsampling method. The interpolation kernel method of downsampling cannot simulate

the characteristics of a physical LR scan, so a new method of downsampling was developed and utilized for this purpose. This technique was employed during the pre-processing stage of dataset creation.

In order to study the effectiveness of these downsampling methods, the experimental SRGAN models were divided into two more groups. One used the standard bicubic interpolation kernel and the other used the new physically representative downsampling technique. In order to fully investigate each of these experiments, individual SRGAN models were trained for each combination of the experimental variables described above. Therefore, this resulted in a total of eight SRGAN models to be tested. From this point onwards, these experimental SRGAN models will be referred to as follows: RHIx2_Interp for the RHIx2 Interpolation Dataset SRGAN models, RHIx2_PhysRep for the RHIx2 Physically Representative Dataset SRGAN models, RHIx4_Interp for the RHIx4 Interpolation Dataset SRGAN models, RHIx4_PhysRep for the RHIx4 Physically Representative Dataset SRGAN models, PPIx2_Interp for the PPIx2 Interpolation Dataset SRGAN models, PPIx2_PhysRep for the PPIx2 Physically Representative Dataset SRGAN models, PPIx4_Interp for the PPIx4 Interpolation Dataset SRGAN models, and PPIx4_PhysRep for the PPIx4 Physically Representative Dataset SRGAN models.

5.4.2 Model Parameter Experiments

The composition of a SRGAN model's architecture serves as another foundational component throughout the model's developmental progress. To begin this research, the original SRGAN model developed by Ledig et al. [14] was chosen as the focus of study. After this SRGAN model was chosen, the architecture was written into code using a modular approach called object oriented programming. Initial testing was conducted to check for viability and experimental variables.

The way in which a SRGAN is built has significant impact on its performance. This thesis aims to investigate the ability of the first SRGAN presented in literature, as in [14], that is considered to be the primary SRGAN architecture when discussed within the machine learning community. Therefore, the foundational architecture from [14] remained unchanged throughout this thesis to

preserve the scope of this study. Instead, different variables within the architecture that control how the model behaves were experimented on. These internal architectural variables were found to affect the performance of the SRGAN model during testing while maintaining the foundational infrastructure of the SRGAN model as it was developed in [14]. The internal architectural variables under investigation are the discriminator filter size (DFS) primarily used as an input to the 2DConv and dense layers within the discriminator neural network, the generator filter size (GFS) primarily used as an input to the 2DConv layers within the generator neural network, and the number of residual blocks (NRB) used within the generator neural network.

In order to investigate how these internal architectural variables affected the behavior of the experimental SRGAN models, all eight SRGAN models were trained multiple times with different combinations of values for the model parameter experimental variables. Reference values for the model parameter experimental variables were obtained from [14] which used a DFS of 64, a GFS of 64 and a NRB of 16 for their SRGAN model. These reference values were used as the reference/basis values for the x4 experimental SRGAN models. However, the x2 experimental SRGAN models used different reference values for the DFS and GFS experimental variables of 32 instead of 64. The set of additional experimental values for these model parameter variables was based around these reference values. Table 5.8 below presents the reference values used alongside all of the sets of experimental values used for each of the model parameter experimental variables for each type of resolution scaling experimental SRGAN model. One note to consider is that the x2 experimental SRGAN models were unable to use a NRB of 128, found in the set of experimental values for the NRB variable with the x4 SRGAN models. Having a NRB of 128 when training the x2 SRGAN models consistently caused an out of memory error to occur due to the limited computational resources available. The set of experimental values for the DFS and GFS model parameter experimental variables is the same for both the x2 and x4 SRGAN models.

Table 5.8: Model Parameter Experimental Variables' Values

Model Parameter Experimental Variable	SRGAN model	Reference Value	Set of Experimental Values
DFS	x2	32	[8, 16, 32, 64, 128]
	x4	64	[8, 16, 32, 64, 128]
GFS	x2	32	[8, 16, 32, 64, 128]
	x4	64	[8, 16, 32, 64, 128]
NRB	x2	16	[4, 8, 16, 32, 64]
	x4	16	[4, 8, 16, 32, 64, 128]

5.5 Training

All experimental SRGAN models were trained on a NVIDIA Tesla P100 GPU using a randomly sampled batch of the training dataset, whose size was determined during the HPO process. Standard SRGAN model training data augmentations prevalently used within the ML community were utilized throughout the training process, namely pixel range scaling in addition to soft and noisy labels are employed. Training was conducted individually for different combinations of the model parameter experimental variables' experimental values for all eight types of SRGAN models. Only one model parameter's values were changed for each training run. The other model parameters' values were held at their corresponding reference values during this time. Thus, 13 individual SRGAN models were trained for each x2 experimental SRGAN model configuration and 14 individual SRGAN models were trained for each x4 experimental SRGAN model configuration, since it had one more experimental value of 128 for the NRB model parameter to test. This resulted in 108 individual experimental SRGAN models being trained, evaluated and analyzed through this thesis work. Training was conducted for 500 epochs for each experimental SRGAN model. Epoch is a term used within the ML community to describe one iteration of a training loop. It can be thought of as the number of times that the experimental SRGAN model "looks" at and

learns on a batch of the training images. The number of training epochs is notably less than other studies in literature, which can be on the order of tens to hundreds of thousands of epochs for a single training cycle, due to the computational resources available.

The training process began with downloading a batch of input image pairs, a randomly sampled set of HR images and their corresponding downsampled LR images, from the training dataset. Both the LR and the HR input image's pixel ranges were scaled to [-1, 1]. The untrained generator then generated an initial SR prediction based on the LR image and the label arrays were built to be input into the discriminator for training. Two label arrays were instantiated. The "valid" label array consisted entirely of 1 values while the "fake" label array consisted entirely of 0 values. Then, both of these label arrays were transformed into soft and noisy labels by either subtracting or adding a random constant between 0 and 0.05 to the "valid" and "fake" label arrays, respectively. This data augmentation technique is used to handicap the discriminator by making its training more difficult. Generally when training SRGAN models, it is found that the discriminator loss converges at a much faster rate than the generator making the adversarial generator loss increase dramatically. With this, the discriminator learns the distribution of the input HR images so that it cannot be fooled by the generator and, thus, the generator does not get to learn through its training process. So, the input label arrays are smoothed by decreasing the "valid" label value and increasing the "fake" label value while inducing some additive noise to the values as well in order to create a better environment for training the generator model with its adversarial loss. The discriminator is then first trained on the HR images with the "valid" label array and then trained on the generated SR images from the untrained generator model with the "fake" label array. At this point, a new "valid" array is made for input into the combined SRGAN model. The VGG19 model outputs an image feature array based on the batch of HR image inputs. Then, the SRGAN model is trained, with the discriminator being set to an untrainable mode, on the batch of input LR/HR image training pairs alongside the new "valid" label array and the image feature array output by the VGG19 model. This process then repeated until the experimental SRGAN model had been trained for 500 epochs. After training, the experimental SRGAN models were evaluated as described in Chapter 5.6. If

any experimental SRGAN model was evaluated to have a significantly lower performance than its related experimental models, two additional training runs were conducted. This was to ensure that the experimental SRGAN models had a fair evaluation and the model did not simply have an inadequate starting point due to the stochastic nature of DL algorithms. An imperative part of the training process as after conducting additional training runs of the same experimental SRGAN model, some of the performances were found to significantly improve.

5.6 Evaluation Metrics

Once all of the experimental SRGAN models were trained, their performance in generating SR images needed to be calculated and compared. In order to do so, this thesis utilizes PSNR, MSE and SSIM as the evaluation metrics used to assess the level of performance of a trained experimental SRGAN model. These are all evaluation metrics that are well-studied throughout literature and utilized within the ML community. PSNR and SSIM, in particular, are used prevalently when evaluating DL SR models as shown in Table 2.1. These evaluation metrics are considered to be standard in the industry; however, it should be taken into consideration that, even though a generated SR image may have the desired attributes of well-performing evaluation metrics (such as a high PSNR, a low MSE and a high SSIM), this does not always result in the SR image looking more similar to the desired HR, ground-truth image. This is due to the inherent ill-posed nature of SR problems as a whole. There exists many solutions to the problem: "What HR image can be generated from this particular LR image?" as the high-frequency details are undefined in the LR image and, thus, could be anything. In general, determining an effective method for quantitatively evaluating the performance of the trained models when developing GAN models, and SR techniques as a whole, is one of infamous challenges that pervades the ML community [57].

PSNR stands for peak signal-to-noise ratio. Appropriately named, it is formulated as the ratio between the maximum power possible in the input and the noise/error between the input and its corresponding ground-truth target. Typically, PSNR is expressed in terms of MSE. MSE stands for means squared error. It is formulated as the sum of the difference, between the ground-truth

target and the input, squared and divided by the length of the input. For evaluating the generated SR images, a pixel-wise difference and mean will be calculated as in Equation 5.1:

$$MSE = \frac{1}{r^2 wh} \sum_{x=0}^{rw-1} \sum_{y=0}^{rh-1} [HR(x,y) - SR(x,y)]^2 \tag{5.1}$$

where the width and height of the HR and SR images are defined as rw and rh, respectively, in order to stay consistent with the notation used earlier in the thesis. The x and y variables denote the pixel location in terms of rows and columns. The HR and SR variables denote the ground-truth target image from the testing dataset, the HR image from the testing image pair, and the super-resolved output image, generated from the LR image from the testing image pair, produced by the trained SRGAN model, respectively. With the MSE equation defined, the PSNR can now be expressed as in Equation 5.2:

$$PSNR = 10 \log_{10} \left(\frac{Max(SR(:,:))^2}{MSE(SR)} \right) \tag{5.2}$$

where $Max()$ denotes the absolute maximum intensity of $SR(:,:)$ which represents all pixels within the input SR image as the colon variable : expresses all values and combinations of x and y. From this, it can be seen that the PSNR and MSE evaluation metrics are inversely proportional to one another. These two evaluation metrics are considered to be the standard for computer vision problems as they are well-studied, have a defined physical meaning, and can be used for optimization. Even so, the ML community as a whole considers these metrics to be unsatisfactory at representing the perceptual visual quality of the images being evaluated. Both PSNR and MSE are not adept at capturing the high-frequency details that distinguish between the perceptual differences of the two images being evaluated. One of the considerations when using these evaluation metrics is that they objectively quantify the error signal's intensity without taking into account how the images differ. Two images that are distorted in different ways could be evaluated as having the same PSNR and MSE, even though one is considered to be of a higher perceptual quality. To this end, many new methods for image quality assessment (IQA) have been developed. One of these

IQA metrics that has been developed has also been used prevalently since its inception; so much so that it is becoming a standard evaluation metric used within the ML community, especially for computer vision problems. This IQA is referred to as SSIM [60]. SSIM stands for structural similarity index measure. It was developed to assess the quality of images based on the differences in their structural information as a whole. The SSIM calculation is expressed as in Equation 5.3:

$$SSIM(HR_j, SR_j) = \frac{\sum_{j=1}^{N} \left[\frac{(2\mu_{HR_j}\mu_{SR_j}+C_1)(2\sigma_{HR_jSR_j}+C_2)}{(\mu_{HR_j}^2+\mu_{SR_j}^2+C_1)(\sigma_{HR_j}^2+\sigma_{SR_j}^2+C_2)} \right]}{N-1} \tag{5.3}$$

in which HR_j and SR_j correspond to the high-resolution and super-resolution images being evaluated, respectively, for image j out of the testing dataset of size N. The SSIM metric is based on the luminance, contrast, and structure of the input images. The luminance is estimated using the mean intensity of the images μ_p which is calculated as $\mu_p = \frac{1}{N}\sum_{j=1}^{N} p_j$. The luminance is expressed as $l(p,q) = \frac{2\mu_p\mu_q+C_1}{\mu_p^2+\mu_q^2+C_1}$. The C_1 variable is a constant meant help with stability of the mean intensities. The contrast of the input images is estimated using the input images' standard deviation σ_p which is calculated as $\sigma_p = \left(\frac{1}{N-1}\sum_{j=1}^{N}(p_j-\mu_p)^2\right)^{\frac{1}{2}}$. The contrast is expressed as $c(p,q) = \frac{2\sigma_p\sigma_q+C_2}{\sigma_p^2+\sigma_q^2+C_2}$ in which the C_2 variable is a small constant. The structure of the input images is estimated by normalizing the inputs with their standard deviation. The structure comparison function is expressed as $s(p,q) = \frac{\sigma_{pq}+C_3}{\sigma_p\sigma_q+C_3}$ in which C_3 is a small constant. Combining the comparison functions together results in the expression shown in Equation 5.3. The SSIM evaluation metric has been remarked as better representing the perceptual quality of images when compared to PSNR and MSE.

These evaluation metrics are the most widely used image quality metrics within the ML community, especially the PSNR and SSIM evaluation metrics for SR tasks as shown in Table 2.1. They are all full-reference metrics meaning that these approaches require a known reference image in order to conduct their evaluation. Even though they are widely used, it is generally accepted in the ML community, especially for computer vision tasks, that there is a need for additional IQAs that better evaluate for the perceptual qualities and features present within the images themselves.

Chapter 6

Results

Chapter 6 outlines the results of the experimental models tested as well as assess their overall performance, comparing them against one another and the baseline methods. This will cover all eight of the different dataset experiments and the three model parameter experimental groups within each of those. A results table is presented that outlines the quantitative evaluations for each SRGAN model configuration. The overall highest performing evaluation for each evaluation metric will be shown in bold. For further comparison, sample SR images generated by each experimental SRGAN will be presented alongside their HR and LR counterparts. A HR image with the appropriate axes and colorbar will also be included to give context in terms of the weather radar physics. Accompanying these figures will be qualitative assessments of the SR image features, in which the ability to emulate the characteristics present within the HR, ground truth image of each SRGAN model configuration will be visually compared. The performance results for the experiments will be discussed for each experimental group. From these results, a comprehensive overview will examine any trends ascertained across the dataset experiments. Supporting sets of graphs will be supplied illustrating the model parameter's progressive effects on the performance of the SRGAN model tested. Then, the SRGAN models with the highest ranking performances will be compared with the baseline model evaluations. Additional tables and figures will be presented that layout the results for both the top performing SRGAN models and each of the baseline methods side by side. Notable strengths of the experimental SRGAN models' abilities in super-resolving weather radar scans will be supplied within the context of real-world applications. From these, the efficacy of utilizing a SRGAN model for super-resolving weather radar images will be discussed at length in Chapter 7.

As described in Chapter 5.4.2, a different set of parameters were used between the x2 resolution scaling SRGAN models and the x4 resolution scaling SRGAN models. For the x2 resolution scaling SRGAN models, the reference parameters used were a DFS of 32, a GFS of 32, and a

NRB of 16. One of these variables were then changed individually, while the other two variables were held constant at the reference value during experimentation. The DFS values were varied within the set [8, 16, 32, 64, 128]. The GFS values were varied within the set [8, 16, 32, 64, 128]. The NRB values were varied within the set [4, 8, 16, 32, 64]. Most of this is maintained for the x4 resolution scaling SRGAN models, except that the reference parameters used were a DFS of 64 and a GFS of 64. In addition, the NRB values were varied within the set [4, 8, 16, 32, 64, 128] which includes the additional NRB value of 128 for testing within the x4 SRGAN model experimental groups. The NRB value could not be set to 128 during the x2 SRGAN model experiments due to the computational resources available.

In order to demonstrate the experimental SRGANs' ability to generate SR weather radar images, a single radar scan was chosen to serve as an example. For both the RHI and PPI experiments, the sample scans were selected from their respective testing datasets. This would ensure that the image was foreign to the SRGAN models, having not been trained on the testing dataset images before, and, thus, provide a better indication as to their general generative abilities. Each of these figures consist of generated SR images of the sample scan alongside their HR and LR pairs. The layout of the sample images for each trial is presented in the same order as their corresponding tables. A HR image with the corresponding axes and colorbar is also shown to provide physical context for the weather radar scan. A meteorological analysis of the weather phenomenon present within the RHI and PPI sample scans is included below. Characteristics of these phenomenon will be further described with terminology used in image analysis such as: high-frequency and low-frequency, color boundaries, object shape, values, and artifacts.

In order to continue, the image analysis terminology must be clearly defined. Frequency is typically thought of in terms of sound or electromagnetism used to describe rapidly oscillating waves in time, such as transmitted radar signals. This time-domain oriented concept can be applied spatially to describe images. High-frequency can be used to describe areas in an image in which significant changes to the pixel intensity are quite common. In this thesis, the term high-frequency will be used to describe regions in the radar scan where there are frequent changes to the reflectivity

values. This would be shown as a region that has many changes in the representative color of reflectivity, either between multiple different colors or back and forth between a couple colors. This is primarily used for describing turbulent areas with many changes in reflectivity as well as the edges of the storm and its prominent features. In contrast, the term low-frequency will be used to describe regions in the radar scan in which the reflectivity remains consistent throughout the region. The background of each of the radar scans can also be considered low-frequency as they should be primarily white, representing no reflectivity data within that area. The term color boundaries is used to describe areas where the main color distinctly changes from one color to another. High-frequency areas in the radar scan will contain more color boundaries than low-frequency areas. The term object shape pertains to the structure and orientation of the storm/storms present within the radar scan, including the object shape of their prominent features. This term will mainly be used when comparing the SR images to their HR and LR pairs in later chapters. The term values refers to how light or dark a color is, with the lowest value corresponding to black and the highest value corresponding to white. This thesis primarily uses this term to compare and contrast between SR images in later chapters. Lighter images will be described as having higher values while darker images will be described as having lower values. The final term, artifacts, is used to describe anomalies in the generated SR images that were not present in the LR or HR image pairs. These are typically found near the edges of the SR images in the form of dark lines but can also take other forms such as black dots speckled throughout the scan or a shadow line following the edge of the storm, for example. The analyses below will use this terminology to analyze the meteorological processes portrayed by sample radar scans. These will be referenced throughout Chapters 6.1 - 6.8.

The sample RHI scan, collected by the CSU-CHIVO radar on December 14, 2018 during the RELAMPAGO campaign, was utilized for all of the RHI experiments detailed below in Chapters 6.1 - 6.4. The prominent features of both the RHI and PPI sample radar scans that will be discussed are exhibited in Figure 6.1. As a note, it is recommended that the figures present in this thesis are viewed digitally to best perceive their high-frequency details and allow for the capability of

zooming into areas of interest as the differences between some SR images are difficult to observe otherwise. A number of the observations discussed below required the use of digital zoom in order to fully perceive them. The sample scan selected contains data at the farthest extent of the radar's range, which shows how the radar beam becomes more spread as the distance from the radar increases. This is exemplified even further in the LR RHIx4_PhysRep scan shown in Figures 6.11 - 6.13. This RHI scan shows two weather systems in close proximity to one another. The first weather system that is closest to the radar contains two updrafts. Updrafts are currents of air that rise upwards and, when present within a storm, carry hydrometeors into the upper levels of the storm. This is shown as the high-frequency regions of high reflectivity within the radar scan. The closest updraft to the radar is much wider, spanning from 0 - 20 km in range. This feature will be referred to as the wide updraft. It has a relatively simple, vertically-oriented object shape with color boundaries consisting of red, orange, yellow, dark green, green, dark blue, to light blue. The other updraft is more vigorous and narrow, consisting of a concentrated area of high reflectivity around the 55 - 65 km range. This updraft will be referred to as the narrow updraft. It functions as a convective core within the first weather system due to its turbulent nature, resulting in its complex object shapes. It has color boundaries consisting of dark red, red, orange, yellow, dark green, green, to dark blue. The second weather system begins around the 70 km range. It appears to be a newly forming system at the time of collection for this RHI radar scan. This feature will be referred to as the second system. The first half of the second system is a low-frequency area from 70 - 105 km in range with color boundaries consisting of mid-green, green, to dark blue. Overall, this region has a basic object shape. The second half of the second system ranging from 105 - 140 km contains more high-frequency components lending to more complex object shapes. It has color boundaries consisting of orange, mid-yellow, yellow, dark green, green, to dark blue.

The sample PPI scan was collected on January 26, 2019 during the RELAMPAGO campaign from the CSU-CHIVO radar. The prominent features of this storm system are displayed in Figure 6.1. It was used as an example for all of the PPI experiments detailed below in Chapters 6.5 - 6.8. This sample PPI radar scan is indicative of a MCS in which multiple storms are converging

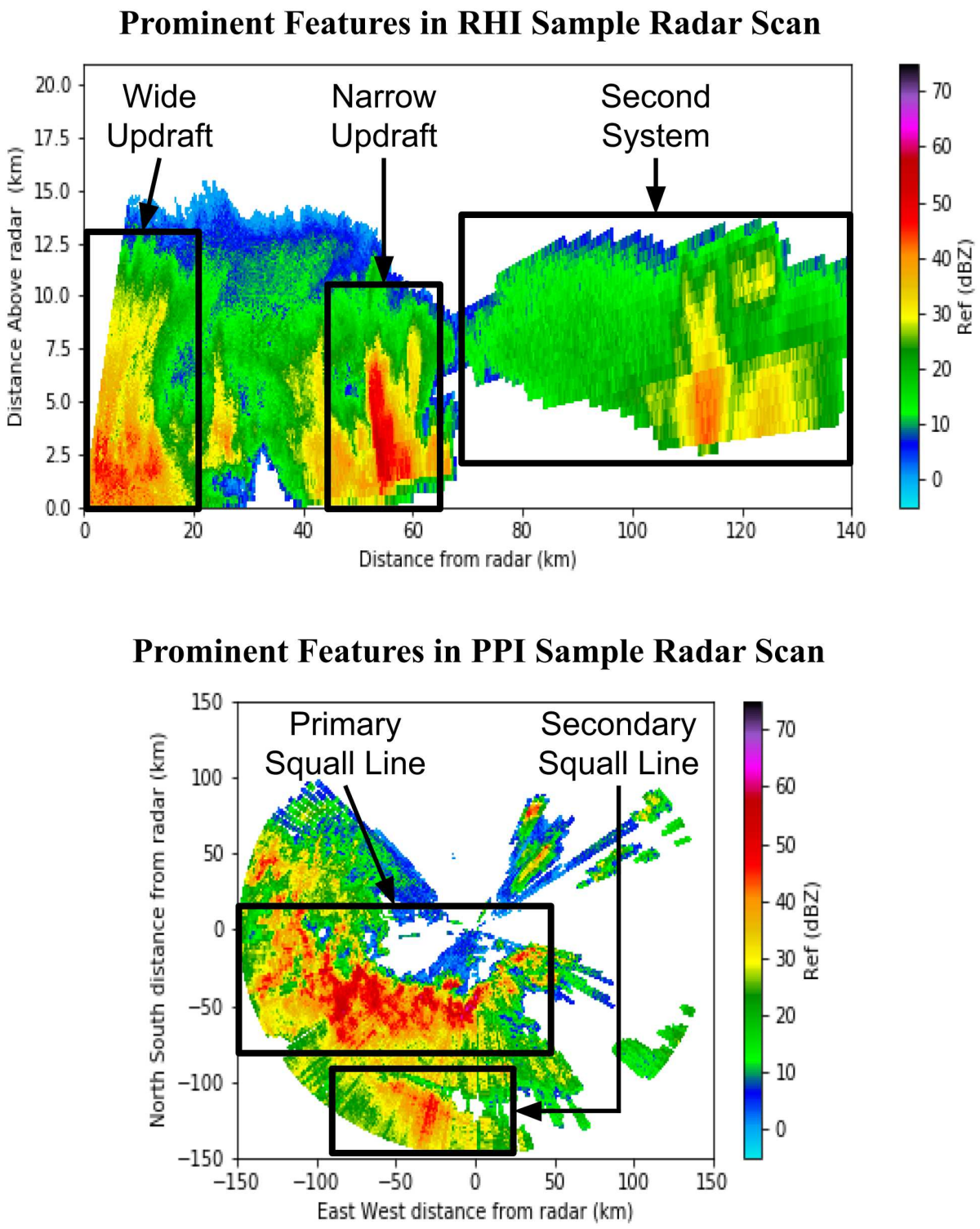

Figure 6.1: Prominent Features in Sample Radar Scans

at the same time. This system was noted as being a backbuilding, terrain-focused MCS moving along a cold front. The area of high reflectivity values reveal a long squall line, a narrow band of storms, that is moving along the cold front. This feature will be referred to as the primary squall line. The primary squall line contains many high-frequency components making up its complex object shape. It has color boundaries consisting of dark red, red, orange, yellow, dark green, green, to blue. A smaller, secondary squall line can be seen developing as well, South of the primary squall line. This feature has a more straightforward object shape containing its high-frequency components. It has color boundaries consisting of red, orange, yellow, to green.

6.1 RHIx2 Interpolation Dataset SRGAN

This experimental group consisted of SRGAN models that were trained with images from the RHIx2_Interp dataset. The results from these tests can be found in Table 6.1. Through evaluation of the DFS experiments for RHIx2_Interp, the reference model's performance, in which the DFS was set to 32, was found to be inadequate as it was the lowest performing across all evaluation metrics within its experimental group. The experimental values of 16 and 128 were the highest performing out of the DFS tests. A DFS of 16 achieved the highest SSIM of 0.905 and matched with the lowest MSE of 0.015 out of its experimental group. When the DFS was set to 128, it was also evaluated as having the lowest MSE while having the highest PSNR of 25.78 within its experimental group. However, it of interest to note that both a DFS of 8 and 16 performed higher in SSIM, with a DFS of 16 having the highest SSIM within the DFS evaluation group. The fastest training time achieved in the DFS experimentation group was for a DFS of 8 with a time of 11:43:55, written in hours:minutes:seconds. The slowest training time in the DFS experimentation group was for a DFS of 128 with a time of 14:17:34. This indicates that the DFS value has a direct affect on the training time.

From Figure 6.2, each of the sample SR images generated by the experimental SRGAN models tested have little variation as to their overall visual perceptibility. This declaration supports the results shown in Table 6.1 since each trial's evaluation is quite close in value to the rest. Even the

Table 6.1: Experimental Results SRGAN: RHI x2 Interpolation Dataset

Test Parameter	Discriminator Filter Size	Generator Filter Size	Number of Residual Blocks	PSNR	MSE	SSIM	Train Time
Reference	32	32	16	25.48	0.016	0.895	11:58:36
Dis. Filter Size	16	32	16	25.67	0.015	0.905	11:50:46
	8	32	16	25.74	0.015	0.900	11:43:55
	64	32	16	25.63	0.016	0.897	12:21:31
	128	32	16	25.78	0.015	0.897	14:17:34
Gen. Filter Size	32	16	16	25.79	0.015	0.900	11:32:55
	32	8	16	25.45	0.016	0.895	11:16:06
	32	64	16	**26.79**	**0.012**	**0.919**	13:45:03
	32	128	16	13.28	0.187	0.796	21:30:43
Number of Residual Blocks	32	32	8	25.66	0.015	0.901	11:52:17
	32	32	4	26.55	**0.012**	0.918	**11:05:59**
	32	32	32	25.48	0.015	0.901	13:02:05
	32	32	64	26.35	0.013	0.912	15:17:22

DFS 32 run, which was stated as being the lowest performing within the DFS experimental group, generated a visually similar SR image to the rest of the experimental SRGAN models tested, which demonstrates the potential of the SRGANs' generative abilities. Generally, the SR images shown in Figure 6.2 all emulate the high-frequency regions, object shapes, and color boundaries present within the HR image. It should be noted that a distinct edge artifact along the lower left corner can

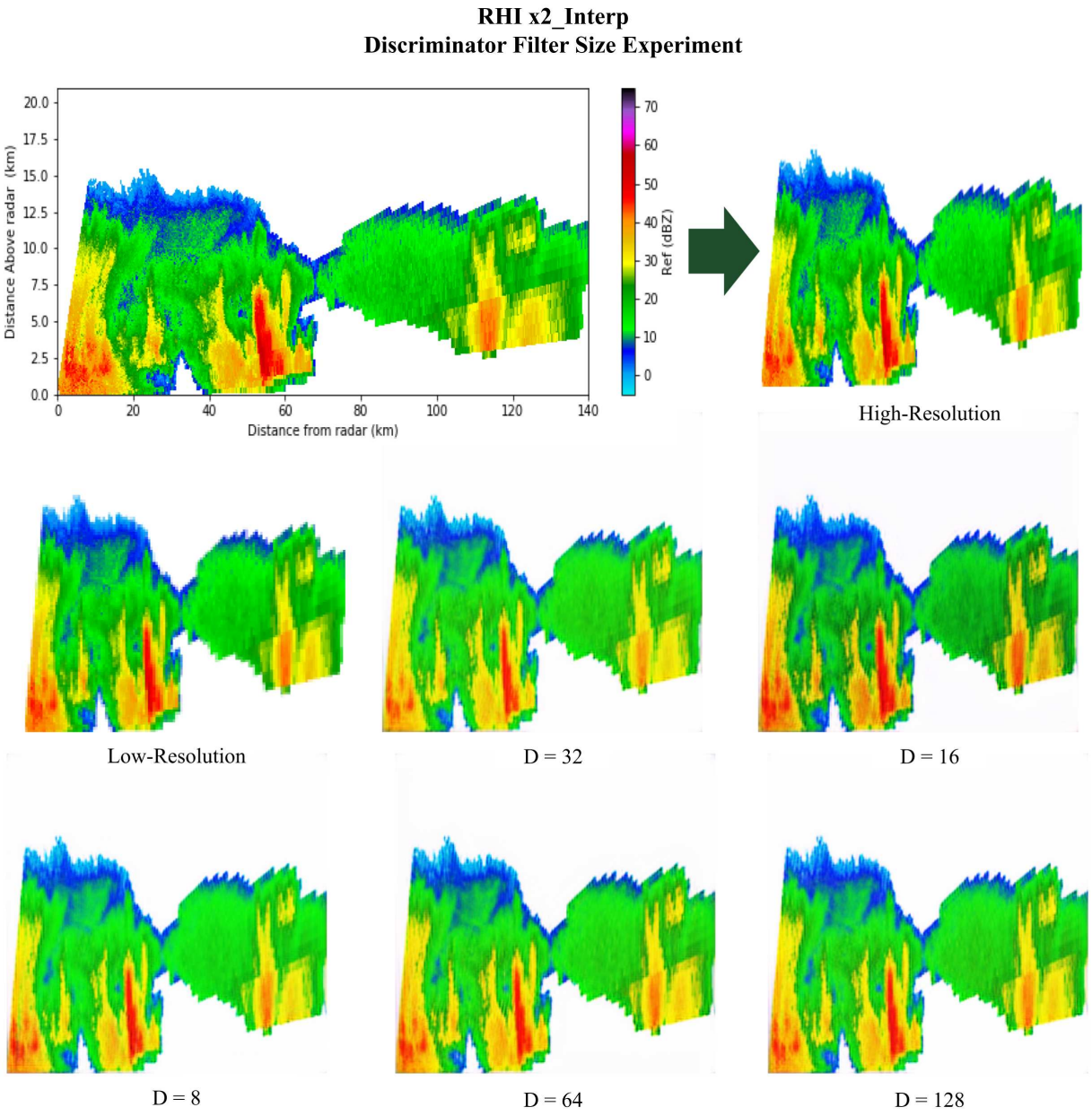

Figure 6.2: RHI x2 Interpolation Dataset: Discriminator Filter Size Experiment

be observed in every SR image, with the artifact present in the DFS 8 SR image being noticeably fainter. This suggests that edge artifacts are less likely to be present when the DFS variable is set to a lower value. The most notable differences are observed within the SR image of the DFS 16 run. The values in the DFS 16 SR image are significantly lower overall than the rest of the SR images, resulting in a darker image. Even though this characteristic is present in the areas of the radar scan that are supposed to be brighter, such as the areas around the narrow updraft and within

the second half of the second system; overall, it still appears to more closely match those in the HR image. The two updraft areas in the first storm system exemplify this assertion. In the wide updraft area, the other SRGAN models generated an image that suggests reflectivity ranging around 30 - 35 dBZ. The DFS 16 run, on the other hand, indicates reflectivity in the area ranging from 35 - 45 dBZ, which is much closer to that found in the HR, ground truth image. Likewise, the narrow updraft feature is also underestimated in the other SR images when compared to the DFS 16 run and the HR image. The object shape of this area is also thicker in the DFS 16 SR image than the other runs, which follows the HR image more closely. This observation would support the previous assertion that the DFS 16 run achieved one of the highest performances out of the DFS trials. It is interesting to note that the DFS 16 run SR image more closely resembles the HR image even when compared to the other highest performing DFS trial, the DFS 128 run. Noting that the DFS 16 run also had the highest SSIM out of all of the RHIx2_Interp DFS tests suggests that a higher SSIM could be an attribute of SR images that more closely resemble the quality of the HR target image or, at least, results in lower values in the image which, in this case, was more similar to that of the HR image.

The GFS experimental group contained a set of parameters that decisively achieved the best performance. A GFS of 64 had the highest PSNR of 26.79, the lowest MSE of 0.012, and the highest SSIM of 0.919 both within its experimental group and across all of the RHIx2_Interp trials. On the other hand, when the GFS was set to 128, the SRGAN model had a significant decrease in its performance. The GFS of 128 was evaluated to have the lowest PSNR of 13.28, the highest MSE of 0.187, and the lowest SSIM of 0.796 both within its experimental group and across all of the RHIx2_Interp trials. These results present the case of a significant difference in performance due to the GFS value being varied. Similarly to the DFS experiments, the GFS value also appears to affect the training time. The smallest GFS of 8 has the fastest training time within the experimental group of 11:16:06 while the largest GFS of 128 has the slowest training time within the experimental group of 21:30:43, which is also the slowest training time overall. These results show a correlation between GFS and training time.

Figure 6.3 contains the sample SR images generated by the RHIx2_Interp SRGAN models with the experimental GFS values. The SR image for the GFS 128 run has a significantly different background compared to the rest of the SR images, being a gray to light blue gradient instead of totally white. This drastic difference mirrors how the GFS 128 run was evaluated to have the lowest performance both within the GFS experimental group and across all RHIx2_Interp trials. Besides the GFS 128 run, the other SR images look very similar to one another, which follows the results found in Table 6.1 as their evaluations are reasonably similar. They all closely emulate the high-frequency areas, the overall object shapes, and the color boundaries of the HR image. Nevertheless, most all of the SR images contain edge artifacts along the lower left corner and the right hand side of the image, except the GFS 64 SR image. The GFS 64 SR image does not have a noticeable edge artifact whatsoever. This finding further supports the results that the GFS 64 run had the highest performance across all of the RHIx2_Interp experiments tested. The GFS 16 SR image and the GFS 64 SR image have important differences that more closely match the HR image when compared to the other SR images. They match more closely to the object shape of the narrow updraft and also have higher estimates for the reflectivity, all more equivalent to those same features found in the HR image. The wide updraft region has reflectivity ranging from 35 - 45 dBZ present in the HR image which is closely emulated in the GFS 16 SR image and the GFS 64 SR image. It is observed that the GFS 16 SR image slightly overestimates these values while the GFS 64 SR image slightly underestimates them. Furthermore, the GFS 16 run is observed to have lower values across the entire image than the rest of the SR images. The color boundaries, especially within the second half of the second system, mark the distinct variations in reflectivity present in this region in the HR image. These are represented more clearly in the GFS 16 SR image and the GFS 64 SR image whereas, in the other SR images, this area's color variations generally have higher values, washing out the variations and making them less distinct. For these reasons, the GFS 16 SR image and the GFS 64 SR image can both be asserted as having the highest visually performing generative abilities out of the GFS experimental group. This is a point of interest because the GFS 64 run was evaluated as having the highest performance due

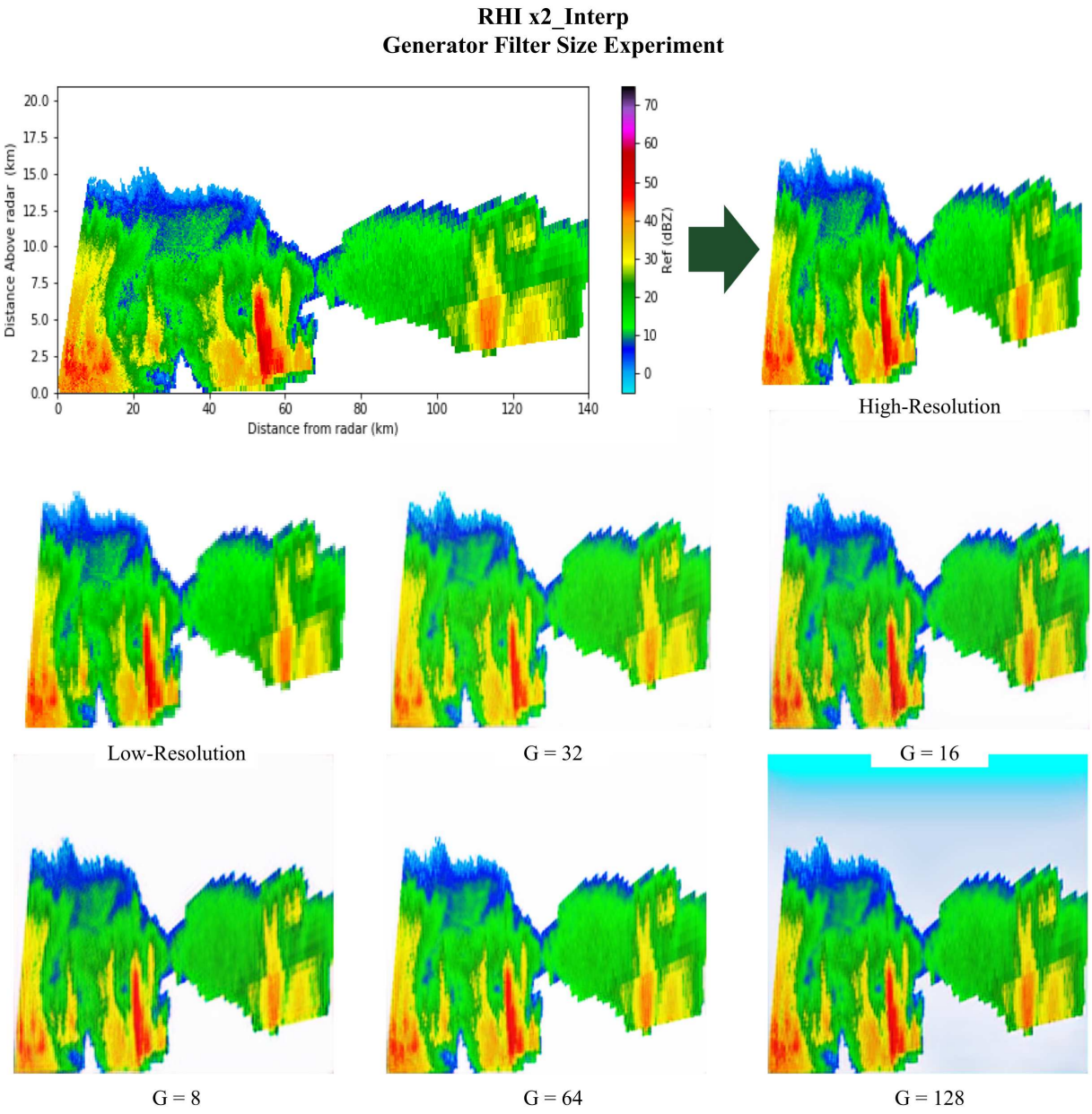

Figure 6.3: RHI x2 Interpolation Dataset: Generator Filter Size Experiment

to its evaluations, both within the GFS experimental group and overall. Likewise, the GFS 16 run had the second highest performance with the second highest PSNR, the second lowest MSE, and the second highest SSIM within the GFS experimental group. Since both the DFS 16 run and the GFS 16 run have both been established as generating quality SR images in terms of their visual representation, it could be suggested that the ratio between the filter sizes used and the low-resolution image size has a noteworthy affect on the overall quality of the SR images generated.

For the NRB parameter, the value of 4 was the highest performing within this experimental group. It resulted in having the highest PSNR of 26.55, the lowest MSE of 0.012, as well as the highest SSIM of 0.918 within its experimental group. It is also observed that a NRB of 4 was evaluated to have a MSE equivalent to the lowest overall MSE for the RHIx2_Interp experiments. In addition, the SSIM achieved by a NRB of 4 for the RHIx2_Interp SRGAN models was very similar to the highest overall SSIM accomplished by a GFS of 64, having a difference of only 0.001. The reference value of 16, the value of 8 and the value of 32 were the lowest performing NRB values within the NRB experimental group. Both a NRB of 16 and 32 were evaluated with the lowest PSNR of 25.48 within their experimental group. Similarly, both a NRB of 8 and 32 had the second highest MSE of 0.015 while the reference value had the highest MSE of 0.016 within the NRB experimental group. The reference value was found to have the lowest SSIM of 0.895 within the NRB trials while the 8 and 32 values had the second lowest SSIM of 0.901 as well. As with the other parameters, the results in Table 6.1 implies a direct connection between the NRB value and the training time. The smallest NRB of 4 has the fastest training time both within the experimentation group and overall. Its training time was 11:05:59. The largest NRB of 64 has the slowest training time within the experimentation group of 15:17:22. It can be seen that as NRB increases, the training time also increases as well.

Figure 6.4 displays the sample SR images for the RHIx2_Interp NRB experimental group. They all closely generate the high-frequency areas, the overall object shapes, and the color boundaries of the HR image, which is logical due to their close evaluations detailed in Table 6.1. Most of the SR images are visually similar to each other, with a few notable differences. The NRB 32 SR image has a marginally lower values in its background which is accentuated in the upper left and lower right corners of the image. This is supported by the evaluation results as the NRB 32 run had one of the lowest performances out of the NRB experimental group. All of the SR images have a noticeable edge artifact in the form of a dark line along the lower left corner and the right hand side of the image boundary, except the NRB 4 SR image where the artifact is significantly fainter. This could be explained by the NRB 4 having the highest performance out of the NRB

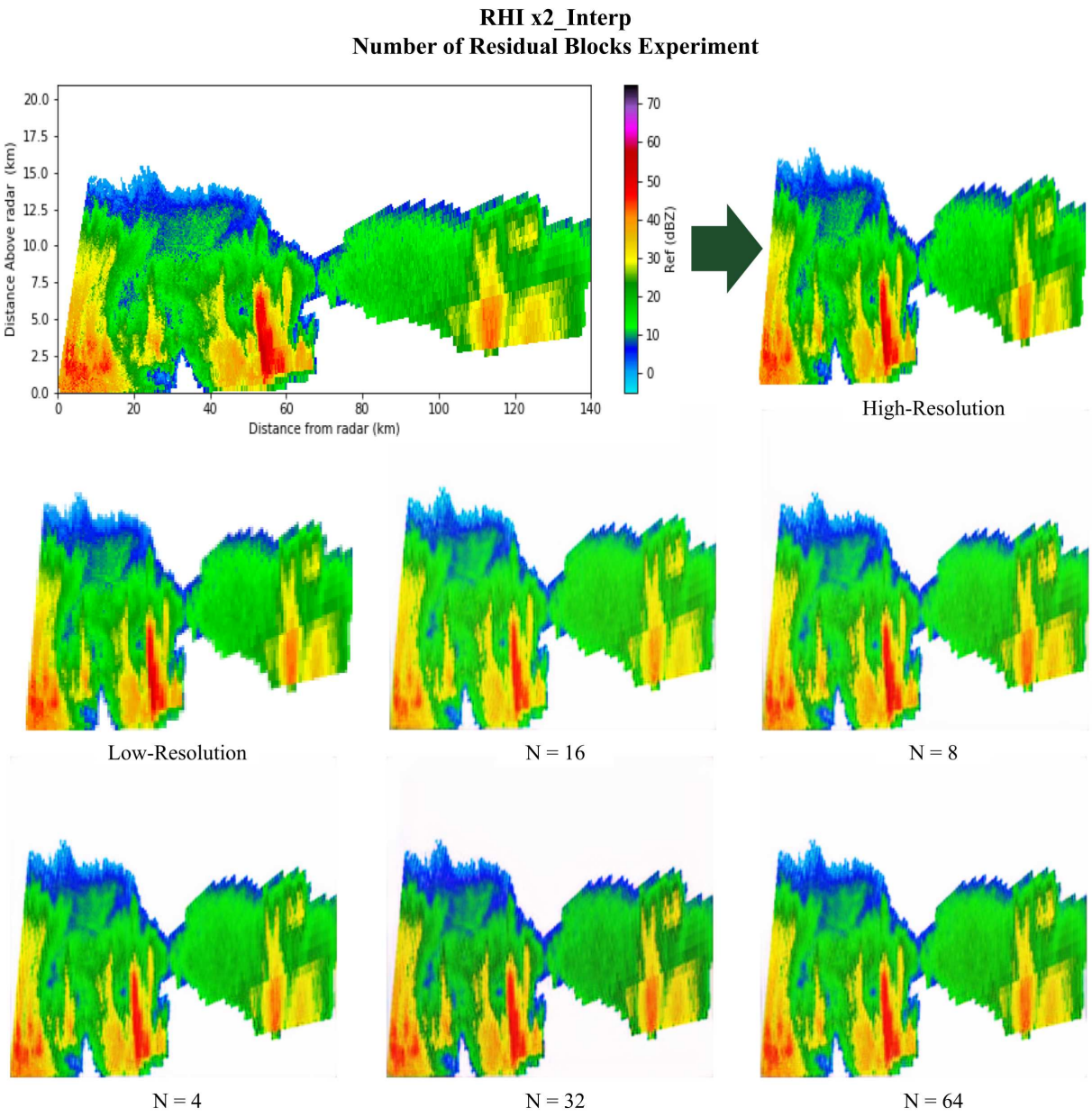

Figure 6.4: RHI x2 Interpolation Dataset: Number of Residual Blocks Experiment

experimental group. It was evaluated as having the lowest MSE overall of the RHIx2_Interp trials as well as having the second highest SSIM evaluation overall. This could suggest that lower NRB values result in more closely matching SR images. The NRB 16 SR image is observed as having the highest values overall, washing out some of the subtle variations in reflectivity present in the HR image. The NRB 32 SR image is observed as having lower values overall. It can be proposed that the NRB 4, NRB 8, and NRB 64 SR images all closely match the quality of the HR image,

with the NRB 4 SR image being the highest performing trial. The NRB 4 SR image balances the variations in reflectivity so that both the low-frequency and high-frequency reflectivity areas in the HR image are maintained. For instance, the areas around the narrow updraft and within the second half of the second system are underestimated to 28 - 35 dBZ in the NRB 16 SR image. The NRB 8, NRB 32, and NRB 64 SR images all overestimate these reflectivity areas as being 35 - 40 dBZ. The NRB 4 SR image is perceived as covering more breadth of the reflectivity range, estimating the reflectivity to be around 30 - 36 dBZ. A similar pattern of the NRB 8, NRB 32, and NRB 64 SR images generally overestimating the reflectivity, the NRB 16 SR image generally underestimating the reflectivity, and the NRB 4 SR images generally having more closely matching reflectivity predictions can be found when looking at the wide updraft feature. Furthermore, the NRB 4 SR image more fully displays the distinct color boundaries of the reflectivity variations within the first half of the second system. The generated NRB 4 SR image is observed to have retained both the low-frequency and the high-frequency reflectivity regions while the other SR images are noted to have generated either underestimated the reflectivity in general, in the case of the NRB 8 and the NRB 16 SR images, or overestimated the reflectivity in general, in the case of the NRB 32 and the NRB 64 SR images. This qualitative analysis is supported by the evaluation results. The SR images that had significantly underestimated or overestimated reflectivity values were the NRB 16 and NRB 32 runs respectively. These were evaluated as having the lowest performances within the NRB experimental group. In contrast, the NRB 4 experiment that was evaluated as having the highest performance within the NRB experimental group also was observed to have the most perceptibly similar SR image compared to the HR image. These findings suggest that setting the NRB parameter to a smaller number will have a more positive affect on the overall performance of the SRGAN model.

The lowest performing set of parameters, in the form [DFS value, GFS value, NRB value], for the RHIx2_Interp tests was determined to be: [32, 128, 16]. Its SR image was significantly altered comparatively as it had a gray to blue gradient instead of the white space background present in the HR image. The set of parameters that ranked the highest due to its evaluations was: [32, 64, 16].

Its SR image closely resembled the HR target image and, noticeably, it had less of an edge artifact than the other SR images within its experimental group. The slowest training was recorded during the [32, 128, 16] GFS experiment whilst the fastest training time was from the parameter set [32, 32, 4]. Each of the training time recorded suggested that the value of each of the parameters tested had a direct affect on the training time as increasing the parameter value also increased training time. The average training time for this SRGAN experimental group was 13:11:55.

6.2 RHIx2 Physically Representative Dataset SRGAN

This experimental group's experiments were conducted with SRGAN models that were trained on the RHIx2_PhysRep image dataset. The evaluation findings are seen as in Table 6.2. The DFS experiments for this experimental group had very little difference in the MSE evaluation for any of the DFS values tested. It was determined that the reference DFS value of 32 was the lowest performing in this experimental group as it had the highest MSE of 0.03, the second lowest SSIM of 0.864, and the second lowest PSNR of 22.96. Nevertheless, it should also be stated that the 64 DFS run had the lowest PSNR of 22.95 within the experimental group while a DFS of 16 had the lowest SSIM of 0.863 within the experimental group. When the DFS parameter was set to 128, the RHIx2_PhysRep SRGAN model achieved the best performance. Its evaluated PSNR was the highest, its MSE was the lowest, and its SSIM was the highest within the experimental group, being 23.12, 0.029, and 0.869, respectively. These observations are consistent with the findings from the RHIx2_Interp tests as a DFS value of 32 was determined to be the lowest performing model while the test where the DFS was set to 128 resulted in the best performing model. From Table 6.2, it appears that training time increases with DFS. The DFS value of 8 had the fastest training time within the experimental group being 12:10:11. The DFS value of 128 had the slowest training time of 14:51:08.

The sample SR images shown in Figure 6.5 are very similar in terms of their overall appearance. They all maintain the overall object shapes as well as the color boundaries present within the HR image. This further supports the results discussed above as their evaluations are close together

Table 6.2: Experimental Results SRGAN: RHI x2 Physically Representative Dataset

Test Parameter	Discriminator Filter Size	Generator Filter Size	Number of Residual Blocks	PSNR	MSE	SSIM	Train Time
Reference	32	32	16	22.96	0.030	0.864	12:27:07
Dis. Filter Size	16	32	16	23.08	0.029	0.863	12:19:05
	8	32	16	23.02	0.029	0.868	12:10:11
	64	32	16	22.95	0.029	0.866	12:53:22
	128	32	16	23.12	0.029	0.869	14:51:08
Gen. Filter Size	32	16	16	22.82	0.030	0.859	12:02:22
	32	8	16	22.55	0.032	0.857	**11:41:13**
	32	64	16	**23.19**	**0.028**	**0.872**	14:15:22
	32	128	16	22.88	0.029	0.867	22:03:08
Number of Residual Blocks	32	32	8	22.39	0.030	0.866	11:55:40
	32	32	4	21.59	0.033	0.865	11:48:40
	32	32	32	23.10	0.029	0.869	13:33:21
	32	32	64	22.93	0.030	0.863	15:49:16

in terms of their calculated value. However, some of the high-frequency details from the HR image are lost. The main affect of losing these high-frequency details can be seen at the farther ranges of the radar scan in the second half of the second system. It is within this region that the affects of the physically representative downsampling method are even more apparent. The resulting LR image is more sectioned within this area, resulting in generated SR images that are

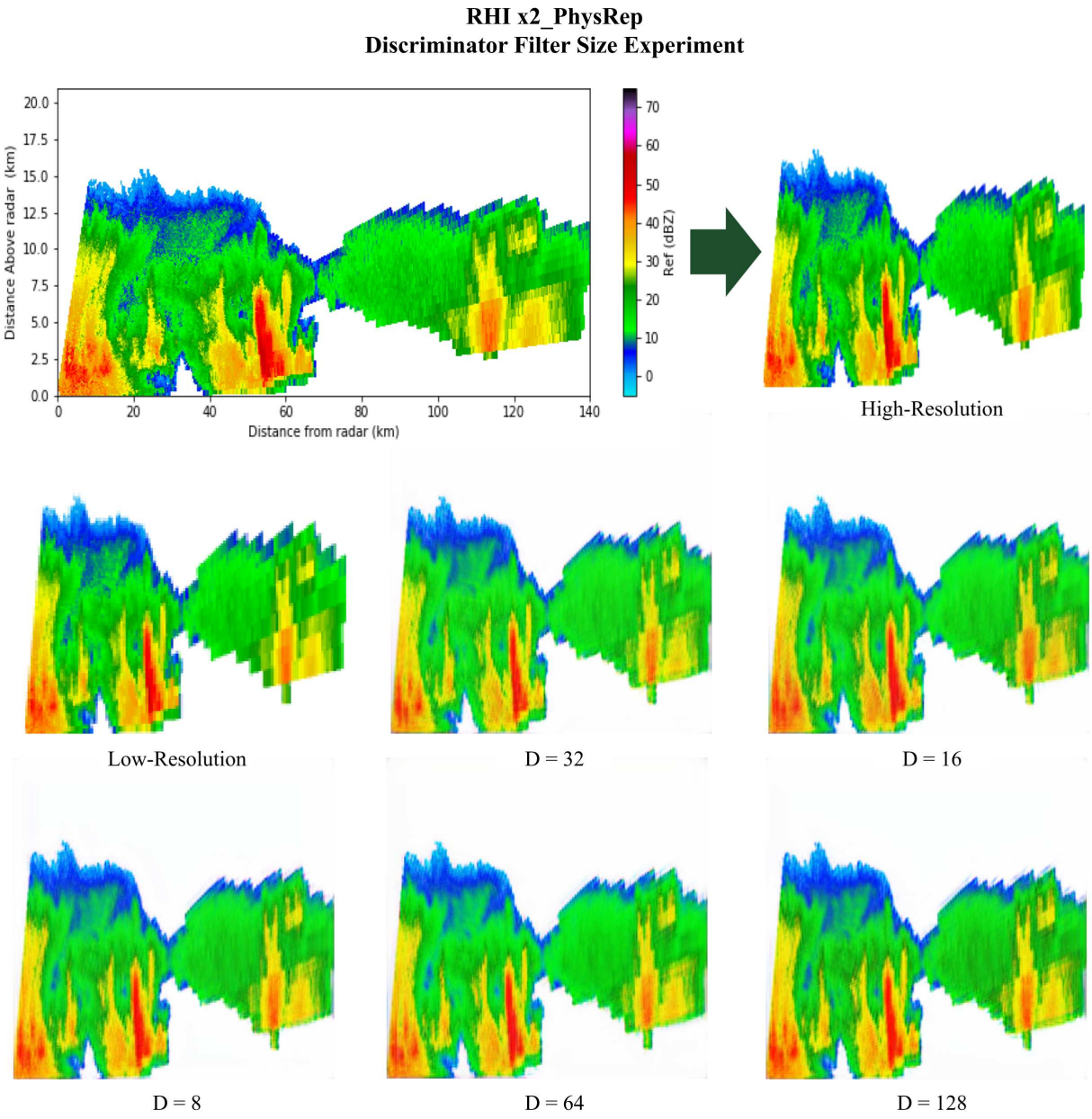

Figure 6.5: RHI x2 Physically Representative Dataset: Discriminator Filter Size Experiment

more blurred with less distinct reflectivity variations at the farther ranges of the radar scan. This can also be observed in the first storm system between 10 - 15 km in height in which the distinct transitions between reflectivity values in the HR image are not fully reconstituted in the sample SR images. Nevertheless, the SRGAN models tested are able to reconstruct enough of these details to preserve most of the pertinent information, such as the overall object shape, range, height, and color boundaries present in the active storm areas of the HR image. All of the SR images are observed

as having edge artifacts in their lower left corners, although the DFS 8 SR image's artifact is much fainter in comparison. This could indicate that a lower DFS value results in less significant edge artifacts. The DFS 32 and DFS 64 SR images are observed to have noticeably higher values overall than the other SR images while the DFS 8, DFS 16, and DFS 128 SR images have noticeably lower values overall, which is perceived as matching the HR image more closely. Overall, the DFS 128 SR image is observed as having been the closest to have a balance of the total array of reflectivity represented, displaying more of the minute, high-frequency variations in reflectivity present in the HR image. This is exemplified when examining the first half of the second system. The DFS 32 and DFS 64 SR images tend to underestimate the reflectivity in this area. The DFS 8 and DFS 16 SR images seem to maintain some of the darker color boundaries, but, in general, they are still observed to underestimate the reflectivity in this area as well. The DFS 128 SR image preserves the high reflectivity present in this region, especially the bold green color boundary within the second half of the second system. It should be noted that the variations in reflectivity present within this area are more difficult to distinguish due to the higher values of the DFS 32 and DFS 64 SR images. All of the SR images are relatively similar with how the narrow updraft is generated, both in terms of the color boundaries and the object shape. Further differences can be found in wide updraft as well as the second half of the second system. In both of these areas, the DFS 32 and DFS 64 SR images tend to have higher values than the HR image, the DFS 16 SR image tends to have lower values than the HR image, and the DFS 8 and DFS 128 SR images are observed to match the HR image more closely, displaying a more representative array of reflectivity overall. When compared visually, the SR images with lower values appear to be more similar to the HR image than the SR images with higher values. These observations further support the results presented in Table 6.2. The SR images with higher values, DFS 32 and DFS 64, were recorded as having the lowest performances within the DFS experimental group. The SR images with lower values, DFS 8, DFS 16 and DFS 128, were all evaluated as having higher performances. The SR image that was determined to be the closest to the HR target image was derived from the DFS 128 experiment. This was also the experiment that had the highest performance out of the DFS experimental group.

Out of the other two experimental trials, the DFS 8 SR image had less of an edge artifact and was able to more closely generate the wide updraft and the second half of the second system. Following this reasoning, it could be argued that the DFS 8 SR image was able match more closely to the HR image overall than the DFS 16 SR image. Both the DFS 8 and DFS 16 experimental tests performed to a similar degree in their evaluations, with the DFS 16 run having a higher PSNR and the DFS 8 run having a higher SSIM. This suggests that the SSIM evaluation metric may be more indicative of higher quality SR images.

When the GFS was set to 64, the RHIx2_PhysRep achieved its best performance overall. The 64 GFS test attained the best results on all of the evaluation metrics both within its experimentation group and when compared to every other parameter combination tested for the RHIx2_PhysRep experiments. It performed with a PSNR of 23.19, a MSE of 0.028, and a SSIM of 0.872. It should be acknowledged that this is the same combination of parameters that performed the highest for the RHIx2_Interp experimental trials as well. Alternatively, when the RHIx2_PhysRep SRGAN model had its GFS set to 8, it was the lowest performing within its experimental group. The lowest PSNR, the highest MSE, and the lowest SSIM out of this experimentation group were all evaluated from the run with a GFS of 8. Its PSNR was 22.55, its MSE was 0.032, and its SSIM was 0.857. The results in Table 6.2 indicate a correlation between GFS and training time. The smallest GFS of 8 had the fastest training time within the experimental group of 11:41:13. This was the fastest training time for the entirety of the RHIx2_PhysRep experimental runs. At the same time, the largest GFS of 128 has the slowest training time, both within the GFS experimental group and across all of the RHIx2_PhysRep tests, of 22:03:08.

Example SR images for the RHIx2_PhysRep GFS experimental trials are given in Figure 6.6. The SR images are all quite similar to one another, which supports the results shown in Table 6.2 as they all had similar evaluations. All of the SR images emulate the general object shapes and color boundaries of the HR image; however, the object shapes are more influenced by the physically representative LR input image. The full affect of this downsampling method can be seen in the high-frequency details of the second half of the second system. The edges of the

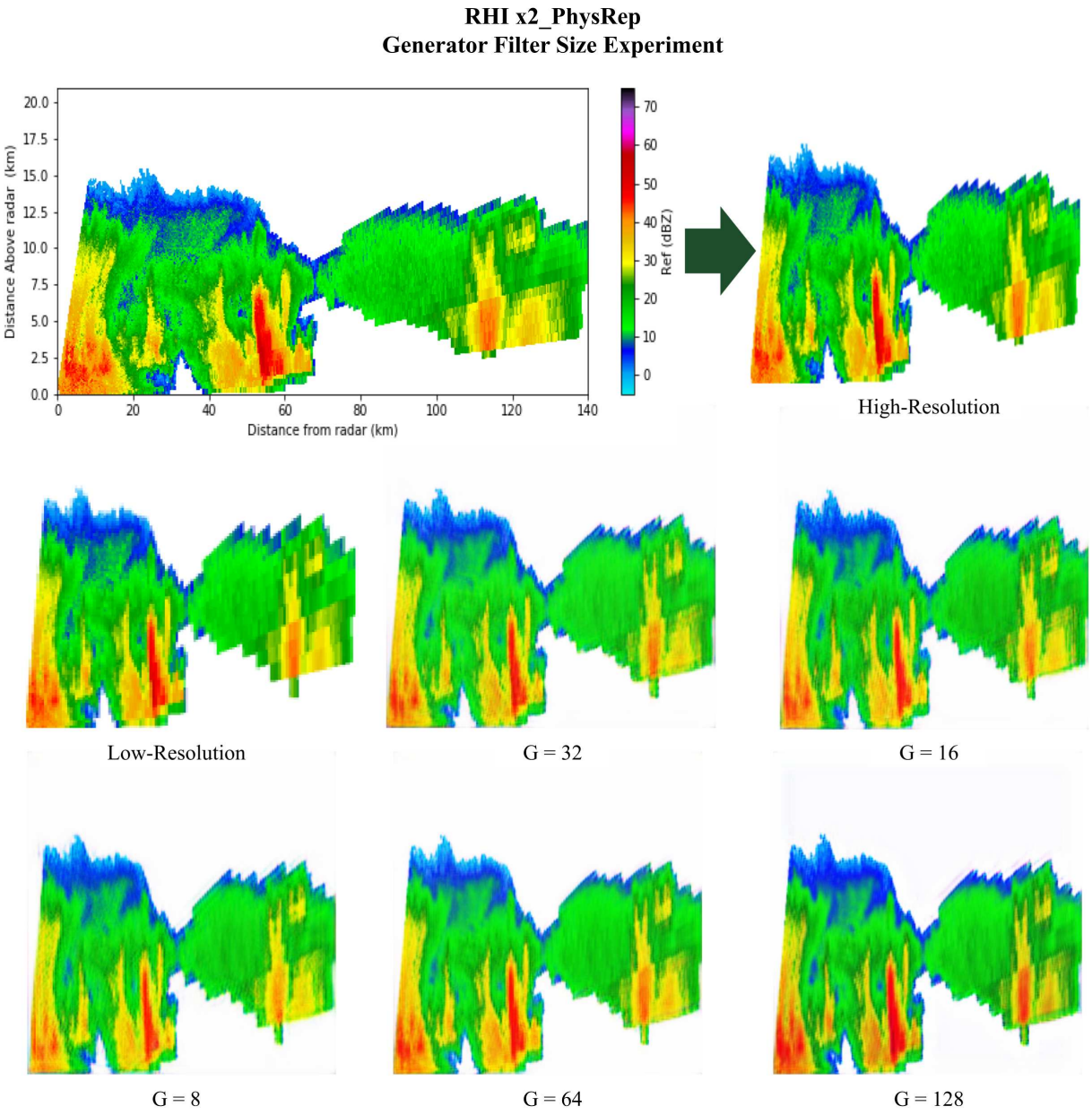

Figure 6.6: RHI x2 Physically Representative Dataset: Generator Filter Size Experiment

storm in the SR images follow the object shape of the LR image rather than the HR image. The high-frequency components within this area of the storm are distorted in the SR images, appearing to be more blurred and pixelated. Although the color boundaries are mostly preserved, the distinction between the boundaries is less clear at the farther ranges of the radar scan. This is a significant difference from the SR images generated from the interpolation dataset. In addition, the GFS experimental SR images appear to be affected by the downsampling method more than

the DFS experimental SR images. All of the SR images are also affected by edge artifacts. The edge artifacts are present in the form of dark lines in the lower left corner as well as shadow lines following along the storm edge in the upper region of the second system. The GFS 64 trial has a much fainter corner artifact while the GFS 32 trial has a much fainter storm edge artifact. The GFS 128 has the lowest values overall which matches more closely to the HR image and results in well-defined color boundaries; however, the background is predominantly more gray and the reflectivity is overestimated in general. For these reasons, the GFS 128 trial is perceived as having the lowest performance in its generative abilities out of the GFS experiments, which implies that increasing the GFS of an SRGAN model results in low quality SR images. This does not reflect its performance determined by the evaluation metrics as it had the second lowest MSE and the second highest SSIM out of its experimental group. In contrast, the GFS 8 and GFS 32 SR images both have higher values overall. This makes the SR images appear more faded in coloration when compared to the HR image. The color boundaries are less distinct and the reflectivity is also underestimated in these SR images, which is exemplified in the wide updraft region. These observations support the evaluations for the GFS 8 run as it was determined to have the lowest performance out of the GFS experimental group. The GFS 32 run, on the other hand, performed quite well on its evaluations overall. The GFS 16 and GFS 64 SR images are perceived as being the closest to the HR image, perceptually. The GFS 16 SR image has slightly lower values, especially within the midtones of the image, which results in more well-defined color boundaries in the second system. But, in terms of visual appearance, the GFS 64 SR image has a closer resemblance to the HR image as it is less affected by the corner artifact, it retains the high-frequency components in the second half of the second system which was most affected by the downsampling method, it appears to be less pixelated, and it appears to better predict the updraft regions as well. This qualitative analysis supports the results found in Table 6.2. The GFS 64 trial had the highest performance out of all of the RHIx2_PhysRep experiments. An interesting note is that the GFS 16 run had the second lowest performance which does not reflect its visual comprehension, similar to the GFS 32 and GFS 128

trials. This suggests that the evaluation metrics do not consistently reflect the perceptibility of the SR images generated except, perhaps, in the extreme cases.

The NRB value that proved to be the best performing within its experimentation group was a NRB of 32. It was evaluated to have a PSNR of 23.10, a MSE of 0.029, and a SSIM of 0.869. While the value of 64 for the NRB tests had the lowest SSIM of the NRB experiments, a NRB of 4 had the lowest ranking in terms of its evaluation. The PSNR for the NRB 4 test was calculated to be 21.59 while its MSE was calculated as 0.033, the lowest PSNR and highest MSE, respectively, both within it experimentation group and across all of the RHIx2_PhysRep experiments. These findings disagree with the previous RHIx2_Interp results on the NRB variable tests. The results in Table 6.2 suggest that the training time is directly impacted by the NRB value. The smallest NRB value of 4 had the fastest training time of 11:48:40 within the NRB experimentation group. Meanwhile, the largest NRB value of 64 had the slowest training time of 15:49:16 within the NRB experimentation group.

Figure 6.7 displays the SR images of the NRB experimental group. Similar to the other RHIx2_PhysRep trials, the NRB tests were able to reconstruct the overall object shapes and color boundaries of the HR image but had more difficulty generating the high-frequency components, especially within the second system. This makes the SR images appear more blurred and pixelated. The NRB experiments are similar to the GFS experiments in that they are more affected by the downsampling method than the DFS experiments. Overall, the visual representation between the SR images is quite similar, which supports the evaluation results. It is apparent, however, that the NRB 4 and NRB 8 tests have a primarily gray background as opposed to the target's white space background, lowering their overall visual quality. These two SR images also have a distinct shadow line artifact along the storm edge in the second system. The second half of the second system in these two NRB test SR images are noticeably more blurred and pixelated than the other SR images. These observations support the evaluation results for the NRB tests. The NRB 4 and the NRB 8 experimental SRGAN models had the lowest performances out of all of the RHIx2_PhysRep experiments. It is of interest to note that, even though they both did not perform well on the PSNR

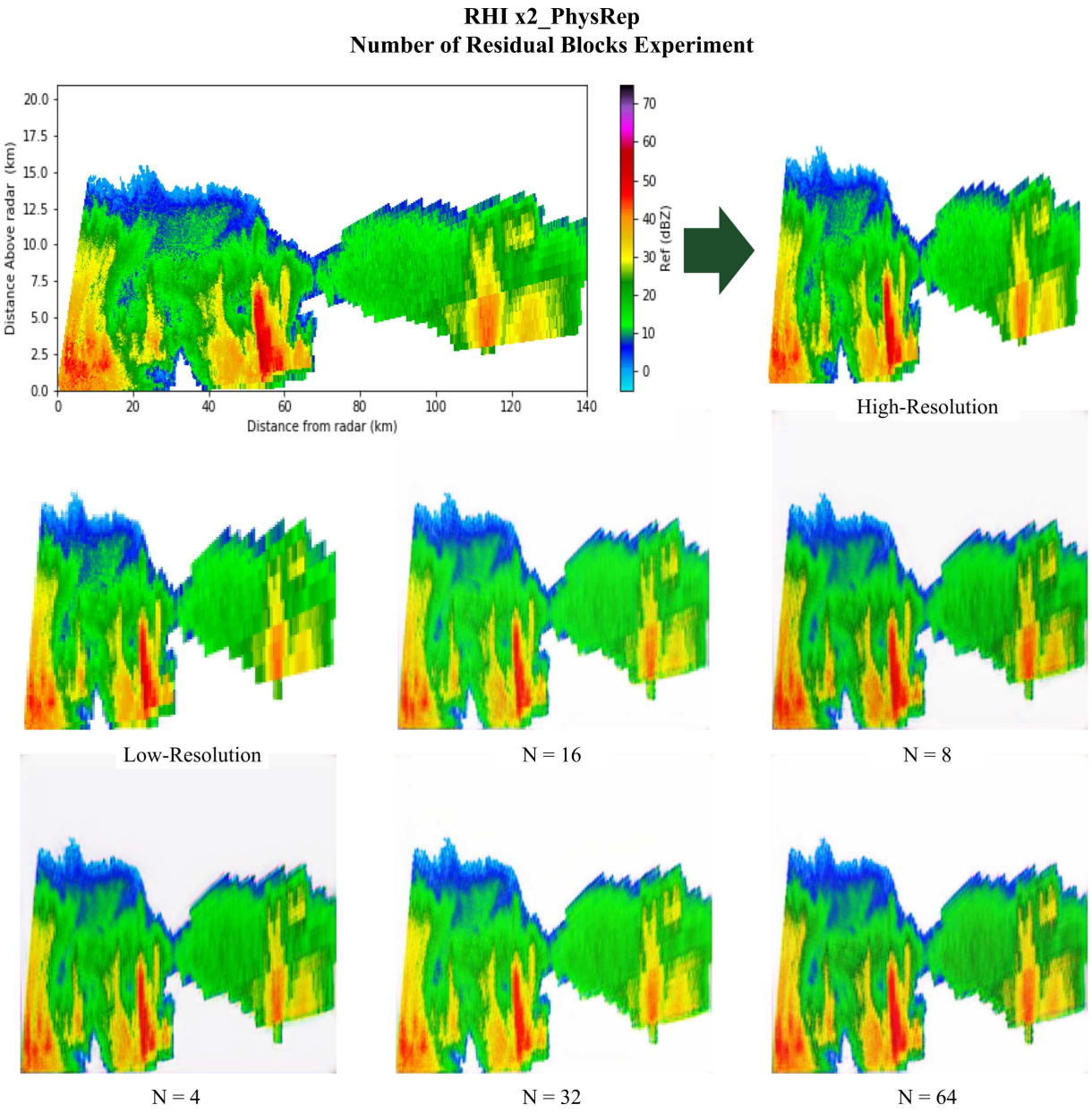

Figure 6.7: RHI x2 Physically Representative Dataset: Number of Residual Blocks Experiment

and MSE evaluation metrics, they both had fairly competitive results on the SSIM metric. This may suggest that the SSIM evaluation metric is not as affected by the low-frequency regions of the SR images. All of the SR images also contain a dark line artifact in the lower left corner and the right hand side of the image. The NRB 32 SR image has the faintest of these artifacts. This is supported by the evaluation results as the NRB 32 run was recorded as being the best performing experimental SRGAN model out of the NRB experimentation group. The NRB 64 SR image

contains lower values, the NRB 16 SR image contains higher values, and the NRB 32 SR image appears to be more balanced overall and is the closest in matching the HR image visually. In general, they have similar object shapes except that the NRB 16 SR image has slightly thinner object shapes and the NRB 64 SR image has slightly thicker object shapes. This can be seen in the narrow updraft area as well as in the second half of the second system. The NRB 16 SR image appears to underestimate the reflectivity within the updraft regions while the NRB 64 SR image appears to overestimate. Even though the color boundaries are more defined in the NRB 64 SR image, the high-frequency details are not as distinct when compared to the NRB 32 SR image. This trend is also observed in the second system as well. This analysis further supports the evaluation results. The NRB 16 and NRB 64 runs had quite similar results on their evaluations while the NRB 32 run had the highest performance out of its experimental group. These statements conflict with the analysis presented for the RHIx2_Interp experiments in which it was suggested that a small NRB value would produce more quality SR images. In addition, even though the NRB 64 run was evaluated to have the lowest SSIM out of the NRB tests conducted, it also had the lowest values overall. This also disagrees with previous suggestions that a higher SSIM is indicative of an SR image with lower values.

The lowest performing set of parameters for the RHIx2_PhysRep experiments was determined to be [32, 32, 4]. The set of parameters that ranked the highest due to its evaluations was: [32, 64, 16]. This is the same set of parameters that had the highest ranking performance for the RHIx2_Interp tests. The slowest training was recorded by the [32, 128, 16] test. The GFS value of 128 also had the slowest training time for the RHIx2_Interp tests. The fastest training time was achieved by [32, 8, 16]. The results from the RHIx2_PhysRep tests further supports that the training times increase as the parameter values increase. The average training time for this SRGAN experimental group is 13:40:46. It should also be noted that all of the RHIx2_PhysRep SRGAN models were not able to reconstruct the high-frequency components within the second system to the same degree as the RHIx2_Interp SRGAN models due to the downsampling method used.

6.3 RHIx4 Interpolation Dataset SRGAN

This experimental group contains the experimental trials for the RHIx4_Interp SRGAN models. The results for these experiments are recorded in Table 6.3. There was a clear DFS value that performed to the highest degree across all evaluation metrics both within the DFS experiment group and for all of the parameters tested for the RHIx4_Interp model. The DFS of 16 obtained a the overall highest PSNR of 24.70, the overall lowest MSE of 0.018, and the overall highest SSIM of 0.890. This SSIM was also achieved by the reference DFS value of 64. This is in contrast to the performance of the DFS value of 8 which had the lowest PSNR of 24.00, the highest MSE of 0.022, and the lowest SSIM of 0.870 within the DFS experimental group. The fastest training time within the experimental group was 13:24:27 which was achieved when the DFS parameter was set to 8. Following a similar trend to the previous experimental groups, the largest DFS value of 128 also had the slowest training time recorded for the DFS experiments of 16:06:07. These results also indicate that the value of the DFS parameters directly affects training time.

The SR images shown in Figure 6.8 are remarkably similar in terms of their overall visual appearance. They all maintain the general object shapes and color boundaries present within the HR image. This follows the results found as their evaluations are reasonably similar. It is also observed that each of the RHIx4_Interp DFS SR images are perceptually more similar to the HR image than their RHIx2_Interp DFS counterparts, especially in terms of values. This is an interesting observations as the experimental RHIx2_Interp SRGAN models had higher performances on the evaluation metrics than the experimental RHIx4_Interp SRGAN models. This could imply that the evaluation metrics do not consistently correspond to visual comprehension. Some of the high-frequency components are distorted in the SR images, especially within the upper levels of the storm and within the second half of the second system. Furthermore, it should be noted that the object shape of the area of high reflectivity in the second half of the second storm appears to be significantly thinner in all of the SR images when compared to the HR image. This might be explained by the four times resolution dataset utilized for these experiments. Further tests will have to analyzed to see if a trend emerges where utilizing the datasets with increased resolution

Table 6.3: Experimental Results SRGAN: RHI x4 Interpolation Dataset

Test Parameter	Discriminator Filter Size	Generator Filter Size	Number of Residual Blocks	PSNR	MSE	SSIM	Train Time
Reference	64	64	16	24.53	0.019	**0.890**	14:29:24
Dis. Filter Size	32	64	16	24.23	0.021	0.882	13:56:22
	16	64	16	**24.70**	**0.018**	**0.890**	13:49:17
	8	64	16	24.00	0.022	0.870	13:24:27
	128	64	16	24.26	0.020	0.883	16:06:07
Gen. Filter Size	64	32	16	24.22	0.021	0.882	12:35:11
	64	16	16	23.92	0.022	0.874	12:04:59
	64	8	16	23.06	0.027	0.856	**11:46:14**
	64	128	16	4.68	1.357	0.252	24:05:46
Number of Residual Blocks	64	64	8	24.50	0.019	0.887	13:17:36
	64	64	4	15.63	0.239	0.768	13:57:54
	64	64	32	24.21	0.020	0.884	14:57:47
	64	64	64	18.72	0.092	0.796	17:17:38
	64	64	128	15.68	0.239	0.769	23:59:43

scaling factors results in SR images with more distorted high-frequency components. All of the SR images contain an edge artifact in the lower left corner and along the right hand side of the images. The DFS 64 run has the least noticeable corner artifact while the DFS 32 run has the faintest side edge artifact. At the same time, the DFS 128 SR image has a slightly gray background. This

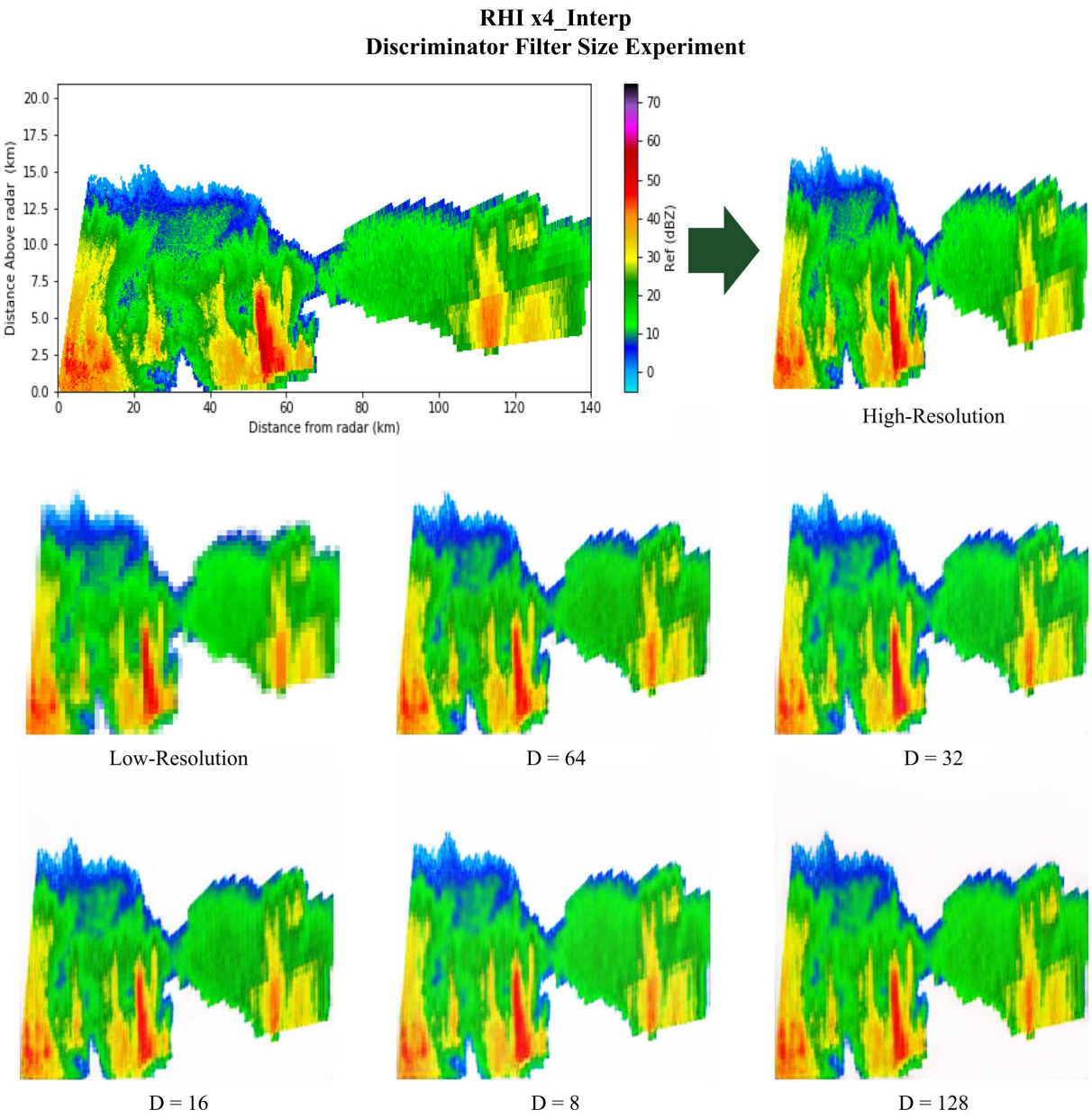

Figure 6.8: RHI x4 Interpolation Dataset: Discriminator Filter Size Experiment

could suggest that increasing the DFS parameter in the SRGAN model tested leads to distortions in the low-frequency regions of the SR image generated. Although this decreases its visual quality, this observation is not reflected in its evaluations as the DFS 128 test performed adequately on all evaluation metrics. The most notable difference between the RHIx4_Interp SR images is that the DFS 8 SR image has slightly higher values, underestimating the reflectivity values in most of the storm's regions overall. The evaluation results concur with the observation that the DFS 8 SR im-

age is the most different from the HR image out of the DFS experimental group. The DFS 128 also has slightly higher values, primarily within the mid-tone green color boundaries as well as the low-frequency components of the second storm system. The DFS 16 and DFS 64 SR images appear to have slightly lower values in general, which is exemplified in the second system, bearing a more similar resemblance to the HR image than the DFS 8 and DFS 128 SR images. It is interesting to note that, while the DFS 16 run had the highest performance both within the DFS experimental group and overall and the DFS 64 run had an equivalent SSIM evaluation, the DFS 32 SR image has the closest visual similarities to the HR target image. The DFS 16 SRGAN model was able to generate the second system quite comparably to the HR image, especially in terms of its object shapes and high-frequency details, but it generally underestimates the reflectivity in the updraft regions. On the other hand, the DFS 64 SR image estimates the reflectivity in the updraft regions well and generally maintains the color boundaries present within the HR image. However, the DFS 64 SR image has distorted object shapes and overestimates the reflectivity within the second half of the second system. The DFS 8, DFS 64, and DFS 128 SR images all have bifurcated object shapes at the top of the area of high reflectivity around 125 km in range and 4 - 7 km in height. The SR image that is the most similar representation of the HR image was produced by the DFS 32 SRGAN model. It appears to have balanced the positive characteristics of the other models. The DFS 32 SRGAN model yielded proper estimations for the reflectivity present in the updraft areas as well as the second system. The color boundaries are well-defined, the high-frequency details are distinct, and the object shapes appear to match those seen in the HR image. It is important to note that these analyses do not directly follow the evaluation results which could imply that the evaluation metrics are alone not sufficient enough to determine the visual quality of an image.

The reference value of 64 for the GFS proved to have the highest ranking within the experimental group in terms of its evaluations. Its PSNR of 24.53, MSE of 0.019, and SSIM of 0.890 were unrivaled by any other GFS value tested. The SSIM attained by the reference value was also the highest SSIM across all of the RHIx4_Interp experiments. On the other hand, the GFS value of 128 performed at the lowest degree both within the GFS experimentation group and overall. It

was evaluated to have the lowest PSNR of 4.68, the highest MSE of 1.357, and the lowest SSIM of 0.252. This also was the GFS value that had the slowest overall training time of 24:05:46. This is over double the fastest overall training time achieved by the GFS value of 8 which clocked in at 11:46:14. These results also suggest that increasing the GFS value also increases training time. Another trend noted is that the evaluations improve as GFS increase until the GFS is set to 128. At that point, the RHIx4_Interp model's performance degrades dramatically.

Figure 6.9 displays the SR images generated by the RHIx4_Interp GFS experimental SRGAN models. At first glance, it is evident that the GFS 128 SR image does not accurately portray the same information as in the HR image. The GFS 128 SRGAN model for this dataset experiment generated an entirely bright blue image. None of the high-frequency details, color boundaries, nor the object shapes have been reconstituted in the SR image. This supports the evaluation results that determined the GFS 128 run to be the lowest performing model out of all the RHIx4_Interp experimental SRGAN models. This further supports the previous suggestion that a higher GFS results in low quality SR images. Otherwise, the rest of the SR images are reasonably similar which supports the evaluation results. The GFS 8 SR image contains significantly higher values overall. This lead to the radar scan's reflectivity being underestimated, especially within the regions with high-frequency components. In addition, the high-frequency details have become distorted, being more blurred overall and obscuring the object shapes. Comparably, the GFS 8 SR image is also not accurately portraying the same information as in the HR image. This follows the evaluation results as the GFS 8 run was the second lowest performing out of the GFS experimental group. This could suggest that a smaller GFS also results in lower quality SR images. The rest of the GFS experimental SRGAN models generated SR images that are representative of the HR target image. All of these SR images contain edge artifacts in the lower left corner with the GFS 16 SR image having the most noticeable artifact. It has higher values in general and underestimates the reflectivity present within the wide updraft and the second half of the second system. Out of the remaining three, the GFS 16 run was evaluated as being the lowest performing, which reflects its visual comparison. The GFS 64 SR image has more defined color boundaries and better estimates

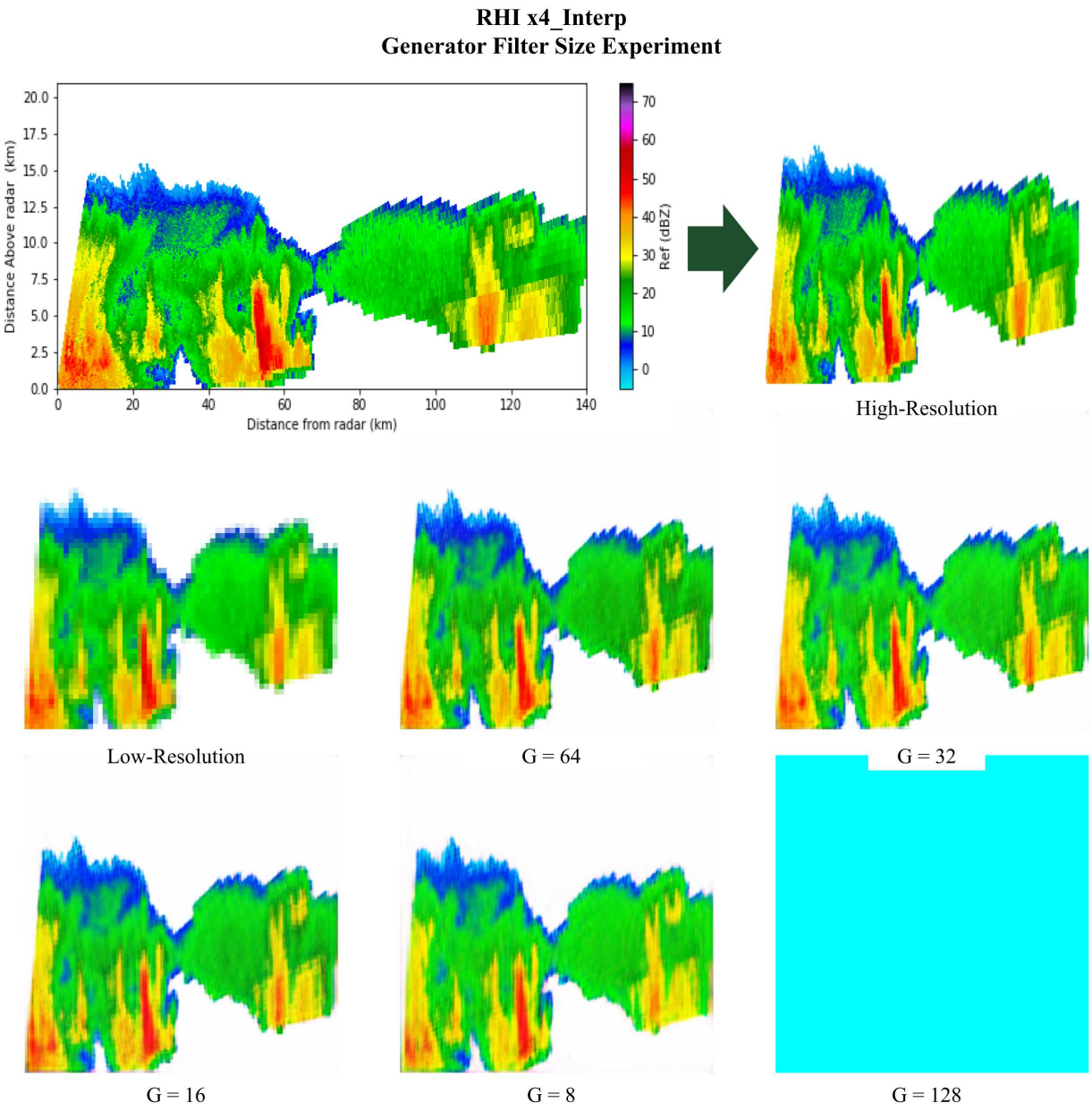

Figure 6.9: RHI x4 Interpolation Dataset: Generator Filter Size Experiment

the reflectivity in the updraft areas when compared to the GFS 32 SR image. At the same time, the GFS 64 SR image generally has lower values, which is exemplified in the low-frequency region of the second system, resulting in overestimation of the reflectivity especially within the green color boundary of both storm systems. The GFS 32 SR image better estimates the reflectivity in these brighter color boundary regions. It also retains the object shape of area of high reflectivity around the 125 km range in the second half of the second system. However, the GFS 32 SR image appears

to underestimate the reflectivity in the updraft areas of the first storm and does not fully represent the dark green color boundaries in general. In terms of their evaluation results, the GFS 64 run had a higher performance on every evaluation metric. Nevertheless, both the GFS 32 and the GFS 64 SR images have quite comparable perceptibility. These analyses further support the idea that the trial that had a higher SSIM, the GFS 64 run, also had an SR image with noticeably lower values. In addition, the GFS 64 run also produced higher quality SR images for the previous data experiments as well which should be noted for further analyses.

In a similar manner to the GFS experiments, the NRB experiments revealed that the reference parameter set was able to achieve the highest performance within the NRB experimentation group. Upon examining Table 6.3, it can be observed that the evaluation performances improved as NRB increased until the NRB reached 16, after which, the performances started to decline. The lowest ranking set of parameters in terms of their evaluation was with the NRB value of 4, with 128 being a close second lowest. The NRB value of 4 was evaluated to have a PSNR of 15.63, a MSE of 0.239, and a SSIM of 0.768. However, the 128 NRB experiment had the slowest training time within the experimental group. Its training time was recorded as 23:59:43. The fastest training time for the NRB experiments was the NRB 8 test with a training time of 13:17:54.

Example SR images for the RHIx4_Interp NRB experimental trials are given in Figure 6.10. It is quite evident that the NRB 4, NRB 64, and NRB 128 experimental SRGAN models are not suitable for generating physically representative SR images. The NRB 64 SR image has oversaturated colors to the point of being neon in color. Most of the high-frequency details cannot be discerned whatsoever. This further supports the evaluation results present in Table 6.3 as the NRB 64 run had one of the lowest performances both within the NRB experimental group and overall. However, it was evaluated as having a higher performance than both the NRB 4 and the NRB 128 tests. This is not reflected in their generated SR images. Most of the information present in the HR image can not be determined from the NRB 64 SR image whereas, in the NRB 4 and the NRB 128 SR images, some of the general information such as basic object shapes, relative color boundaries, and some high-frequency components from the HR image are observable. Both the NRB 4 and NRB 128

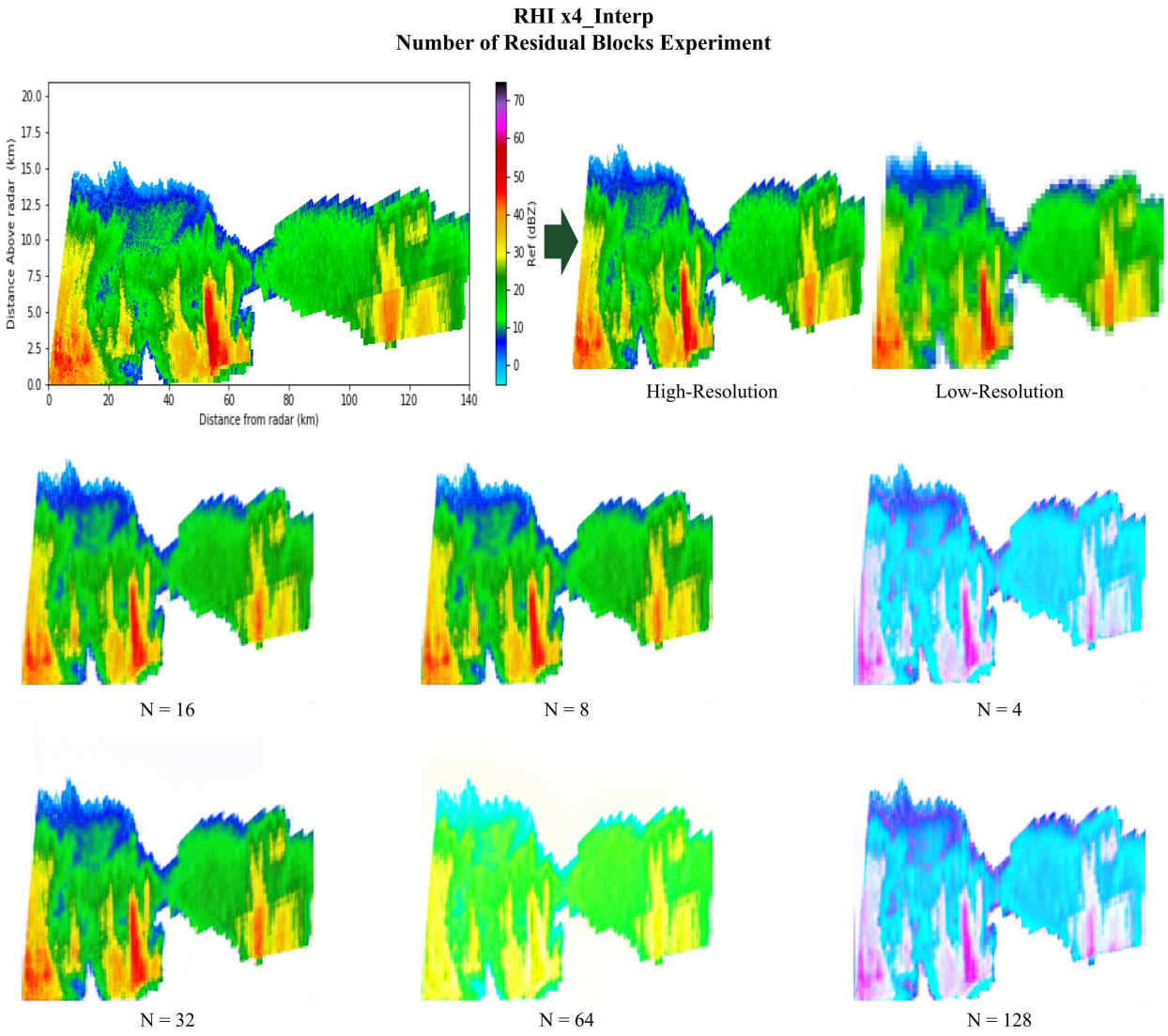

Figure 6.10: RHI x4 Interpolation Dataset: Number of Residual Blocks Experiment

SR images contain significantly higher values causing their SR images to appear faded. Although their general object shapes and high-frequency components are visible, the colors shown do not fully represent their corresponding reflectivity quantities. These observations further support the evaluation results because the NRB 4 and NRB 128 trials both had two of the lowest performances out of all the RHIx4_Interp experiments. In contrast, the NRB 8, NRB 16, and NRB 32 SR images are quite similar in their appearance when compared to the HR image which is supported by their relatively close evaluation results. Their high-frequency details, object shapes, and color boundaries are all very similar perceptually. Upon closer inspection, the NRB 32 SR image appears to be

closer to the HR image with respect to its estimations of the reflectivity within the wide updraft as well as the object shapes present within the second half of the second system. This does not follow the results reported for the NRB experimental group. The NRB 32 run was evaluated as having a lower performance on the evaluation metrics when compared to both the NRB 8 and the NRB 16 runs. Nevertheless, the NRB 32 SR image is also observed as having better representations for the areas of low reflectivity. In particular, this can be seen within the green color boundaries of the storm systems. The NRB 8 and NRB 16 SR images contain slightly lower values in these color boundaries, overestimating the reflectivity in these regions. There are not many distinguishing characteristics that differentiate between the NRB 8 and the NRB 16 SR images. Even so, the NRB 8 SR image contains slightly lower values in general which can be perceived primarily when comparing the high reflectivity within the updraft areas. In terms of their evaluations, the NRB 16 run had a slightly higher performance in both PSNR and SSIM. These observational analyses imply that the evaluation metrics do not consistently reflect a SR image's perceptibility, except for the extreme cases for which the evaluation metrics give a reasonable indication of the SR image's likeness compared to the HR image.

The lowest ranking performance of the RHIx4_Interp tests was [64, 128, 16]. This GFS value also had the lowest ranking performance for the RHIx2_Interp tests. The parameter set that achieved the highest ranking performance was [16, 64, 16]. The [64, 128, 16] test had the slowest training time, similar to the other RHI SRGAN model tests previously mentioned, and the [64, 8, 16] test had the fastest training time. The average training time for this SRGAN experimental group is 15:24:53.

6.4 RHIx4 Physically Representative Dataset SRGAN

This experimental group focuses on SRGAN models trained on the RHIx4_PhysRep dataset. The results for these experiments can be found in Table 6.4. When the DFS was set to its smallest value of 8, the evaluation performance was greatly reduced when compared with the rest of the evaluation group. The PSNR for this experiment was 17.85, the MSE was 0.109, and the SSIM

Table 6.4: Experimental Results SRGAN: RHI x4 Physically Representative Dataset

Test Parameter	Discriminator Filter Size	Generator Filter Size	Number of Residual Blocks	PSNR	MSE	SSIM	Train Time
Reference	64	64	16	20.24	0.056	0.809	15:15:56
Dis. Filter Size	32	64	16	19.99	0.049	0.815	14:37:03
	16	64	16	**20.77**	0.048	0.825	16:06:59
	8	64	16	17.85	0.109	0.782	13:29:57
	128	64	16	20.63	0.051	0.824	16:21:22
Gen. Filter Size	64	32	16	19.46	0.073	0.801	12:39:04
	64	16	16	20.40	0.052	0.816	12:17:22
	64	8	16	20.55	0.050	0.822	**12:00:26**
	64	128	16	7.58	0.708	0.176	24:37:29
Number of Residual Blocks	64	64	8	20.09	0.058	0.822	15:05:17
	64	64	4	20.33	0.049	0.825	14:36:56
	64	64	32	20.39	0.053	0.821	15:11:42
	64	64	64	20.75	0.048	**0.828**	18:28:44
	64	64	128	20.63	**0.045**	0.823	24:12:26

was 0.782. Incidentally, this DFS trial was able to have the fastest training time out of the DFS experiments, completing its training within 13:29:57. The DFS 16 experimentation test was able to outperform all other DFS experiments tested in all of the evaluation metrics. It is also one of the highest performing sets of parameters throughout all of the RHIx4_PhysRep experiments,

performing to a similar degree as the NRB 64 test. They both had the second lowest MSE overall and while DFS 16 had a higher PSNR, NRB 64 had a higher SSIM. The slowest DFS variable in terms of training time was when DFS was 128. This test took 16:21:22 to complete its training cycles. Unlike the previous experimental groups, the varied nature of the RHIx4_PhysRep model training times do not indicate an explicit correlation between DFS and training time.

The SR images presented in Figure 6.11 were generated by the RHIx4_PhysRep DFS experimental SRGAN models. The DFS 8 SR image distinctly differs from the rest of them as the colors have been considerably oversaturated. This makes the high-frequency details and color boundaries quite difficult to decipher. The rest of the SR images have similar visual appearances which is supported by the evaluation results. However, the DFS 32 SR image is also of lower quality due to its gray background and the pixelation of the high-frequency details in the second half of the second system. It also has significantly lower values overall. These observations support the results found in Table 6.4 as both the DFS 8 and DFS 32 runs had the two lowest performances out of the DFS experimental group, with the DFS 8 run having the absolute lowest performance of the DFS runs. All of these SR images contain edge artifacts in their lower left corners and have distortions to the high-frequency components of the second system. The downsampling method used to create the LR image segments this region considerably, making it more difficult for the SRGAN models to properly reconstruct the second system's details. The DFS 16 SR image significantly underestimates the reflectivity throughout the radar scan, containing higher values overall. Because of this, the color boundaries are difficult to distinguish between. It also contains shadow line artifacts along the edges of the storms, making the SR image appear blurred. It is of great interest to note that the DFS 16 run was noted as being one of the highest performing SRGAN models out of all the RHIx4_PhysRep experiments conducts. Nevertheless, this does not reflect the visual quality of the SR images it generates as it does not properly reconstruct the reflectivity information present in the HR image. These qualitative analyses suggest that decreasing the DFS parameter also decreases the perceptibility of their SR images. The DFS 64 and DFS 128 SR images also contain shadow line artifacts along the storm edges; however, they do not decrease the quality of the image

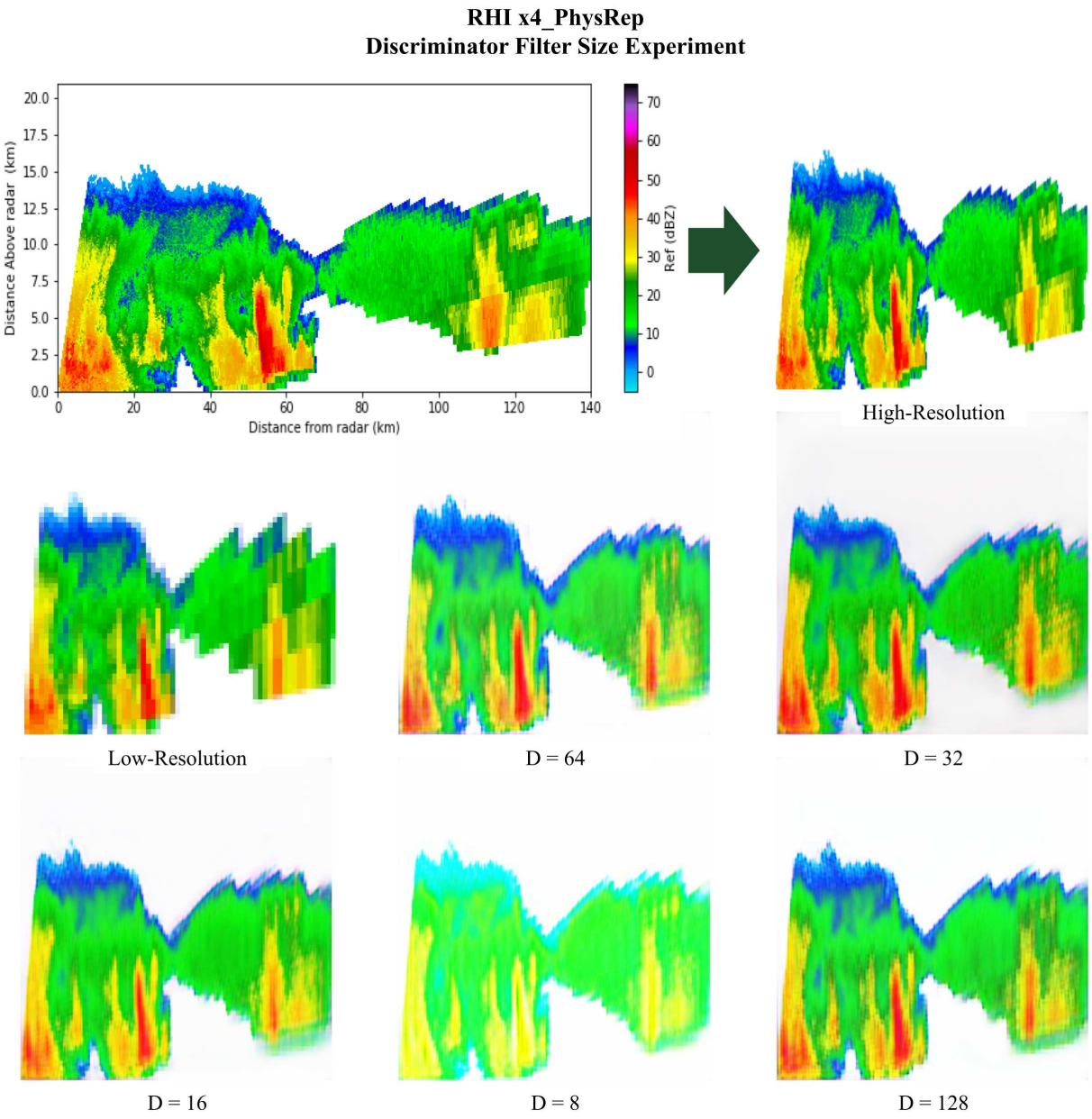

Figure 6.11: RHI x4 Physically Representative Dataset: Discriminator Filter Size Experiment

to the same extent as the DFS 16 SR image. When comparing the DFS 64 and the DFS 128 SR images, the DFS 64 SR image is observed to have more blurring in the second half of the second system, primarily in the lower altitudes around 2.5 - 5.0 km above the radar. The color boundaries are more difficult to perceive and, while the reflectivity is noticeably overestimated in the wide updraft area, the reflectivity is generally underestimated throughout the other regions of the radar scan. On the other hand, the DFS 128 SR image has significantly more pixelation throughout the

whole radar scan which, in turn, makes the color boundaries blend together and the object shapes less defined. It also has a more prominent edge artifact in the lower left corner of the image. In terms of image quality, the DFS 64 SR image appears to be more readable while the DFS 128 SR image is a better representation of the reflectivity present in the HR image. When comparing their evaluation results, the DFS 128 run performed higher on all of the evaluation metrics tested. From these observations, it could be suggested that a larger DFS results in more representative SR images.

For the GFS experiments, the performance of the 128 GFS run was significantly degraded compared to all other parameters under investigation within the RHIx4_PhysRep experiments. Its PSNR was the overall lowest, calculated to be 7.58, its MSE was the highest at 0.708, and its SSIM was the lowest at 0.176. In addition, it had the slowest training time recorded as 24:37:19. This is over double the fastest overall time recorded which was achieved by the GFS of 8 experiment. It was trained in 12:00:26 and also received the best evaluations out of all of the GFS tests. Its PSNR was 20.55, its MSE was 0.050, and its SSIM was 0.822. These trials also show a distinct correlation between GFS and training time as larger GFS values result in longer training times.

The RHIx4_PhysRep GFS SR images are shown in Figure 6.12. Two of the SR images stand out due to their low visual quality. The GFS 32 SR image contains significantly higher values overall. This washes out the color boundaries and obscures the object shapes within the storms. The reflectivity information present in the HR image is mostly lost, especially the high-frequency regions such as the updrafts and the second half of the second system. In a similar vein, the GFS 128 SR image's background is primarily a dark gray which lowers the overall quality of the image. There is also considerable pixelation and generally lower values throughout the SR radar scan as well as black dot artifacts speckled within the updraft areas. Nevertheless, some of the high-frequency components and object shapes can be perceived in the GFS 128 SR image making it visually superior to the GFS 32 SR image. This is an interesting note as the GFS 128 run was evaluated as having the lowest performance in every evaluation metric out of all the RHIx4_PhysRep experiments tested, even though the storms' main features are more readily observed in the GFS

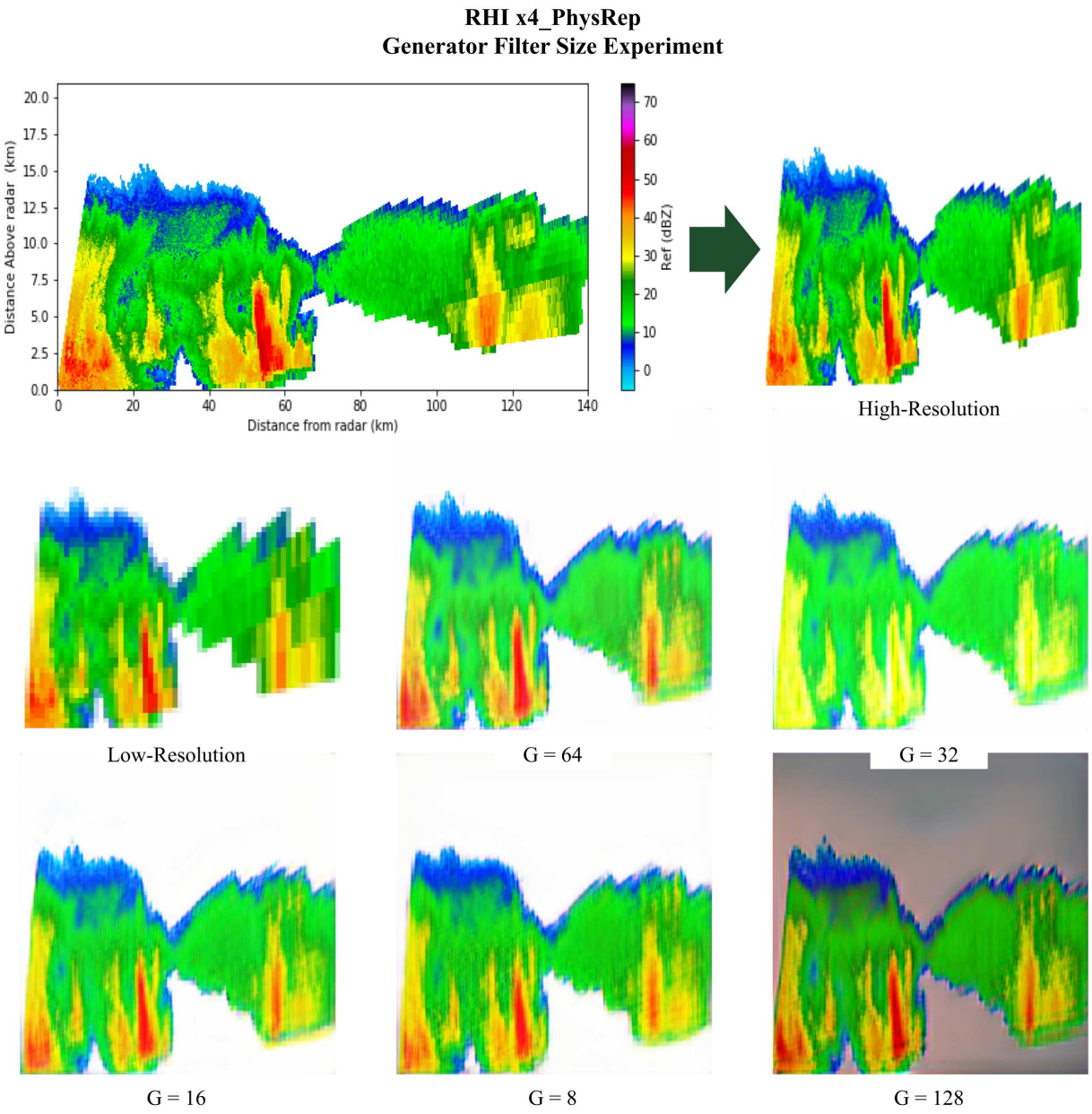

Figure 6.12: RHI x4 Physically Representative Dataset: Generator Filter Size Experiment

128 SR image than in the GFS 32 SR image. This suggests that the evaluation metrics do not reliably indicate how the SR image will be perceived. The rest of the SR images are quite similar perceptually which is supported by their close evaluation results. The GFS 8 SR image has noticeably lower values and more pixelation in general, distorting the high-frequency components and object shapes in the storms. It also underestimates the reflectivity in the updraft areas and has distinct edge artifacts in the corners of the image. These observations also do not support the evalu-

ation results as the GFS 8 run had the highest performance out of the GFS experimental group. The GFS 16 SR image contains significant pixelation distortion throughout the radar scan, has a more severe edge artifact in the lower left corner, and has a storm edge artifact that makes it appear that there is a region of reflectivity around 30 dBZ at the bottom of the second half of the storm around 3.5 km altitude above the radar. The GFS 64 SR image also has a storm edge artifact along the bottom of the second half of the second storm making this area somewhat blurred. It also slightly overestimates the reflectivity present in the wide updraft area. However, the GFS 64 SR image is perceived as having the closest representation of reflectivity present in the HR image when compared to the rest of the SR images. The object shapes are the most defined, the color boundaries are the most distinguishable, and the high-frequency components are the least distorted within the GFS 64 SR image. The characteristics observed from the SR images from this experimental group do not support the evaluation results recorded in Table 6.4. This could imply that other evaluation metrics are required, or need to be developed, to accurately measure an image for its visual perceptibility and comprehension. These observations do, however, support previous deductions that the GFS parameter being set to 64 is absolutely ideal for generating the most representative SR image.

The NRB value of 8 was found to be lacking in its performance, compared to the other NRB values tested, due to its low PSNR of 20.09 and its high MSE of 0.058. However, markedly, it did have a higher SSIM than the reference NRB value which had the lowest SSIM evaluated at 0.809 out of the NRB experiments. The NRB value of 64 obtained the highest ranking out of the NRB experiments group. It had the highest PSNR of 20.75, which was also the second highest PSNR overall, the second lowest MSE at 0.048, and the highest SSIM at 0.828. The lowest MSE, for both the NRB experiments and overall, was evaluated at 0.045 from the NRB 128 trial. Although it obtained the lowest MSE, it also had the slowest training time out of the NRB tests, and the second slowest overall, at 24:12:26. The fastest training time for the NRB values tested was 14:36:56. These results follow the previous experimental groups that generally support the assertion that increasing the NRB also increases the training time.

Figure 6.13 displays the SR images for the NRB experimental SRGAN models. All of the SR images are quite similar in terms of their general object shapes and color boundaries. This is supported by their similar evaluation results. This set of SR images has a distinct victor for being the most similar to the target HR image, the NRB 64 SR image. The NRB 4 and NRB 128 SR images display significantly lower values overall, have intense pixelation of the high-frequency components present in the second system, and have gray backgrounds. Out of these two, the NRB 4 SR image is distorted to a higher degree. On the other hand, the NRB 8 and NRB 32 SR images contain notably higher values in general, have less defined color boundaries, have shadow line edge artifacts along the bottom edge of the second system, and underestimate the reflectivity throughout the radar scan. The visual quality of the NRB 8 SR image is affected by these distortions to an even greater extent than the NRB 32 SR image. Since the NRB 4 and NRB 8 SR images were determined as having the lowest perceptibility comparatively, this could imply that a smaller NRB results in lower quality SR images. These observations follow the evaluation results as the NRB 4 and NRB 8 runs had the lowest performances out of the NRB experimental group. Despite this, it is interesting to note that they also had higher SSIM evaluations than their corresponding pair for this analysis. This shows that the SSIM does not dependably reflect the perceptibility of the SR image being evaluated. This is also supported by the NRB 16 SR image as, even though it had the second lowest performance in the PSNR and MSE metrics as well as the lowest SSIM performance out of the NRB experimental group, it generates more of the features present in the HR image than the NRB trials mentioned previously. Nevertheless, it is also observed as overestimating the reflectivity in the wide updraft, not fully representing the color boundaries present in the high-frequency areas, and having considerable blurring along the bottom edge of the second half of the second system. Therefore, the NRB 64 is perceived as being the closest representation of the HR image. This follows the evaluation results which found the NRB 64 run to be the highest ranking SRGAN model tested out of the NRB experimental group. Its color boundaries are the most defined and the object shapes in the first storm are mostly reconstructed. Most of the storms' features portrayed in the HR image can be perceived in the NRB 64 SR image. However, at the

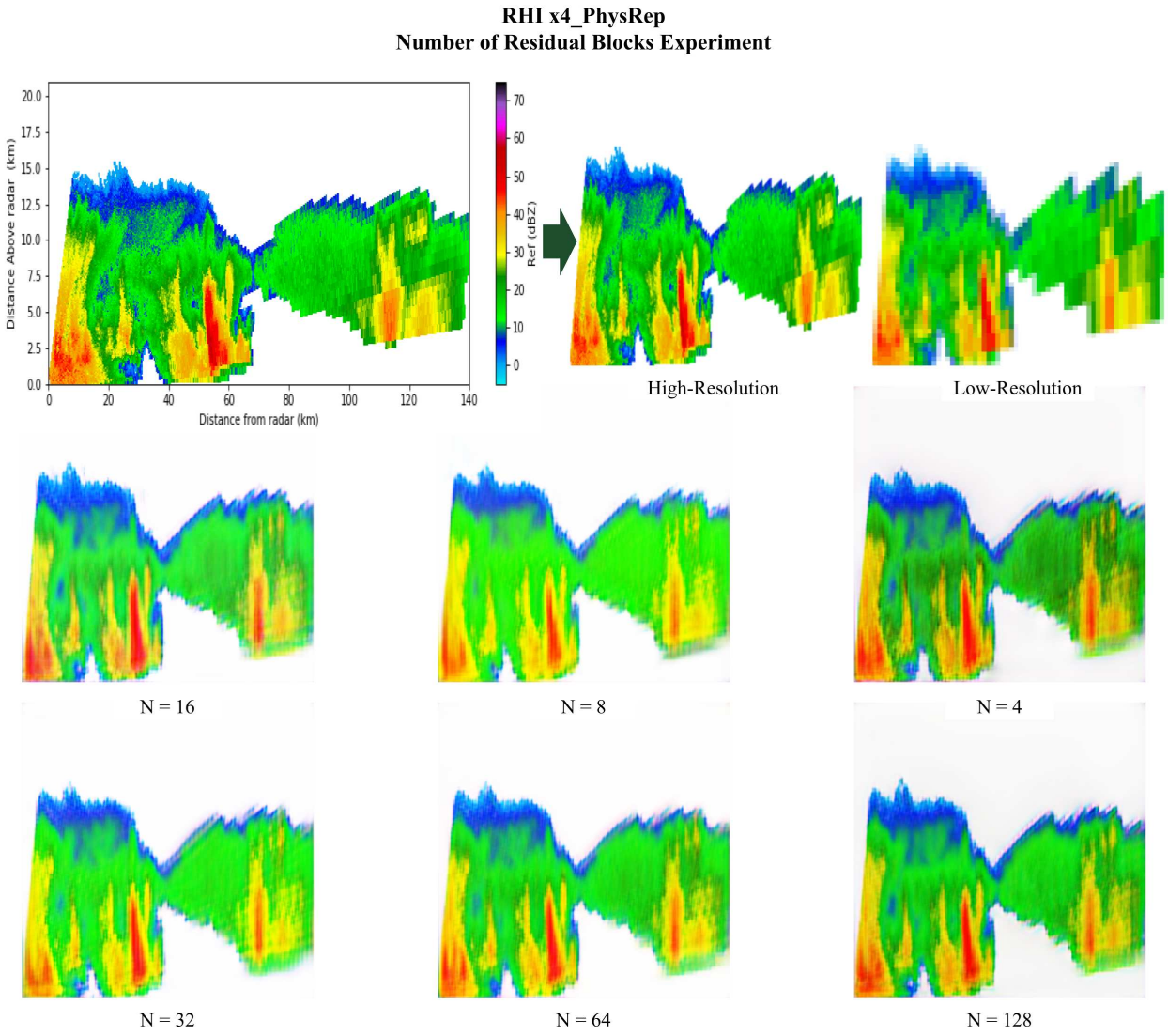

Figure 6.13: RHI x4 Physically Representative Dataset: Number of Residual Blocks Experiment

same time, the predicted reflectivity in the updraft regions is underestimated, the mid-tone green color boundary in the first storm is not fully represented, the object shapes in the second half of the second system are distorted and unclear, and there is a prominent, dark line edge artifact in the lower left corner as well as blurring in the bottom edge of the second system.

The set of parameters that had a substandard performance overall was [64, 128, 16]. This was the same GFS value that had the lowest performance for the RHIx2_Interp and the RHIx4_Interp experiments. The sets of parameters that were able to outperform all other RHIx4_PhysRep model variations tested were [16, 64, 16] and [64, 64, 64]. The [16, 64, 16] configuration also had the

highest ranking performance for the RHIx4_Interp trials. The slowest training time was recorded by [64, 128, 16]. The GFS of 128 test have proved to have the slowest training time across all RHI SRGAN models tested. The fastest training time was recorded by [64, 8, 16]. A GFS of 8 has also had the fastest training time for 3 out of the 4 different RHI SRGAN model configurations tested. The average training time for this SRGAN experimental group is 16:04:20.

6.5 PPIx2 Interpolation Dataset SRGAN

This experimental group is comprised of SRGAN models that were trained on the PPIx2_Interp dataset. These results are laid out in Table 6.5. In this case, the reference value utilized was evaluated as having the lowest PSNR of 23.54, the highest MSE of 0.024, and the lowest SSIM of 0.900 out of the DFS tests run. On the other hand, the DFS value of 16 outperformed all of the other sets of parameters within the PPIx2_Interp experiments. A DFS of 16 achieved the highest overall PSNR of 25.98, the lowest overall MSE of 0.013, and the highest SSIM of 0.937. The slowest training time of 8:11:46 out of the DFS trials was recorded by the largest DFS value of 128 while the fastest training time of 6:42:00 was recorded by the smallest DFS value of 8. This further implies that DFS has a direct affect on the time required to train a SRGAN model.

From the SR images presented in Figure 6.14, it is clear to see that the DFS 32 SR image is the least similar to the ground truth image when compared with the rest of the SR images. It has lower values overall which, in turn, results in overestimations in the reflectivity present throughout the storm system. In addition, the high-frequency areas and color boundaries are less defined making the finer object shapes more difficult to perceive. The object shapes of the squall are also much thicker than in the HR image. This follows the evaluation results as the DFS 32 run had the lowest performance on each of the evaluation metrics out of the DFS experimental group. The rest of the SR images are quite similar in their visual perceptibility which is supported by their close evaluation results. All of the PPIx2_Interp DFS SR images have a dark line edge artifact that also contains blurring around the areas on the left and bottom edges of the SR image. The DFS 128 SR image has the least discernible edge artifacts, perhaps due to its higher values in general.

Table 6.5: Experimental Results SRGAN: PPI x2 Interpolation Dataset

Test Parameter	Discriminator Filter Size	Generator Filter Size	Number of Residual Blocks	PSNR	MSE	SSIM	Train Time
Reference	32	32	16	23.54	0.024	0.900	6:49:54
Dis. Filter Size	16	32	16	**25.98**	**0.013**	**0.937**	6:42:35
	8	32	16	25.10	0.015	0.926	6:42:00
	64	32	16	24.42	0.017	0.918	6:50:11
	128	32	16	25.68	0.014	0.935	8:11:46
Gen. Filter Size	32	16	16	25.65	0.014	0.935	6:34:26
	32	8	16	22.77	0.026	0.889	**6:28:47**
	32	64	16	5.54	1.127	0.354	7:33:33
	32	128	16	4.72	1.351	0.263	12:35:44
Number of Residual Blocks	32	32	8	22.59	0.024	0.903	6:37:10
	32	32	4	25.27	0.015	0.927	6:50:35
	32	32	32	24.88	0.015	0.926	7:05:08
	32	32	64	2.42	2.304	-0.219	8:42:11

However, this also results in the SR image appearing to underestimate the reflectivity throughout the radar scan. The high-frequency areas in the squall lines are not fully generated as the darker color boundaries such as the dark red, orange and dark green boundaries are not represented in the DFS 128 SR image. These observations do not fully support the evaluation results as, even though the DFS 128 SR image is determined to be one of the least similar to the HR image, the

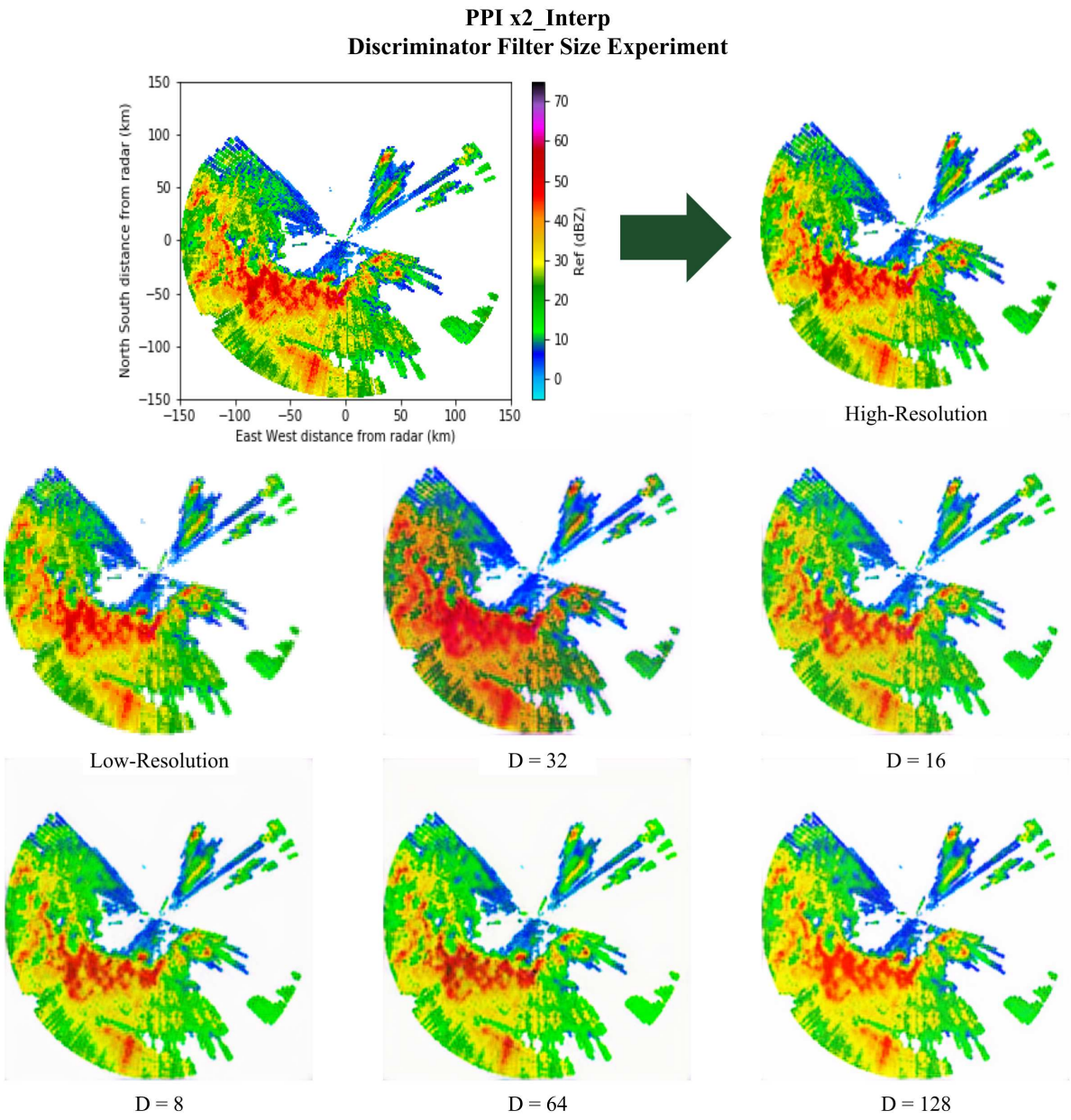

Figure 6.14: PPI x2 Interpolation Dataset: Discriminator Filter Size Experiment

DFS 128 run had the second highest performance on all the evaluation metrics out of the DFS experimental group. When observing the DFS 64 SR image, it is apparent that the high-frequency components are not properly generated. Many of the intricate variations in reflectivity within the primary squall line are reduced so that the many color boundaries within the HR image are, instead, represented by a single color boundary. Particularly the red, orange and dark green color boundaries. Furthermore, the DFS 64 SR image contains black spot artifacts speckled throughout

the primary squall line region. The DFS 8 and DFS 16 SR images also have these black spot artifacts but they are much less prominent than in the DFS 64 SR image. The DFS 64 experimental test was outperformed by both the DFS 8 and the DFS 16 runs, with the DFS 16 experimental SRGAN model having received the highest ranking both out of the DFS experimental group and all the PPIx2_Interp SRGAN models overall. The evaluation results provide further support to the following qualitative observations. Out of the remaining DFS experiments, the DFS 16 SR image is considered to more accurately portray the characteristics of the HR image. The DFS 8 SR image is comprised of slightly higher values, which generally underestimates the reflectivity. The darker color boundaries such as the dark red, orange, and dark green boundaries are also not fully represented. The DFS 16 SR image displays more of these characteristics and tends to estimate the reflectivity more closely to the HR image. Nevertheless, the orange color boundaries in both of the squall lines are not properly generated and the high-frequency details are not as defined. In addition, the DFS 16 SR image overestimates the areas around the points of highest reflectivity of about 63 dBZ within the primary squall line so that these areas appear noticeably larger. It is interesting to note that setting the DFS parameter to 16 achieved the highest visual performance for both the PPIx2_Interp and the RHIx2_Interp experiments and that the SR images generated by these experimental SRGAN models had some of the lowest values overall out of their respective experimental groups.

When the GFS was set to 16, the PPIx2_Interp SRGAN achieved the highest ranking in terms of its performance out of the GFS experiments across all evaluation metrics. Its PSNR was evaluated at 25.65, its MSE was 0.014, and its SSIM was 0.935. The GFS values of 64 and 128 performed inadequately on each of the evaluation metrics, with 128 ranking as the lowest performing GFS value tested. It was evaluated as having the lowest PSNR of 4.72, the highest MSE of 1.351, and the lowest SSIM of 0.263. This was also the second lowest performing set of parameters out of all of the PPIx2_Interp tests, with a GFS of 64 being the third lowest performing. These results suggest that a higher GFS value will prove to be counterproductive to training an effective SRGAN that can produce super-resolution images in the same likeness as their high-resolution target images,

especially when considering PPIx2_Interp datasets. The 128 GFS trial also had the longest training time of 12:35:44. The 8 GFS trial had the shortest training time of 6:28:47 both out of the GFS experimental group and overall. These results also show that increasing GFS increases training time as well.

The example SR images for the GFS experimental group are shown in Figure 6.15. So far, this set of SR images are the most visually different from one another which is supported by the wide spread in their evaluation results. Out of these, it is evident that the GFS 128 SR image was unable to generate any of the characteristics of the HR image. It is merely a blank image of a single hue. The GFS 8 and GFS 64 experimental SRGAN models also generated inadequate representations of the HR image. The GFS 8 SR image contains significantly higher values in general, blending the color boundaries together, making the object shapes appear smaller and underestimating the reflectivity throughout the radar scan. It also has dark spot artifacts speckled throughout the primary squall line in addition to dark line edge artifacts lining every side of the SR image. The GFS 64 SR image has a yellow background as opposed to the white space background of the HR image. Furthermore, dark area artifacts are spread around the edges of the storm system obscuring the object shapes. If one were to ignore these characteristics, the GFS 64 SR image might have been in the running for the closest representation of the HR image out of the GFS experimental group. These qualitative analyses are supported by the evaluation results. The GFS 128 and GFS 64 runs had significantly lower performances when compared with most of the other PPIx2_Interp experiments, except for one. The GFS 8 run also had an inadequate performance, albeit more comparable to the other GFS experimental trials. Out of the remaining SR images, the GFS 16 SR image is determined as being the closest resemblance to the HR image. This statement is reinforced by the evaluation results. The GFS 16 run had the highest performance out of the GFS experimental group, with the GFS 32 run following close behind. The GFS 32 SR image is observed as having significantly lower values throughout the radar scan. This makes the SR image appear to overestimate the reflectivity and have larger, less defined object shapes in general. In addition, the brighter color boundaries such as the yellow and green boundaries are not well represented in the

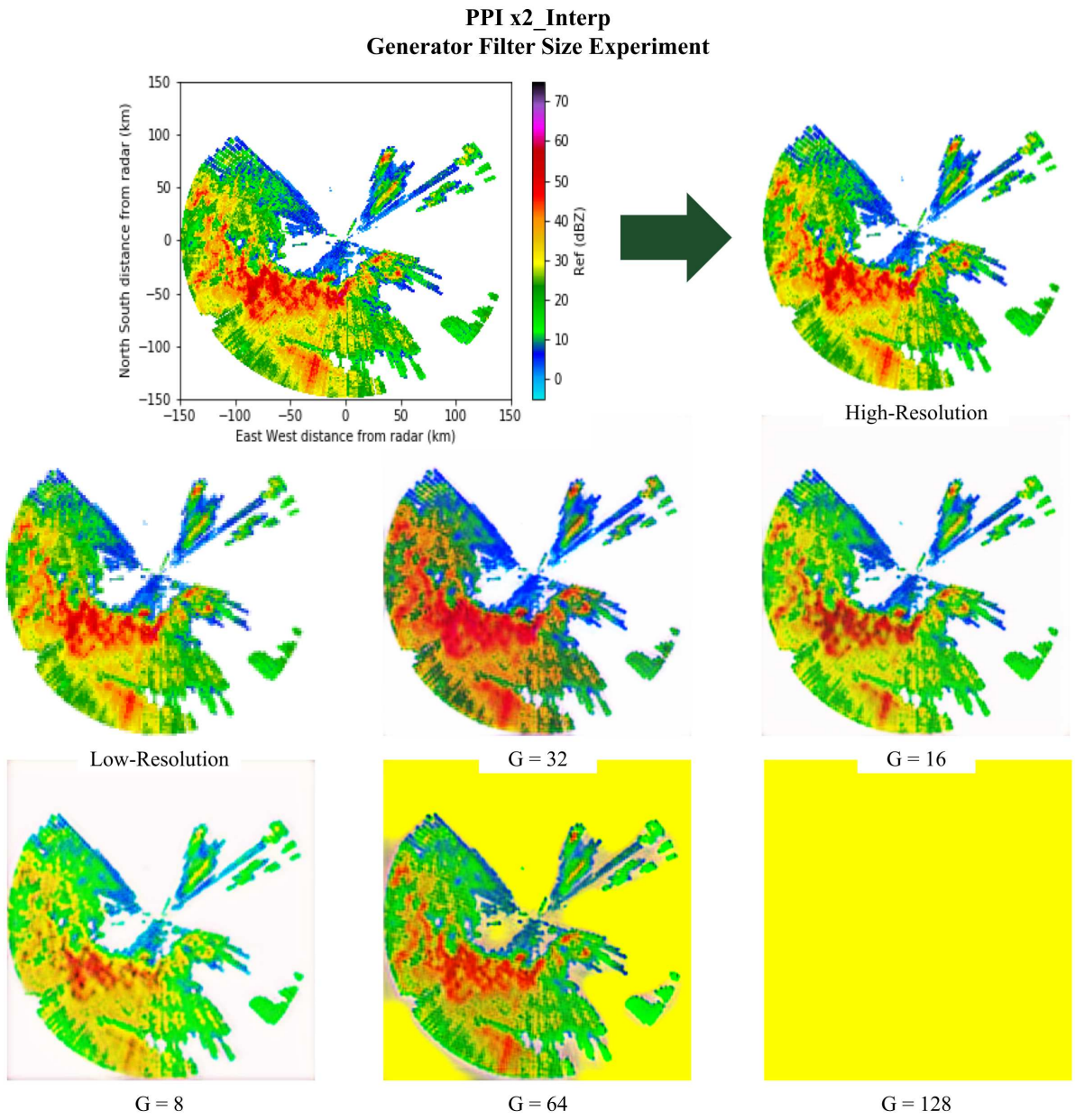

Figure 6.15: PPI x2 Interpolation Dataset: Generator Filter Size Experiment

GFS 32 SR image. Although the GFS 16 SR image is declared to be the closest representation to the HR image when compared with the rest of the GFS SR images, it leaves much to be desired in terms of its generative abilities. The high-frequency details within the primary squall line are not properly portrayed, dark spot artifacts can be found within the primary squall line, and some of the color boundaries are missing such as the dark green, orange, and red boundaries in both the squall line areas in the GFS 16 SR image. Both qualitative analyses from the PPIx2_Interp and the

RHI x2_Interp GFS experimental groups asserted that the GFS 16 SRGAN model had generated the closest representations to the HR image. This is a significant finding as, so far, the type of radar scan used has not affected the outcome of the relative perceptibility of this parameter when compared with the others in its experimental group.

The value that proved to obtain significantly substandard evaluations across all evaluation metrics over all of the PPIx2_Interp tests was a NRB of 64. When the NRB was set to 64, it was evaluated as having the lowest overall PSNR of 2.42, the highest overall MSE of 2.304, and the lowest overall SSIM of -0.219. It also had the longest training time of 8:42:11 out of the NRB experimental group. The shortest training time in the NRB experimental group was recorded by the smallest NRB value of 8 with a training time of 6:37:10. The NRB value that had the highest performance in its evaluations out of this experimental group was the smallest value with a NRB of 4. It obtained a PSNR of 25.27, a MSE of 0.015, and a SSIM of 0.927. This MSE was also evaluated by the NRB of 32 run.

Figure 6.16 contains the SR images for the NRB experimental group. This set of SR images are quite different from one another in terms of their visual representation which is supported by the spread in their evaluation results. Even though some of the storm features can be observed, the NRB 64 SR image has a blue background and is comprised of significantly lower values that make the object shapes and high-frequency components more difficult to perceive. Similarly, the NRB 16 also has significantly lower values in general that make the object shapes of the squall lines appear larger, the reflectivity estimations higher than the ground truth, and some of the brighter color boundaries to be missing such as the yellow and green boundaries. It also has considerable dark line edge artifacts along the left and bottom edges of the SR image. The NRB 16 run had a fairly comparable performance to the rest of the PPIx2_Interp experimental models. However, it had one of the lowest SSIM evaluations which further refutes previous analyses suggesting that higher SSIM evaluations indicate lower values in the SR image generated overall. The NRB 64 run was found to be the lowest performing SRGAN model out of all the PPIx2_Interp experiments. This evaluation is quite fascinating in that the NRB 64 SR image had a lower performance quan-

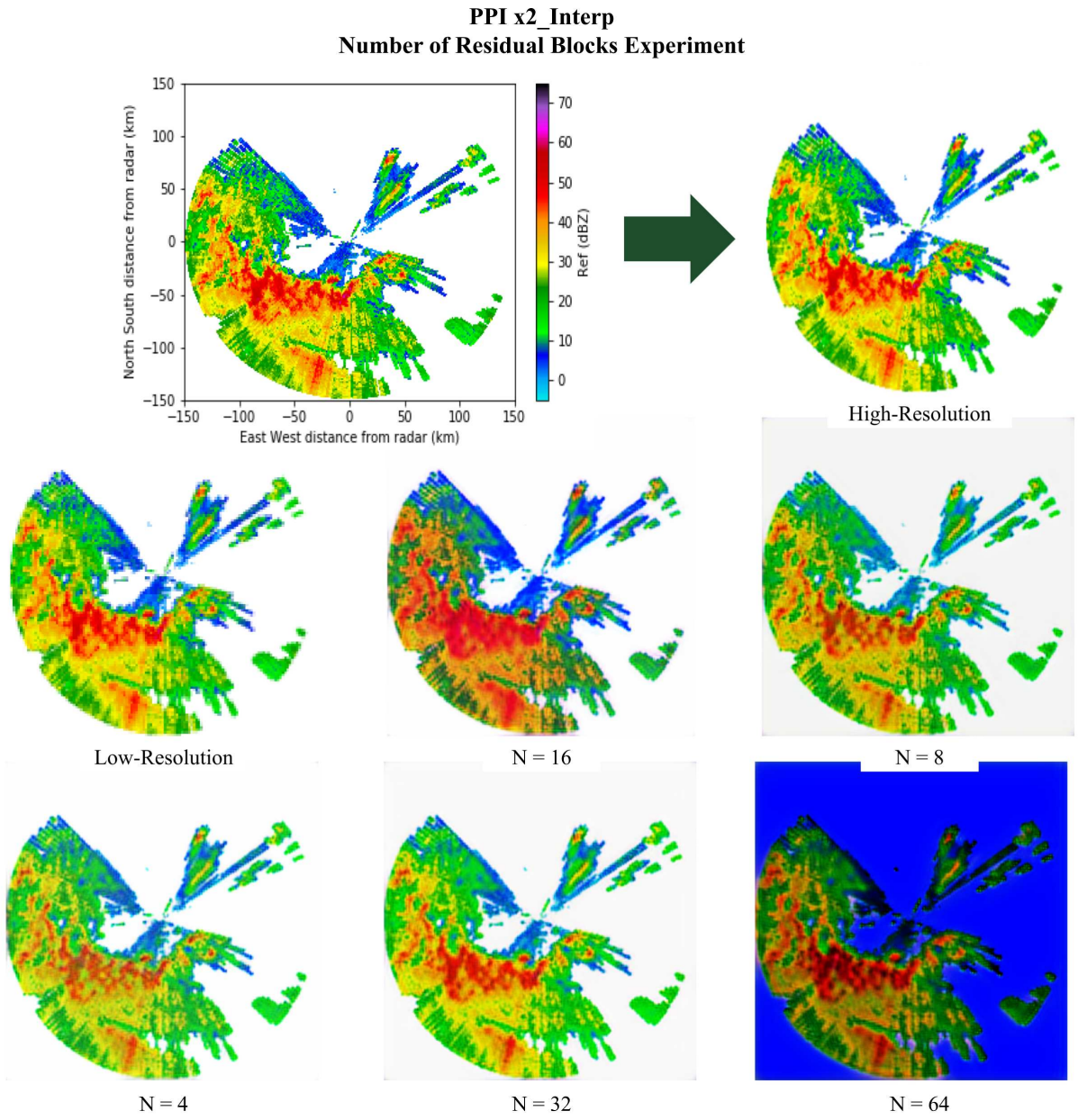

Figure 6.16: PPI x2 Interpolation Dataset: Number of Residual Blocks Experiment

titatively than the GFS 128 SR image consisting of a single color, even though more of the storm features can be recognized in the NRB 64 SR image. The NRB 4 SR image is the least affected by the edge artifacts; however, the colors displayed appear faded as they are much less vibrant than the HR image. Because of this, the details in the high-frequency components and color boundaries are obfuscated making the squall line features more difficult to discern. These observations do not reflect the evaluation results of the NRB 4 run as it was professed as being the highest

performing NRB experimental model. When observing the NRB 8 SR image, it is noted that the high-frequency details of the squall lines are less defined and appear to be slightly blurred as well in these particular regions. This could be due to having less defined color boundaries. Some of the brighter color boundaries are also not fully represented such as the yellow and green boundaries. In addition, the NRB 8 SR image has a slightly gray background. Nevertheless, the NRB 8 SR image is one of the closest representations to the HR image in terms of its perceptibility when compared to the other NRB experimental tests, except for the NRB 32 SR image. The NRB 8 run was evaluated as having the second lowest performance out of the NRB experimental group while the NRB 32 run was evaluated as having the second highest performance out of the NRB experimental group. These qualitative analyses suggest that the evaluation metrics are not alone sufficient for determining visual comprehension when comparing between images. Despite the slightly grey background and noticeable edge artifacts along the left and bottom borders, the NRB 32 SR image is a close resemblance to the HR image. It possesses the most defined high-frequency components and displays most of the color boundaries. But some of the object shapes are still not properly characterized and some of the color boundaries are still not represented such as the dark green and orange boundaries.

The lowest ranking performance of the PPIx2_Interp experiments was [32, 32, 64]. The parameter set that achieved the highest ranking performance was [16, 32, 16]. The [32, 128, 16] test had the slowest training time and the [32, 8, 16] test had the fastest training time. This is similar to results from the RHI experiments where the smallest GFS value had the fastest training time and the largest GFS value had the slowest training time for a majority of the results recorded. The average training time for this SRGAN experimental group is 7:31:05.

6.6 PPIx2 Physically Representative Dataset SRGAN

This experimental group consisted of SRGAN models that were trained with images from the PPIx2_PhysRep dataset. The results for these experiments are shown in Table 6.6. The reference value used for the DFS parameter had the lowest ranking performance out of the DFS experimental

Table 6.6: Experimental Results SRGAN: PPI x2 Physically Representative Dataset

Test Parameter	Discriminator Filter Size	Generator Filter Size	Number of Residual Blocks	PSNR	MSE	SSIM	Train Time
Reference	32	32	16	16.12	0.100	0.801	7:05:39
Dis. Filter Size	16	32	16	21.45	0.038	0.885	6:32:21
	8	32	16	21.01	0.042	0.880	6:56:18
	64	32	16	21.61	0.037	0.884	7:17:55
	128	32	16	21.40	0.039	0.883	8:26:07
Gen. Filter Size	32	16	16	21.68	0.037	0.888	**6:22:16**
	32	8	16	21.44	0.038	0.883	6:38:51
	32	64	16	2.27	2.381	-0.224	7:25:15
	32	128	16	5.17	1.221	0.321	13:11:56
Number of Residual Blocks	32	32	8	21.56	0.037	0.888	6:48:40
	32	32	4	21.61	0.037	0.887	6:38:53
	32	32	32	**21.89**	**0.035**	**0.893**	7:13:52
	32	32	64	21.82	**0.035**	0.892	8:23:38

group. It had the lowest PSNR of 16.12, the highest MSE of 0.100, and the lowest SSIM of 0.801. The DFS value of 64 achieved the highest PSNR of 21.61 and the lowest MSE of 0.037 out of the DFS experimental group. It was also evaluated to have the second highest SSIM of 0.884. The highest SSIM out of the DFS experimentation group was achieved by a DFS of 16. This was also the DFS value that had the fastest training time of 6:32:21 out of the DFS trials. The slowest

training time for the DFS values tested was recorded by the DFS of 128 test. It had a training time of 8:26:07.

The SR images in Figure 6.17 appear very similar in terms of their generic object shapes and color boundaries, although none of them generated all of the color boundaries. This means that some of the storm's reflectivity is not indicated in these SR images. Further more, there are entire pieces of the storm system, around 0 - 50 km on the East-West axis and around 0 - 100 km on the North-South axis as well as the farther ranges in the top left quadrant of the HR image, that are not represented in any of the PPIx2_PhysRep SR images. The cause of this is likely due to the physically representative downsampling method as these regions of the storm are missing from the corresponding LR image as well. This exemplifies one of the limitations of the SRGAN model in that the SR images generated are dependent upon the form of the LR input image that it is given. The DFS 32 SR image was the least representative of the HR target image, primarily due to its background not being white space and the black dot artifacts dotted throughout the primary squall line. In addition, the DFS 32 SR image has higher values in general and is noticeably pixelated making the high-frequency components difficult to discern. This is supported by the evaluation results as the DFS 32 run was determined to be the lowest performing out of the DFS experimental group. Besides this SR image, the rest of them are quite similar in their visual perceptibility which is supported by their close evaluation results. All of the SR images contain noticeable black dot artifacts in the primary squall line, with the DFS 32 and DFS 16 SR images having the most prominent artifacts while the DFS 8 SR image has the faintest. The DFS 8 SR image contains significantly higher values overall. This washes out the color boundaries and makes the object shapes within the storm system less defined. Additionally, the low reflectivity regions are significantly underestimated throughout the storm. The DFS 16 SR image has significantly lower values in general, dismissing the lighter color boundaries while overestimating the reflectivity throughout the storm. This also makes the high-frequency details in the primary squall line less distinct while widening the object shape in the secondary squall line. However, the DFS 16 SR images gives more proper representations of the low reflectivity present throughout the storm when

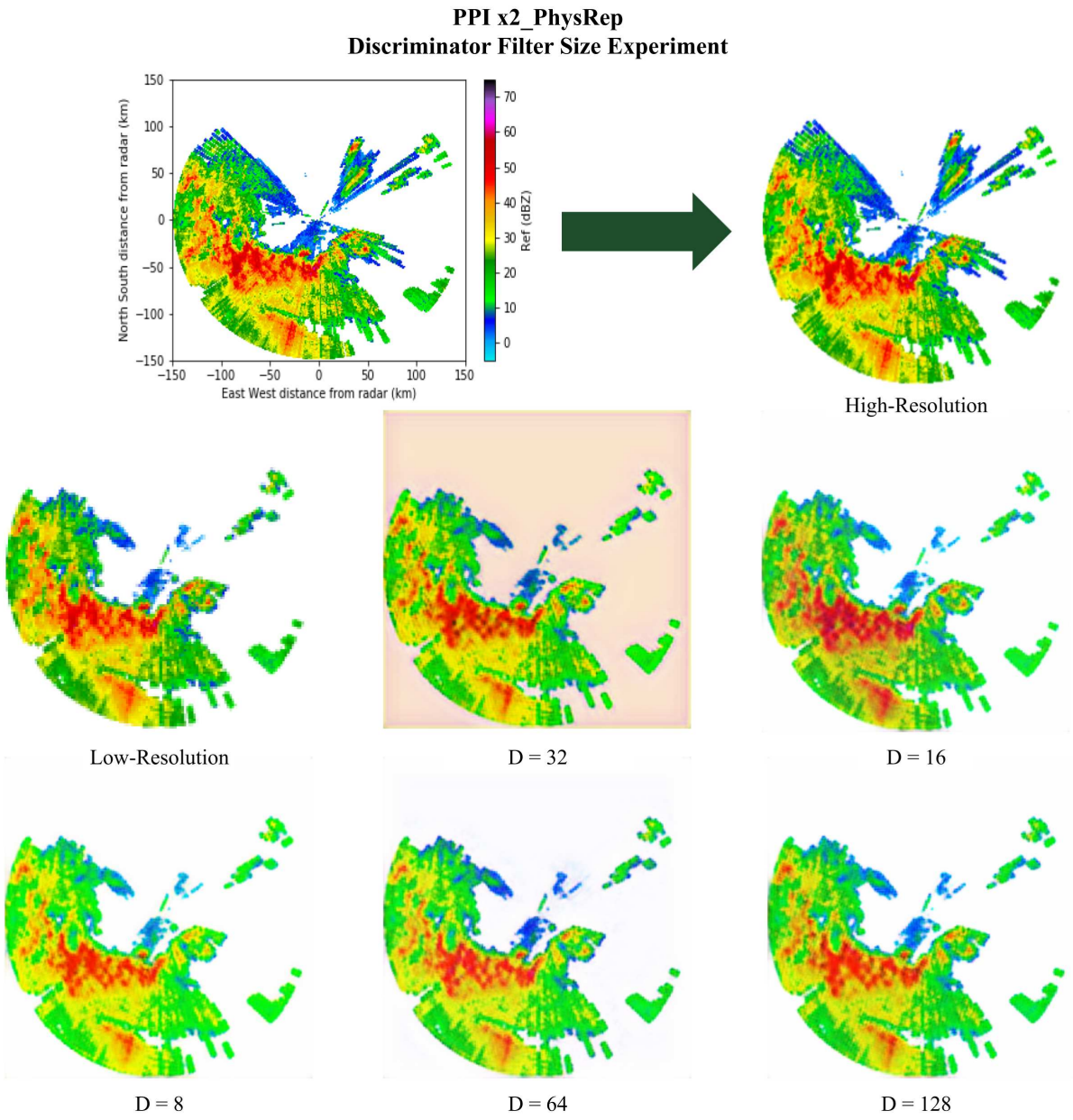

Figure 6.17: PPI x2 Physically Representative Dataset: Discriminator Filter Size Experiment

compared with the rest of the SR images. If the high-frequency details in the squall line features were not so distorted, the DFS 16 SR image would be the closest representation of the HR image. The remaining two SR images are quite similar in terms of their visual appearance. The DFS 128 SR image has lower values and more fully displays the color boundaries present in the HR image when compared to the DFS 64 SR image. The DFS 64 SR image has more defined high-frequency details and object shapes than the more blurred DFS 128 SR image. Additionally, the

black dot artifacts are much fainter and the small areas of high reflectivity around 63 dBZ are more accurately represented in the DFS 64 SR image. However, the DFS 128 SR image more fully represents the reflectivity present when compared to the DFS 64 SR image, especially throughout the areas of lower reflectivity. For these reasons, the DFS 128 SR image is considered to better represent the reflectivity of the HR image overall. These observations are not completely supported by the evaluation results. The DFS 32 and DFS 8 runs had the lowest performances out of the DFS experimental group. Meanwhile, the DFS 64 run had the highest performance out of the DFS experimental group, even though this was not reflected in its visual appearance. The DFS 16 and DFS 128 runs performed quite similarly on their evaluations which makes sense due to their more suitable representations of the lower reflectivity throughout the storm system.

The largest GFS values of 64 and 128 were evaluated to have the lowest ranking performances out of all of the PPIx2_PhysRep experiments, with the GFS of 64 having the lowest ranking performance overall. It had the overall lowest PSNR of 2.27, the overall highest MSE of 2.381, and the overall lowest SSIM of -0.224. The 128 GFS run had the second lowest overall PSNR of 5.17, the second highest overall MSE of 1.221, and the second lowest overall SSIM of 0.321. These values are significantly dissimilar to the rest of the evaluations obtained by the other PPIx2_PhysRep tests. It also suggests that lower GFS values will improve the overall performance of PPIx2_PhysRep SRGAN models. The GFS of 128 also had the slowest overall training time of 13:11:56. When the GFS was set to 16, it was recorded as having the fastest overall training time of 6:22:16. This was also the GFS value that had the highest ranking performance within the GFS experimentation group. It achieved the highest PSNR of 21.68, the lowest MSE of 0.037, and the highest SSIM of 0.888.

Many of the SR images in Figure 6.18 are significantly distorted. The SR images within this set are quite different from one another which follows the wide spread in their evaluation performances. The GFS 64 and GFS 128 SR images retained the least amount of information from the HR image. They both have solid color backgrounds instead of white space, contain prominent dark spot artifacts throughout the squall line features, do not fully represent the color boundaries

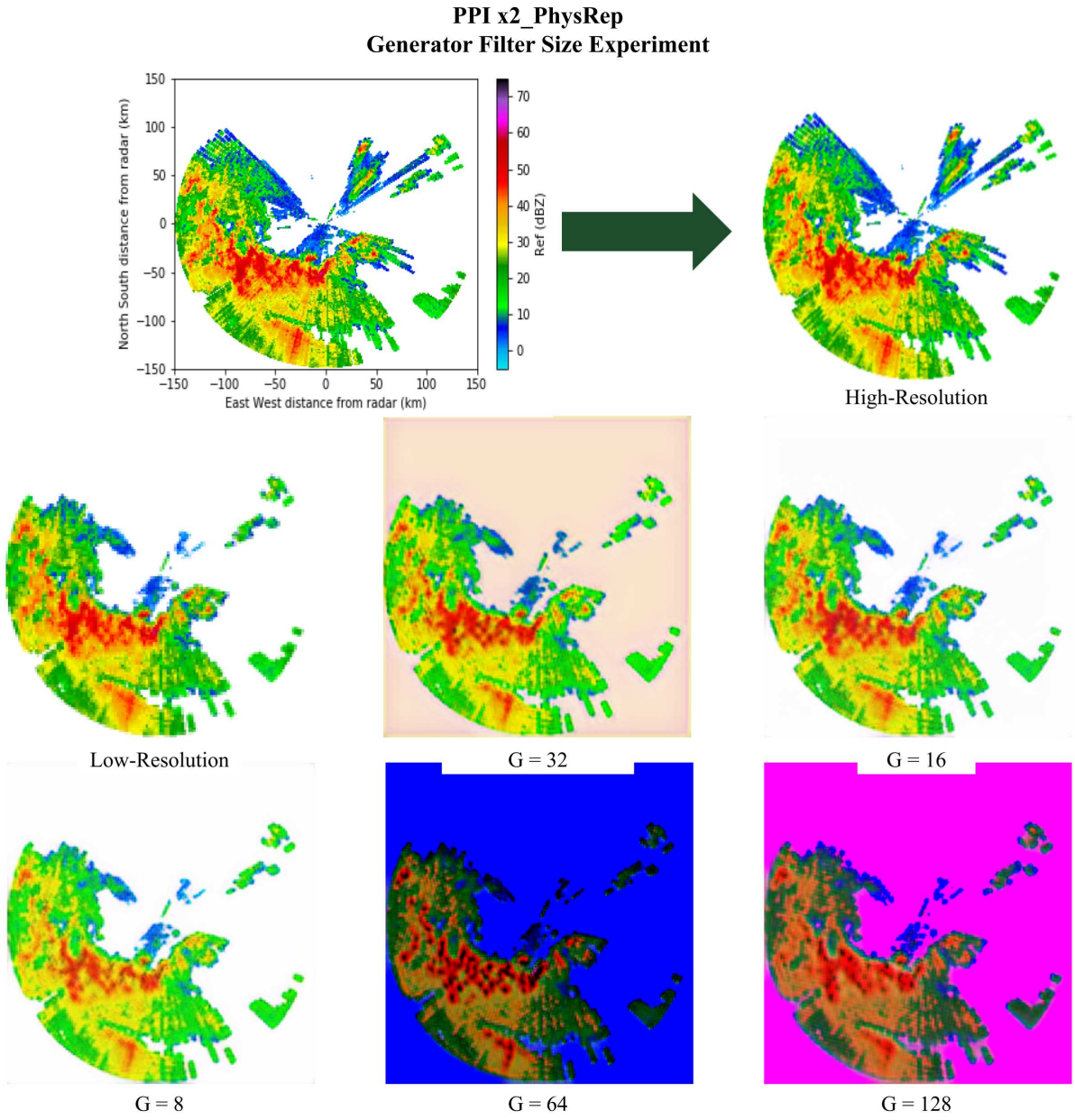

Figure 6.18: PPI x2 Physically Representative Dataset: Generator Filter Size Experiment

present in the HR image, have lost most of the high-frequency details, have thicker object shapes and have lower values that significantly overestimate the reflectivity throughout the storm. This further validates the evaluation results as both the GFS 64 and GFS 128 runs had the absolute lowest performances out of all the PPIx2_PhysRep experiments. Out of the remaining three, the GFS 16 SR image is the closest representation of the HR image. The GFS 8 SR image contains generally higher values that underestimate the reflectivity present in the storm. In addition, not all of the

color boundaries are represented and the high-frequency details are not well-defined. The GFS 32 SR image is quite pixelated, there are prominent dark spot artifact within the primary squall line, and it has a slightly gray background as opposed to the white space. It is quite evident that the GFS 16 SR image is the most visually comprehensive SR image of this set. These observations are supported by the evaluation results as the GFS 16 SR image had the highest performance out of the GFS experimental group. The GFS 8 run was close behind while the GFS 32 run performed considerably lower, conceivably due to its more prominent dark spot artifacts and background. The GFS 16 SR image has the most suitable predictions for the reflectivity present throughout the storm, its dark spot artifacts are relatively imperceptible, and most all of the color boundaries are represented. Nevertheless, it is also noticeably pixelated and the high-frequency details and object shapes are still not as distinct as in the HR image. This analysis concurs with the other RHIx2 and PPIx2 investigations that determined the GFS 16 SR images to be one of the closest visual representations of the HR image out of their respective experimental groups.

For this experimental group, the reference value of 16 for the NRB parameter had the lowest PSNR of 16.12, the highest MSE of 0.100, and the lowest SSIM of 0.801. However, the highest ranking performance for the NRB parameter, both within the NRB experimental group and overall, was achieved by a NRB of 32. It had the overall highest PSNR of 21.89, the overall lowest MSE of 0.035, and the overall highest SSIM of 0.893. The NRB value of 64 was able to have an equivalent MSE when evaluated. The fastest training time of 6:38:53 was recorded by the 4 NRB run. The slowest training time of 8:23:38 was recorded by the 64 NRB run. These results suggest that increasing the value of the NRB parameter has the effect of increasing training time.

Figure 6.19 contains the SR images for the NRB experimental group. Initial observations will reveal the NRB 16 SR image to be the most dissimilar to the HR target image due to its brown background and the heavy dark spot artifacts within the primary squall line. Most of the color boundaries have been reduced so much so that the object shapes and high-frequency details less defined, although these are relatively comparable to the other SR images, except for the NRB 32 SR image. Most notably, the object shapes and, thus, the high-frequency details are significantly

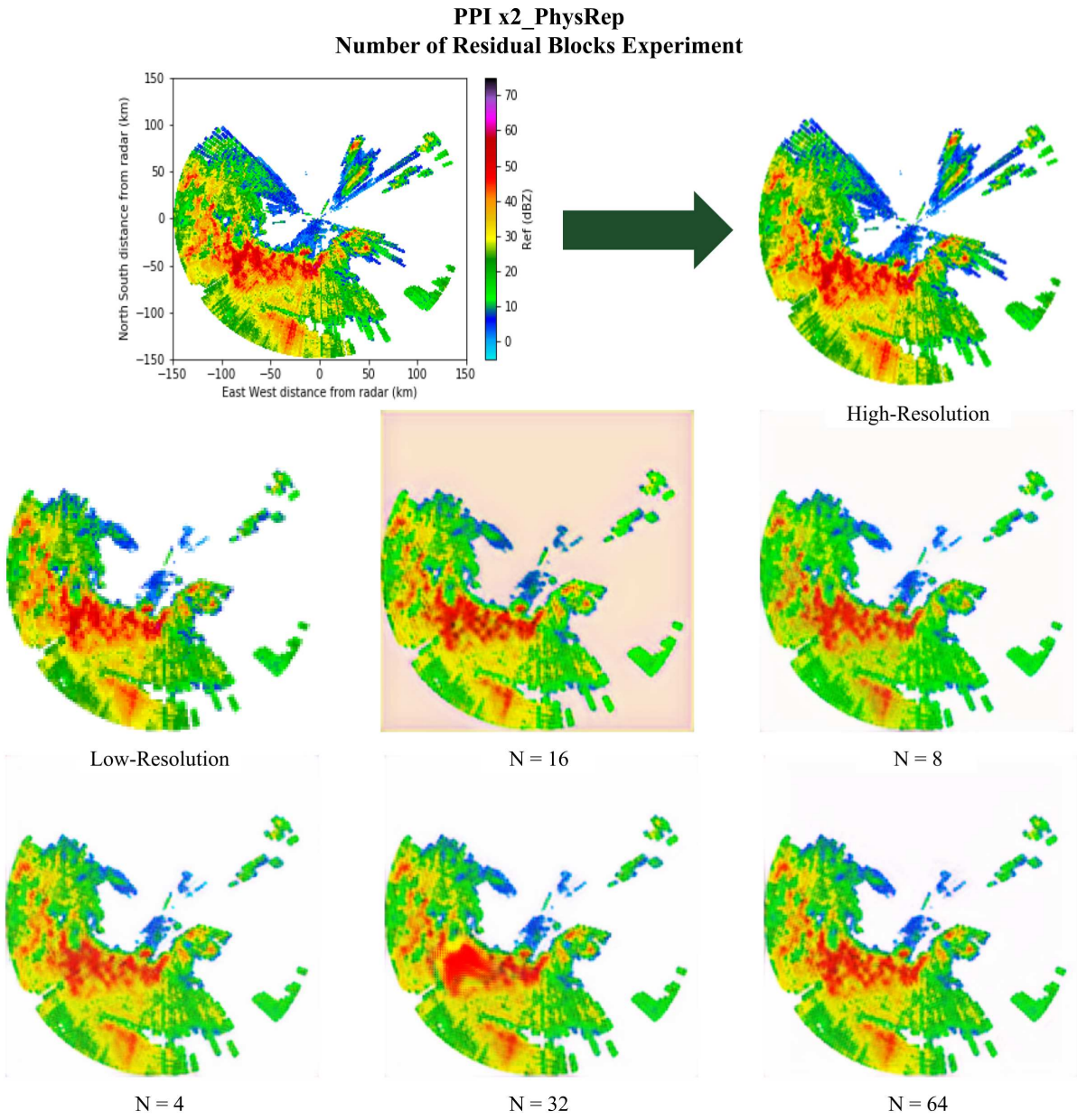

Figure 6.19: PPI x2 Physically Representative Dataset: Number of Residual Blocks Experiment

distorted in the NRB 32 SR image. A majority of the storm structure within this area has been considerably altered from the HR ground truth. So much so that the original feature is unrecognizable in its generated counterpart. Furthermore, a substantial amount of blurring is found alongside a checkerboard-esque artifact which affects the overall visual perceptibility of the primary squall line. In terms of representation, the NRB 16 SR image more fully represents the reflectivity and meteorological features present within the radar scan, despite its background, when compared with

the NRB 32 SR image. This is an important observation to note as the NRB 32 run was determined to be the highest performing out of all the PPIx2_PhysRep experiments, which does not support its visual comprehension. Even though its SR image better retains the storm structure than the NRB 32 SR image, the NRB 16 run was evaluated to be the lowest performing out of the NRB experimental group. These observations suggest that the evaluation metrics do not reliably indicate the perceptual quality of the SR images generated. All of the other SR images are quite similar when comparing their ability to generate the object shapes and high-frequency components of the HR image. They all also contain shadow artifacts along the storm's edge, faint dark spot artifacts as well as edge artifacts along the left and bottom image borders. In general, the NRB 8 SR image has less defined color boundaries and significantly underestimates the reflectivity present in the low-frequency regions around the squall line features. The NRB 4 and NRB 64 SR images are nigh indistinguishable from one another. The main differences between them that can be observed are that the NRB 64 SR image's black spot and shadow artifacts are more apparent while the NRB 4 SR image's color boundaries and high-frequency components are less defined due to blurring and the low-frequency areas around the squall lines contain noticeable pixelation. The NRB 4 SR image is also less affected by the edge artifacts. Overall, these observations follow the evaluation results as the NRB 8, NRB 4, and NRB 64 received the highest evaluations in ascending order, respectively, with the exception of the NRB 32 trial.

The lowest performing set of parameters for the PPIx2_PhysRep experiments was determined to be [32, 64, 16]. The set of parameters that ranked the highest due to it evaluations was [32, 32, 32]. The slowest training was recorded by the [32, 128, 16] test, which is consistent with the previous experiments in which the largest GFS value tested also had the slowest training time. The fastest training time was achieved by the parameter set [32, 16, 16]. The average training time for this SRGAN experimental group is 7:37:03.

6.7 PPIx4 Interpolation Dataset SRGAN

This experimental group's experiments were conducted with SRGAN models trained on the PPIx4_Interp image dataset. The results for these experiments are shown in Table 6.7. The highest performing model out of all of the PPIx4_Interp tests had the DFS set to 16. This experiment run was evaluated to have the highest overall PSNR of 23.18, the lowest overall MSE of 0.024, as well as the highest overall SSIM of 0.904. It should be noted that this is the same parameter configuration that had the highest ranking performance out of the RHIx4 SRGAN models tested. Other experiments also had a MSE of 0.024 such as the DFS of 32 run as well as three other tests within the NRB experimental group. The lowest performing set of parameters within the DFS experimental group was the reference value run. With the DFS value set to 64, the PPIx4_Interp model had the lowest PSNR of 22.57, the highest MSE of 0.027, and the lowest SSIM of 0.891. The fastest training time was achieved by the smallest DFS value of 8 within the DFS experimental group. It had a training time of 7:58:59. The longest training time within the DFS experimental group was recorded as 9:09:42 by the largest DFS value tested of 128. These results suggest a correlation between increasing DFS values resulting in increasing training times.

All of the SR images in Figure 6.20 are quite similar in terms of their general color boundaries, high-frequency details and object shapes. This is supported by the evaluation results as all of these runs had relatively comparable performances on each of the evaluation metrics. The set of DFS SR images all contain dark spot artifacts, pixelation in the lower-frequency regions, and edge artifacts on the left and bottom image boundaries to varying degrees. The DFS 16 and DFS 32 SR images are observed as having the most prominent dark spot artifacts. They also generally underestimate the reflectivity present within the dark green color boundary of the HR image and have less defined object shapes within the squall line features. The visual perceptibility of both the DFS 64 and the DFS 128 SR images is more affected by pixelation than the DFS 8 SR image. In addition, their dark spot artifacts are more noticeable and the DFS 64 SR image, in particular, has a slightly blue background artifact present in the low-frequency area around 50 - 100 km on the North-South axis. From these observations, it is clear that the DFS 8 experimental SRGAN model generated the most

Table 6.7: Experimental Results SRGAN: PPI x4 Interpolation Dataset

Test Parameter	Discriminator Filter Size	Generator Filter Size	Number of Residual Blocks	PSNR	MSE	SSIM	Train Time
Reference	64	64	16	22.57	0.027	0.891	8:21:50
Dis. Filter Size	32	64	16	23.14	**0.024**	0.901	8:09:25
	16	64	16	**23.18**	**0.024**	**0.904**	8:08:51
	8	64	16	22.69	0.027	0.894	7:58:59
	128	64	16	22.81	0.026	0.896	9:09:42
Gen. Filter Size	64	32	16	22.23	0.030	0.884	7:41:24
	64	16	16	21.93	0.032	0.877	**7:05:16**
	64	8	16	21.26	0.038	0.866	7:21:55
	64	128	16	22.93	0.026	0.897	14:02:08
Number of Residual Blocks	64	64	8	22.90	0.025	0.896	8:12:33
	64	64	4	23.10	**0.024**	0.902	8:01:17
	64	64	32	22.89	0.026	0.896	8:50:35
	64	64	64	23.11	**0.024**	0.902	9:37:06
	64	64	128	23.17	**0.024**	0.903	11:06:29

representative SR image when compared to the HR target image. It more closely characterizes the low-reflectivity areas around the squall lines, namely in the dark green color boundary, is the least affected by artifacts, and has the most defined object shapes and high-frequency details in the squall line features when compared against the other SR images in this experimental group.

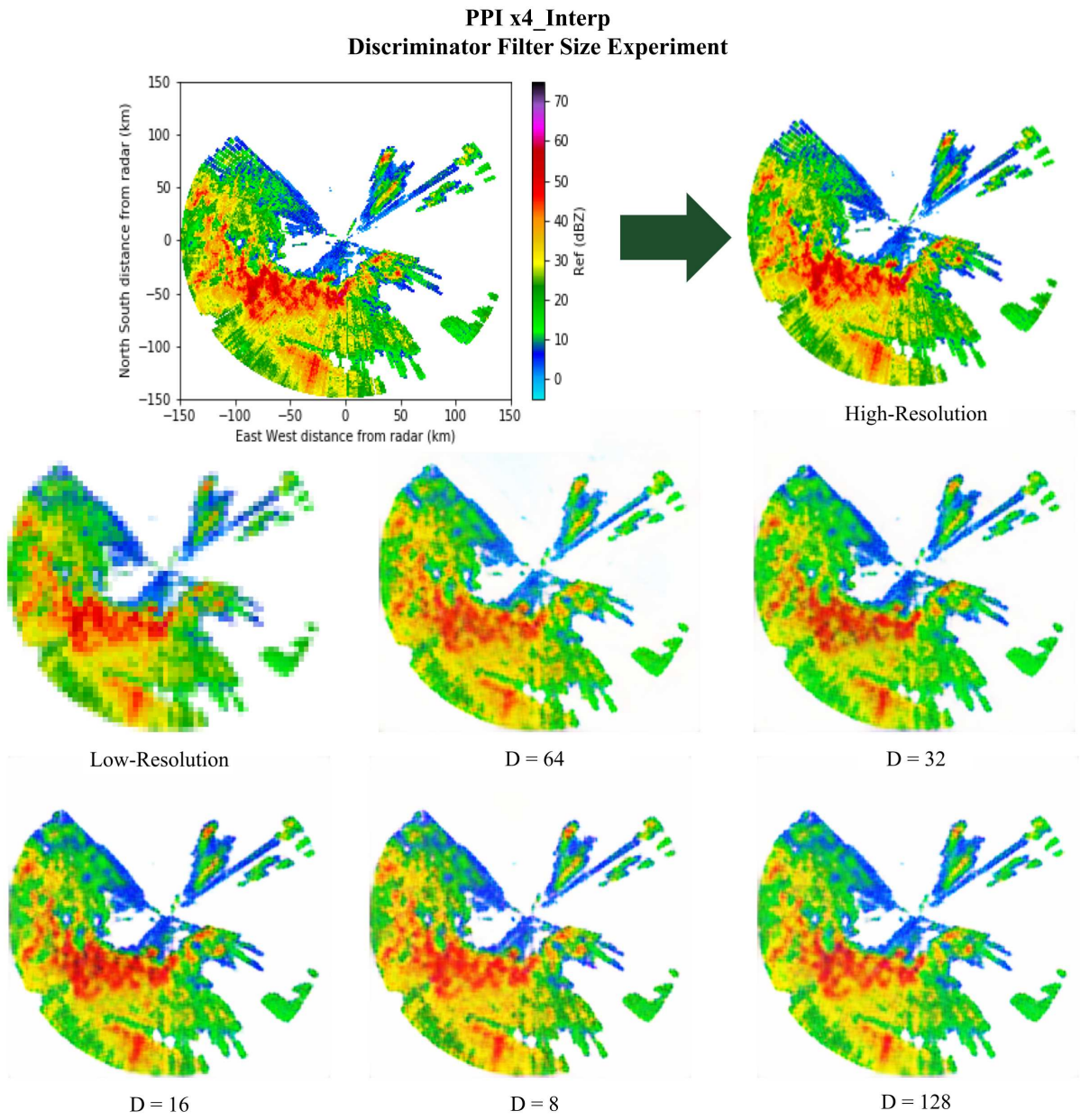

Figure 6.20: PPI x4 Interpolation Dataset: Discriminator Filter Size Experiment

However, these qualitative analyses are not supported by the evaluation results. The DFS 8 trial, observed as being the closest representation of the HR image, was determined to be the second lowest performing experiment out of the DFS experimental group while the DFS 16 trial, observed as having generated one of the most visually distorted SR images, was evaluated as having the highest performance out of the DFS experimental group. This further validates the assertion that

the evaluation metrics are not fully capable of accurately signifying the visual quality of the SR images generated by SRGANs.

The GFS value that outperformed the other GFS experimental trials was a GFS of 128. Within its experimental group, it had the highest PSNR of 22.93, the lowest MSE of 0.026, and the highest SSIM of 0.897. This is in contrast to the RHIx4 SRGAN models tested in which the GFS 128 run had the lowest ranking performance on each of the evaluation metrics overall. Although it had the highest performance within its experimental group, the slowest training time occurred during the GFS 128 run. This was the slowest training time out of all of the PPIx4_Interp runs with a time of 14:02:08. The fastest training time was achieved by the GFS value of 16 with a training time of 7:05:16. Out of all of the PPIx4_Interp trials, the lowest overall PSNR of 21.26, the highest overall MSE of 0.038, and the lowest overall SSIM of 0.866 was recorded for when the GFS value was set to 8.

Figure 6.21 displays the set of SR images for the GFS experimental group. Although their evaluation results are quite comparable to one another, the visual characteristics of these SR images vary considerably from one to the other. Many of these SR images are also significantly distorted from the HR target image. They all contain dark spot artifacts in the squall line features as well as edge artifacts along the left and bottom borders of the image. The GFS 8 and GFS 16 SR images are extensively pixelated and blurred, so much so that the high-frequency components and object shapes of the squall lines cannot be discerned. The GFS 8 SR image appears to have higher values in the lower frequency areas, most of the color boundaries are not displayed, and the reflectivity and object shapes, particularly in the squall line features, are not representative of the HR image. The GFS 16 SR image has slightly better values and color boundaries but the reflectivity and object shapes are still not indicative of the actual storm structure present in the HR image. These observations are supported by the evaluation results as both the GFS 8 and the GFS 16 trials had the lowest performances out of all the PPIx4_Interp experiments tested. The GFS 32 SR image has higher values overall which underestimates the reflectivity being represented. Some of the color boundaries have not been reconstituted and the object shapes within the primary squall line

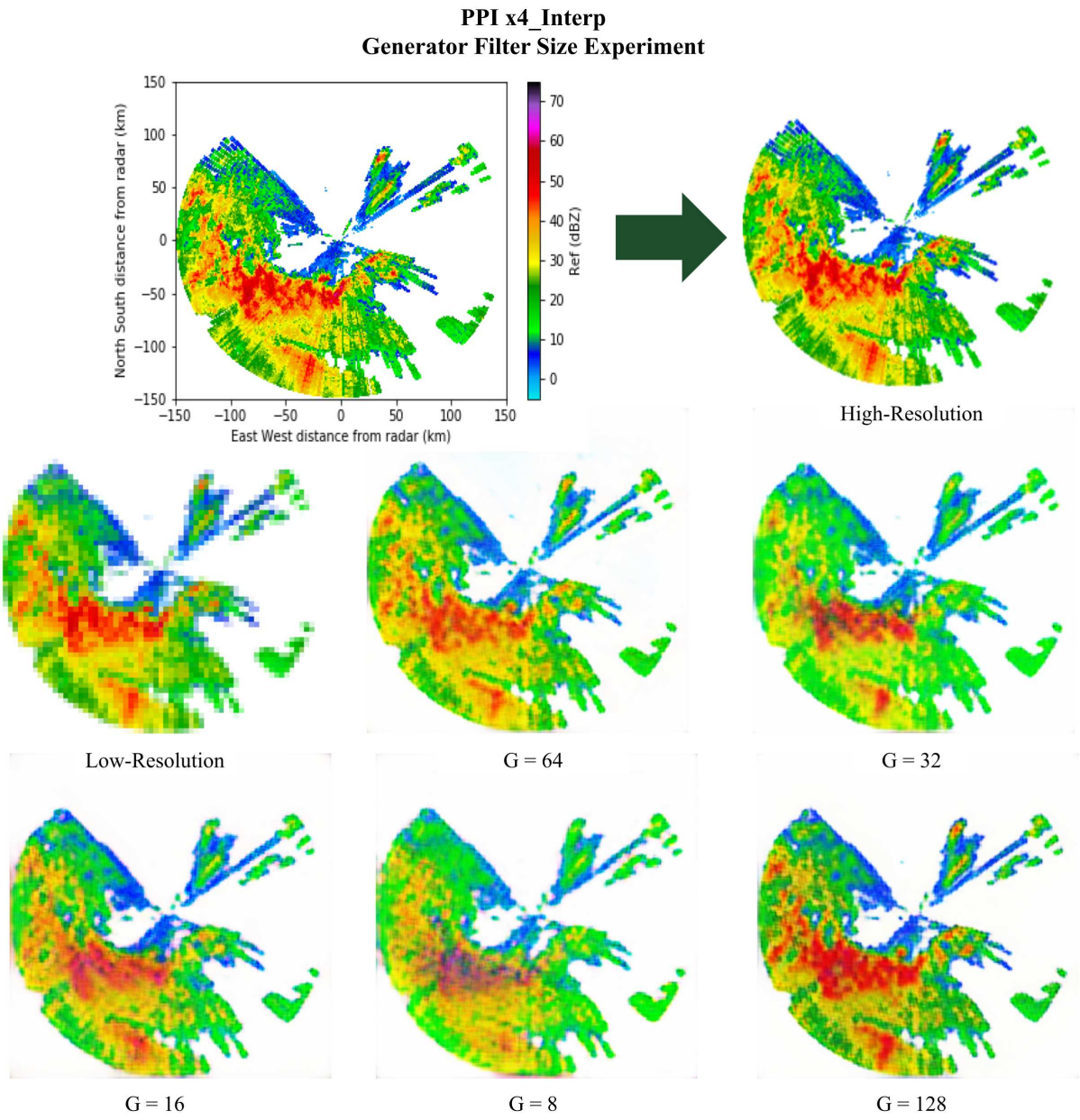

Figure 6.21: PPI x4 Interpolation Dataset: Generator Filter Size Experiment

appear to meld together, obscuring the high-frequency components of the HR image. The GFS 32 SR image also has the most prominent dark spot artifacts in the primary squall line out of the GFS sample SR images. The pros and cons of the final two PPIx4_Interp GFS SR images are dependent upon the context. The GFS 64 SR image exhibits the most distinct object shapes and high-frequency components, representative of the HR image. The GFS 128 SR image has lower values in general which, in turn, more closely portrays the reflectivity present in the areas of lower

reflectivity throughout the radar scan. However, with regard to the overall visual comprehension, the storm features appear more distinct and decipherable in the GFS 64 SR image. The GFS 128 SR image is quite pixelated and the object shapes in the squall line features are thicker and not as defined, making the complex high-frequency components less perceptible. These qualitative analyses are mostly supported by the evaluation results as the GFS 64 and GFS 128 trials had the highest performances out of the GFS experimental group. However, the GFS 128 was evaluated as having a higher performance across all evaluation metrics. This further corroborates the assertion that the evaluation metrics do not consistently reflect the visual quality of the SR images being evaluated.

The 4, 64, and 128 NRB runs had an equivalent evaluation on the lowest overall MSE of 0.024. The NRB value that was evaluated as having the highest performance out of the NRB experimental group was the NRB 128 run. It had the highest PSNR of 23.17, the lowest overall MSE, and the highest SSIM of 0.903 out of the NRB trials. This test value also had the longest training time within the NRB experimental group of 11:06:29. The NRB value that had the fastest training time was when the NRB parameter was set to 4, the smallest NRB value tested. The results in Table 6.7 further supports that the NRB parameter has a direct impact on the training time for the PPIx4_Interp SRGAN models. The set of NRB parameters that had the lowest ranking performance on the evaluation metrics was the reference value test, which also had the lowest ranking performance out of the DFS experimental group.

The example SR images for the NRB experimental group are found in Figure 6.22. This set of SR images have quite comparable visual perceptibility with regards to their general object shapes, high-frequency details and color boundaries. In general, this is in agreement with the results presented in Table 6.7 as the NRB experimental SRGAN models all performed to a similar degree. They all contain edge artifacts along the left and bottom border of the image as well as dark spot artifacts within the primary squall line. The NRB 32 SR image possesses the most prominent dark spot artifacts and has the least defined high-frequency details and object shapes within the primary squall line. The secondary squall line presented in the NRB 32 SR image is also noticeably wider

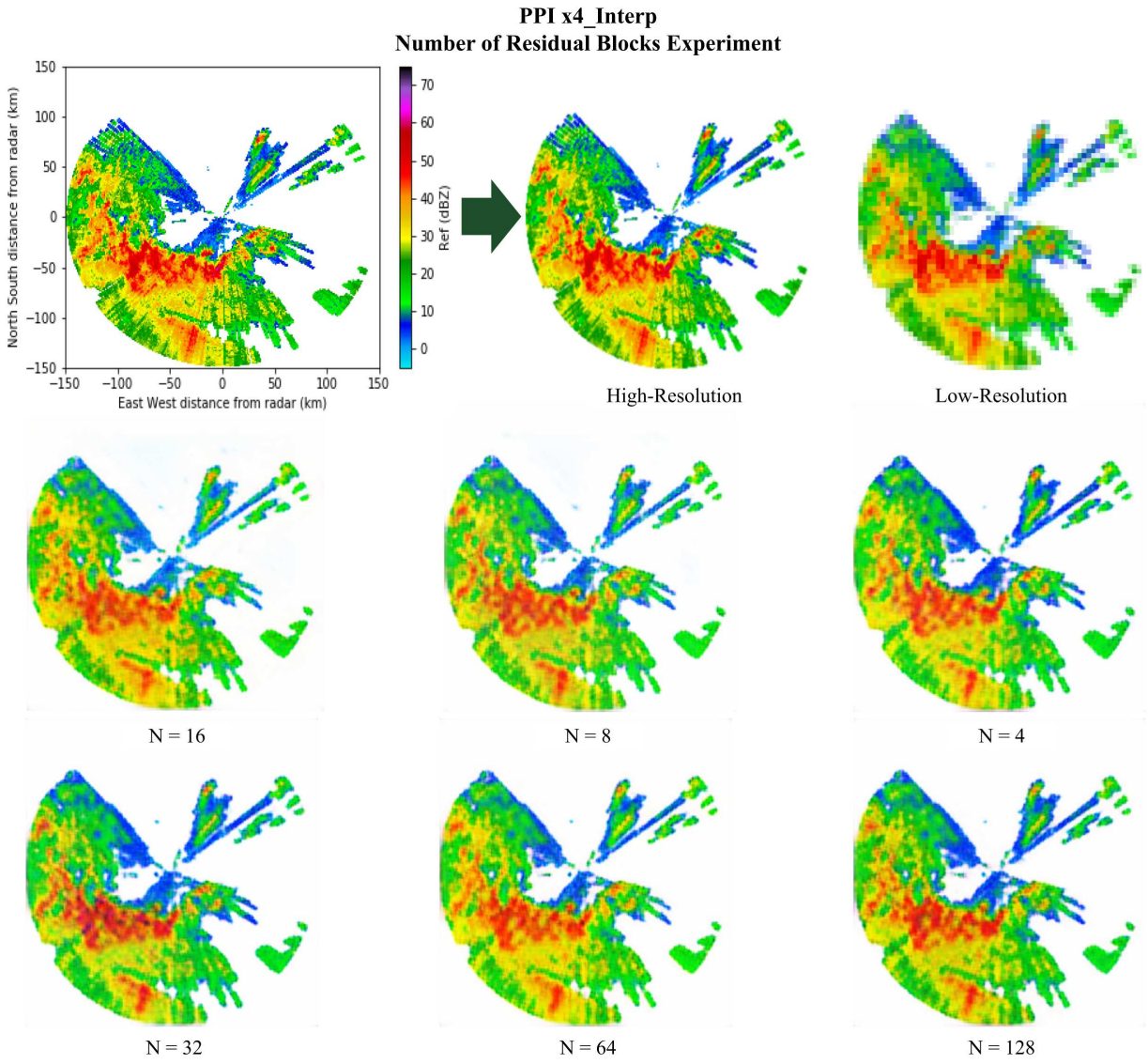

Figure 6.22: PPI x4 Interpolation Dataset: Number of Residual Blocks Experiment

than in the HR image. The NRB 16 SR image is more pixelated and contains lower values than the other SR images, especially within the areas of low reflectivity surrounding the squall line features. This SR image along with the NRB 8 SR image contain a light blue background artifact within the area of low-frequency between -50 and 50 km in the East-West axis and 0 - 100 km in the North-South axis. The NRB 8 SR image also has a noticeably wider secondary squall line and has a more blurred primary squall line than the rest of the SR images. These observations are supported by the evaluation results as the NRB 8, the NRB 16 and the NRB 32 tests had the lowest performances

out of the NRB experimental group. The NRB 64 SR image generally has higher values which results in underestimations in the reflectivity, primarily within the areas of low reflectivity around the squall line features. It also does not represent all of the color boundaries present within the HR image. The NRB 4 and NRB 128 are the closest representations of the HR image. The NRB 4 SR image generally exhibits higher values whilst the NRB 128 SR image generally exhibits lower values which better portrays the color boundaries and reflectivity present within the areas of low reflectivity around the squall line features. Both of these SR images have object shapes in the primary squall line that have merged together, but the NRB 128 SR image has more distinct object shapes and more defined high-frequency details than the NRB 4 SR image. For these reasons, the NRB 128 SR image is observed as having achieved the highest perceptual quality out of the NRB experimental group. These observations further validate the evaluation results as the NRB 4, the NRB 64 and the NRB 128 trials had the highest evaluations out of the NRB experimental group with the NRB 128 trial having the second highest performance out of all the PPIx4_Interp experiments.

The set of parameters that ranked the lowest out of the PPIx4_Interp experiments due to its evaluations was [64, 8, 16]. The highest performing set of parameters was determined to be [16, 64, 16]. This was the same GFS trial that had the highest performance out of the RHIx4 SRGAN models tested. The slowest training time recorded was from the [64, 128, 16] test, which is consistent with all of the previous tests conducted thus far. The fastest training time was achieved by [64, 16, 16]. The average training time for this SRGAN experimental group is 8:50:32.

6.8 PPIx4 Physically Representative Dataset SRGAN

This experimental group's experiments focuses on SRGAN models that were trained on the PPIx4_PhysRep image dataset. The results for these experiments are shown in Table 6.8. When the DFS was set to 16, the PPIx4_PhysRep SRGAN model was able to achieve the highest PSNR of 19.35 as well as the lowest MSE of 0.063 out of its experimental group; however, it also had the second lowest SSIM out of the DFS trials. The highest SSIM of 0.839 was achieved by the

Table 6.8: Experimental Results SRGAN: PPI x4 Physically Representative Dataset

Test Parameter	Discriminator Filter Size	Generator Filter Size	Number of Residual Blocks	PSNR	MSE	SSIM	Train Time
Reference	64	64	16	19.19	0.062	0.839	8:14:23
Dis. Filter Size	32	64	16	18.97	0.065	0.829	7:47:10
	16	64	16	19.35	0.059	0.832	7:51:36
	8	64	16	19.12	0.063	0.833	7:43:17
	128	64	16	19.24	0.061	0.834	9:28:32
Gen. Filter Size	64	32	16	19.09	0.064	0.835	7:05:57
	64	16	16	18.96	0.064	0.823	6:56:47
	64	8	16	18.30	0.076	0.819	**6:41:04**
	64	128	16	9.10	0.497	0.439	13:52:48
Number of Residual Blocks	64	64	8	19.22	0.062	**0.840**	8:06:01
	64	64	4	**19.38**	**0.058**	0.834	7:50:55
	64	64	32	18.81	0.064	0.814	8:40:24
	64	64	64	19.30	0.059	0.833	9:57:50
	64	64	128	18.94	0.062	0.814	13:11:20

reference DFS value of 64. The lowest ranking performance due to its evaluations was the DFS 32 run. It had the lowest PSNR of 18.97, the highest MSE of 0.065, and the lowest SSIM of 0.829 out of the DFS experimental group. The smallest DFS value of 8 had the fastest training time out

of the DFS experimental group of 7:43:17. The largest DFS value of 128 had the slowest training time out of its experimental group of 9:28:32.

Figure 6.23 presents all of the SR images for the PPIx4_PhysRep DFS experimental group. This set of experiments all had comparable performance on their evaluations which is reflected in the similarity of their general object shapes and high-frequency components. Similar to the PPIx2_PhysRep experiments, entire pieces of the storm system, around 0 - 50 km on the East-West axis and around 0 - 100 km on the North-South axis as well as the farther range bins in the top left quadrant of the HR image, are not represented in any of the PPIx4_PhysRep SR images. The physically representative downsampling method is likely responsible for this as these regions of the storm are missing from the corresponding LR images as well. This demonstrates one of the limitations of the SRGAN model, especially when applying them for super-resolving weather radar images. The DFS sample SR images all contain slightly gray backgrounds, edge artifacts on the left and bottom borders of the images, and shadow artifacts along the storm's edges. The DFS 32 SR image has significantly higher values overall. The reflectivity represented is typically underestimated, some of the color boundaries are missing, and the high-frequency details are less defined. This SR image also contains a slight blue background artifact in the North and East regions of the radar scan, a characteristic that it shares with the DFS 8 SR image. The DFS 8 SR image has the lowest values overall and tends to overestimate the reflectivity, primarily within the squall line regions. Furthermore, the high-frequency details are less defined and the object shapes are thicker and tend to meld together in the primary squall line area. This supports the evaluation results as the DFS 32 and the DFS 8 tests were determined as being the lowest performing out of the DFS experimental group. Out of the remaining three SR images, the DFS 16 SR image was determined to be the most visually comprehensive. Even though the DFS 64 contains lower values, is noticeably pixelated, has melded object shapes in the primary squall line, an indistinct object shape for the secondary squall line, and does not represent all of the color boundaries from the HR image, it also better estimates the reflectivity throughout the storm than the DFS 128 SR image. The DFS 128 SR image has more defined object shapes but also contains higher values and significantly

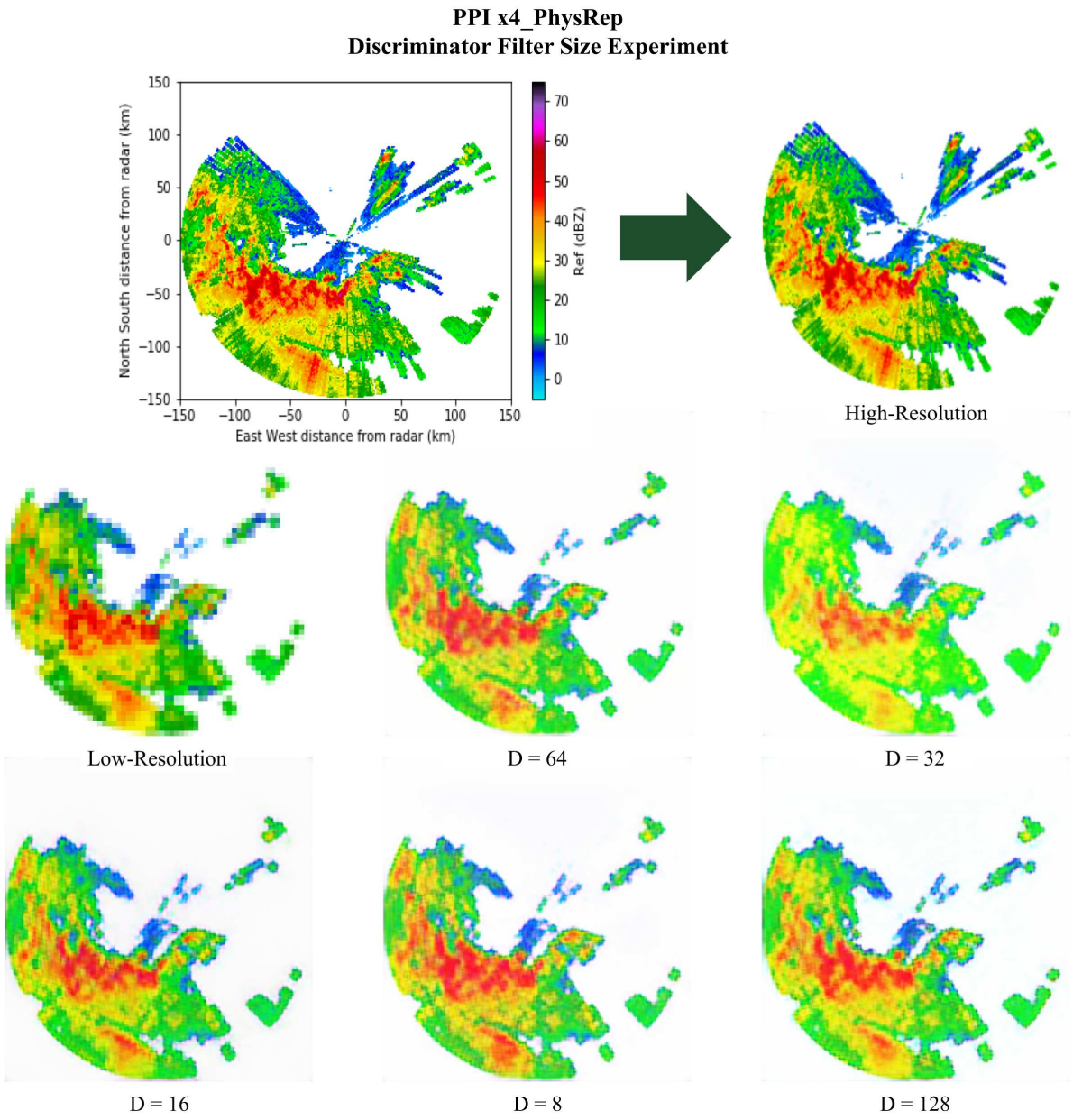

Figure 6.23: PPI x4 Physically Representative Dataset: Discriminator Filter Size Experiment

underestimates the reflectivity, especially within the squall line features including the regions of low reflectivity surrounding them. The DFS 16 SR image portrays a balance between these two, having the most distinct object shapes, especially when observing the squall line features, the best estimations for the reflectivity overall and more fully representing the color boundaries. Nevertheless, there is still noticeable pixelation throughout the DFS 16 generated radar scan, while the DFS 128 SR image is least affected by pixelation, in addition to a considerable shadow artifact along

the storm's edges, whereas the shadow artifact in the DFS 64 SR image is almost imperceptible. However, these characteristics do not have a significant impact on the overall perceptibility of the DFS 16 SR image making it the highest quality SR image, visually, out of the DFS experimental group. These observations also support the evaluation results overall as the DFS 64, the DFS 128, and the DFS 16 trials were evaluated as having comparable marks on the evaluation metrics, with the DFS 16 SRGAN model receiving the highest evaluation out of the DFS experimental group. This is an outlier as all of the previous PhysRep DFS experimental groups observed that the DFS 128 SR image was the most representative of the corresponding HR target image.

For the GFS experimental tests, the GFS run that was evaluated as having the best performance in terms of its evaluation metrics was the reference value run. It had the highest PSNR of 19.19, the lowest MSE of 0.062, and the highest SSIM of 0.839 out of the GFS experimental group. The lowest performing value for GFS was 128. It had the lowest overall PSNR of 9.10, the highest overall MSE of 0.497, as well as the lowest overall SSIM of 0.439. The fastest training time both within the GFS experimental group as well as across all PPIx4_Interp runs was the GFS 8 run. It had a training time of 6:41:04. On the other hand, the slowest training time, both within the GFS experimental group and overall, was recorded by the GFS 128 run with a training time of 13:52:48. These results indicate that increasing the GFS value also increases the training time.

Many of the SR images portrayed in Figure 6.24 are significantly distorted when compared to the HR ground truth image. It should be noted that these SR images are also missing a piece of the storm system shown in the HR image. Overall, the numerical values of the GFS trials' evaluations are quite similar, except for the GFS 128 trial which was determined to be the lowest performing out of all the PPIx4_PhysRep experiments. However, it is evident that the GFS 8 SR image is the least representative of the HR target image as it is missing the squall line features in their entirety. The GFS 16 SR image also has reduced visual quality due to the substantial amount of pixelation throughout the generated radar scan making the object shapes almost indistinguishable. Even though it has well-defined object shapes in the squall line features, the GFS 128 SR image has a yellow background, a moderate level of pixelation as well as dark spot artifacts speckled

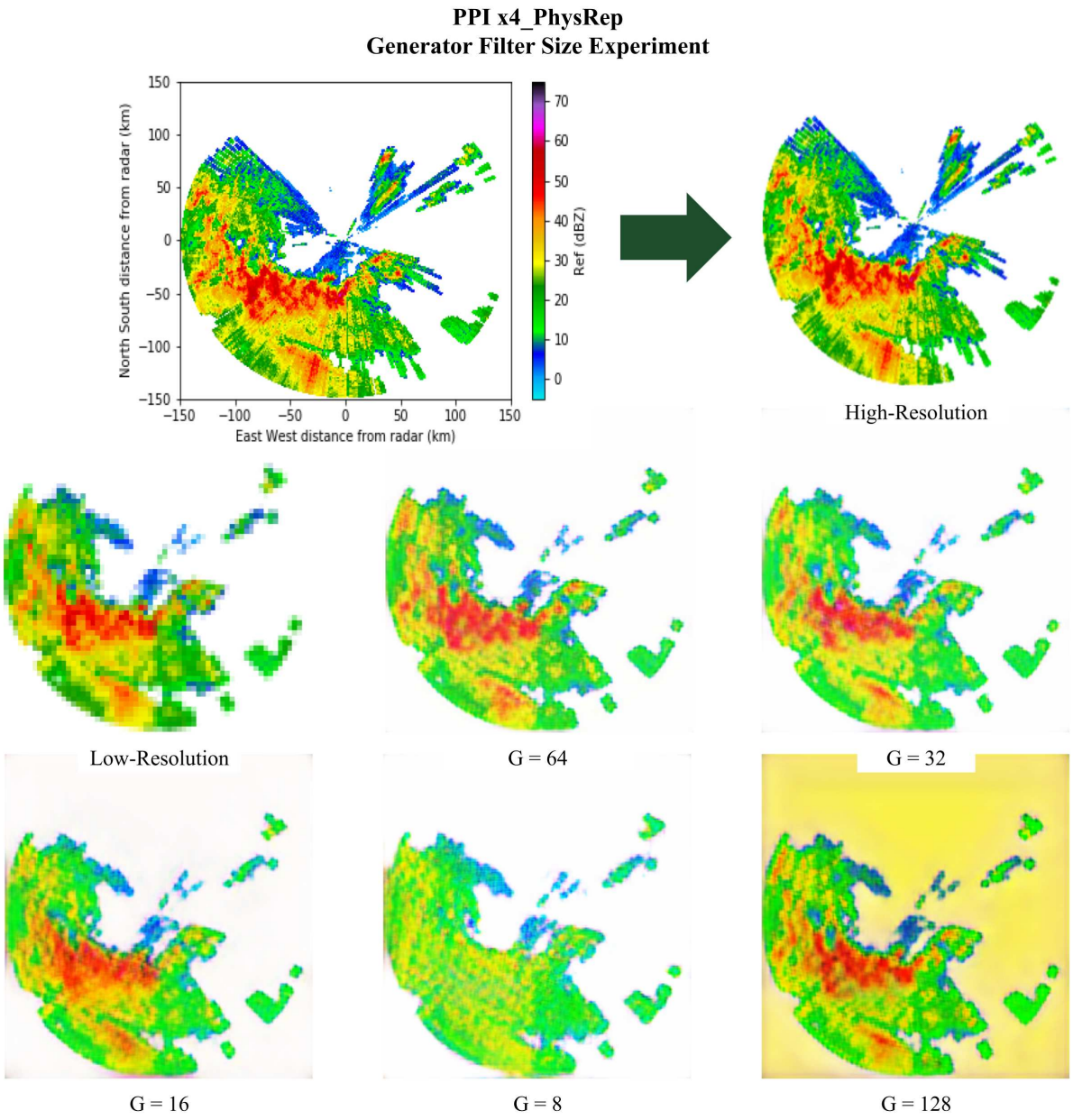

Figure 6.24: PPI x4 Physically Representative Dataset: Generator Filter Size Experiment

throughout the primary squall line. Between the final two, the GFS 64 SR image is the most representative of the HR image. The GFS 32 SR image has slightly higher values, giving it a more faded appearance, and tends to underestimate the areas of low reflectivity surrounding the squall line features. In addition, it contains dark spot artifacts as well as numerous instances of overestimations in the form of pink and purple spots throughout the primary squall line. While the GFS 64 SR image is observed as being the most visually similar to the HR image out of the

GFS experimental group, its overall quality is still found to be inadequate. It contains noticeable pixelation, does not fully represent all of the color boundaries from the HR image, contains blurring that muddles the high-frequency components of the primary squall line, and does not reconstruct the object shape of the secondary squall line accurately. Nonetheless, the GFS 64 SR image is observed as displaying the closest approximation to the HR image out of the GFS experimental group. These observations support the evaluation results presented in Table 6.8. The GFS 64 trial had the highest outcome on its evaluations out of the GFS experimental group. However, it should be noted that the GFS 128 test had a drastically lower performance when compared to the rest of the GFS experiments, even though its generated SR image portrays more of the storm's pertinent meteorological information than the GFS 8 SR image. Most likely, this is a result of the differently colored background. This supports previous assertions that the evaluation metrics do not consistently indicate the visual quality of the SR images generated.

The NRB 4 run had the overall highest ranking performance for the PPIx4_PhysRep SRGAN model tests. It had the highest overall PSNR of 19.38 and the lowest overall MSE of 0.058. Its SSIM was evaluated to be 0.834 which was not the highest SSIM recorded. The NRB 8 run proved to have the highest overall SSIM of 0.840. The NRB value tested that had the lowest ranking performance in terms of its evaluations was when the NRB parameter was set to 8. It was evaluated to have the lowest PSNR of 18.81, the highest MSE of 0.064, and the lowest SSIM of 0.814 out of its experimentation group. The NRB 128 run also had an equivalent SSIM evaluation. The NRB experimentation group was found to have training times that increased as the NRB value was increased. The fastest training time of the NRB experimental group was 7:50:55 from the NRB 4 experimental test. The slowest training time of the NRB experimental group was 13:11:20 from the NRB 128 experimental test. The training times recorded follow the trend that increasing the NRB value will lead to slower training times.

The sample SR images for the NRB experimental group are displayed in Figure 6.25. Overall, these SR images share many similarities regarding their generic object shapes and high-frequency details. This is reflected in the results shown in Table 6.8 as their evaluations are quite similar

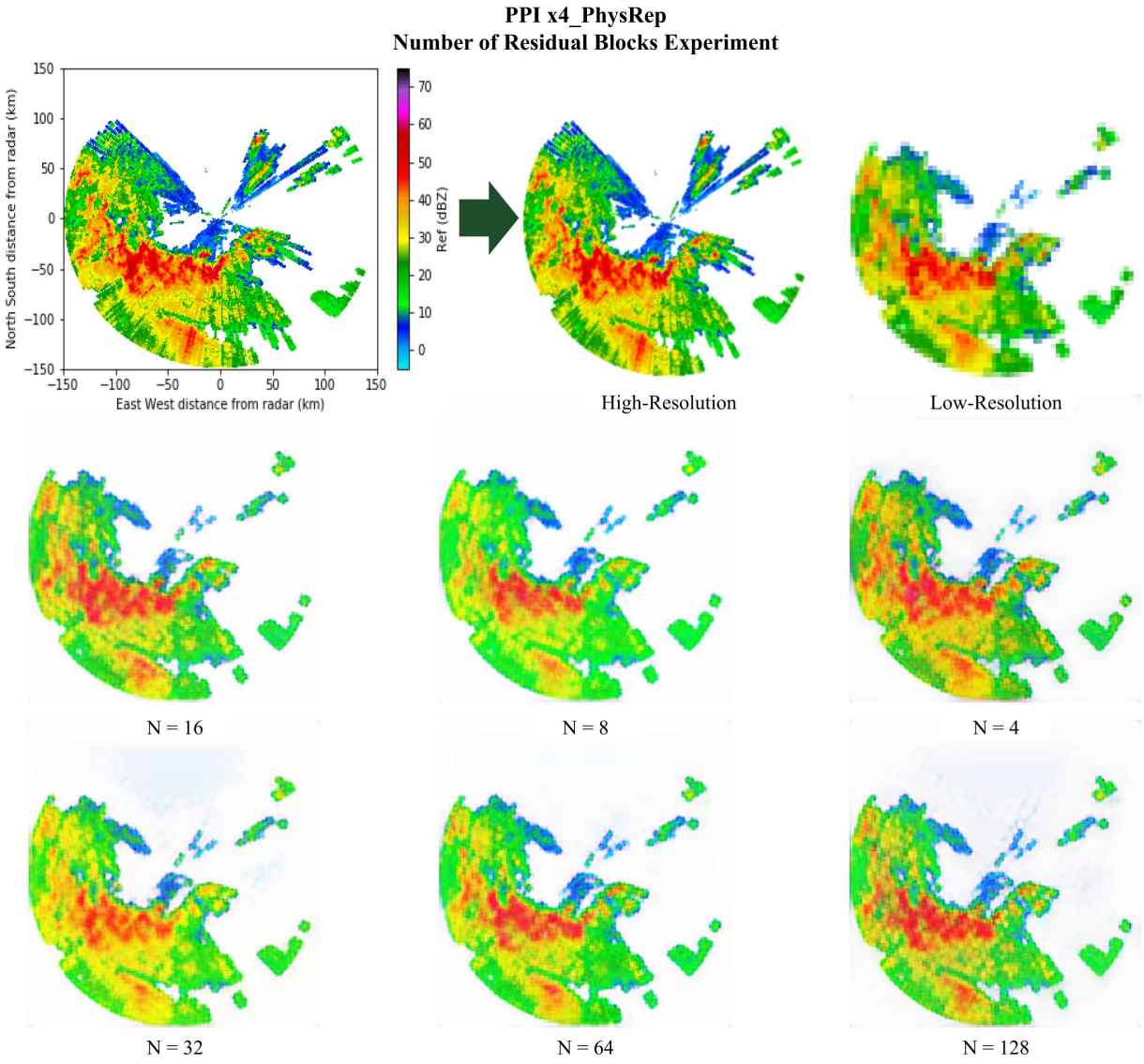

Figure 6.25: PPI x4 Physically Representative Dataset: Number of Residual Blocks Experiment

to one another as well. All of the NRB SR images have edge artifacts on the left and bottom borders of the images to varying degrees and several of them also contain slight background and shadow artifacts reside along the storm edges within the generated radar scans, except for the NRB 8 and NRB 16 SR images. However, it is quite evident that the values of the NRB 8 SR image are significantly higher which underestimates the reflectivity being represented and masks several of the color boundaries found within the HR image. The NRB 32 SR image also contains notably higher values which underestimates the areas of low reflectivity in and around the squall

line features and also does not display all of the color boundaries. It also has quite prominent background and shadow artifacts that can also be observed in the NRB 4, the NRB 64, and the NRB 128 SR images. The NRB 4 SR image generally has lower values, which more closely resembles the HR image overall when compared to the other SR images. However, the high-frequency components are not well-defined, dark spot artifacts can be found in the primary squall line, and there is a significant overestimation of the reflectivity within the primary squall line. This overestimation is located around -90 km on the East-West axis and -50 on the North-South axis where an intensity spot of around 65 dBZ is displayed in the SR image which should be around 55 dBZ according to the HR image. The NRB 64 SR image has well-defined object shapes in the primary squall line when compared to the other SR images but also contains dark spot artifacts and generally underestimates the reflectivity in the areas surrounding the squall line features. Even though it has lower values which more closely matches the HR image, the NRB 128 SR image is considerably pixelated throughout the generated radar scan making the high-frequency components difficult to discern while also containing dark spot artifacts in the primary squall line. Therefore, the NRB 16 SR image is observed as being the closest in resemblance to the HR ground truth image. This is primarily due to its lack of artifacts, its more distinct object shapes, as well as its more fully represented color boundaries. Nevertheless, it still does not have well-defined high-frequency details, does not completely reconstruct the storm's object shapes, especially of the secondary squall line, and generally underestimates the color boundaries surrounding the squall line features. This observational analysis is not fully supported by the evaluation results. The evaluation metrics used determined that the NRB 4 SR image was the highest performing out of all the PPIx4_PhysRep experimental SRGAN models even though it contains many distorting artifacts and is significantly pixelated. The NRB 16 SR image, on the other hand, was in the bottom half in terms of its evaluation performance despite its lack of artifacts. This further suggestes that the evaluation metrics do not consistently reflect the visual quality of the SR images under investigation. It should be noted that the NRB 16 trial was evaluated as having the second highest

SSIM overall. This could suggest that a higher SSIM is indicative of less artifacts within the SR image generated.

The set of parameters that had a substandard performance overall was [64, 128, 16]. This is the same GFS value that proved to have the lowest ranking performance for 3 out of the 4 RHI SRGAN model configurations tested. The set of parameters that were able to outperform all other PPIx4_PhysRep model variations tested was [64, 64, 4]. The slowest overall training time was recorded by [64, 128, 16] while the fastest overall training time was recorded by [64, 8, 16]. Setting the GFS parameter to its largest value consistently resulted in having the slowest training times across all SRGAN models tested, RHI and PPI. Generally, this trend is also observed for most all of the experiments conducted with the fastest training time being recorded by the smaller GFS values tested. The average training time for this SRGAN experimental group is 8:49:09.

6.9 Comprehensive Overview

By themselves, the individual evaluation results and visual comparisons are specific to their type of experiment. This chapter will go beyond this specificity by coalescing the quantitative and qualitative analyses discussed previously with further cross-experiment analyses. This will form a comprehensive overview from which insight into the nature of utilizing SRGAN models for weather radar scan super-resolution can be extrapolated. In order to accomplish this, graphs that show the parameter evaluation trends will be presented and discussed. Figures 6.26 - 6.31 will display the PSNR, MSE, and SSIM evaluation results against the quantities of the architectural parameters, DFS, GFS, and NRB, being tested for each experimental dataset used. The figures will be grouped according to the radar scan type and the architectural parameter under investigation. This way, any patterns concerning the SRGANs' behavior in generating super-resolved images for the particular weather radar scan types of interest can be determined.

Figure 6.26 presents a summary of the results for all the RHI DFS experiments conducted during this thesis study. From these, it is quite apparent that the interpolation dataset experiments had significantly higher performances overall than the physically representative dataset experiments. A

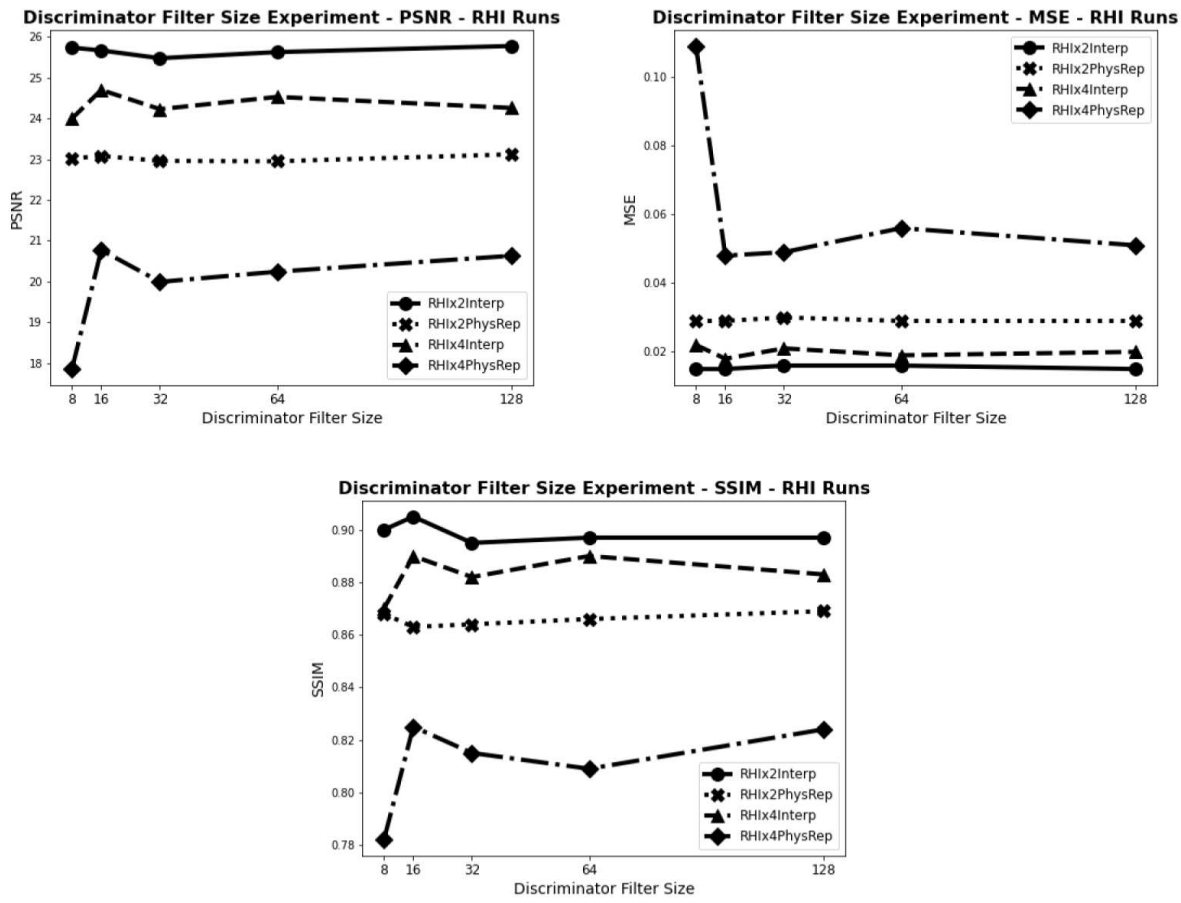

Figure 6.26: Summary of Results for the RHI DFS Experiments

possible explanation for this would be that the SRGAN model architecture from [14] was originally designed with the consideration that the downsampling method used to create the LR input dataset would be interpolation. To this end, further development of the SRGAN architecture that better accommodates other downsampling methods will need to be explored, especially when developing SRGAN models with actual low-resolution weather radar scans as inputs. Overall, the variations in the DFS parameter do not affect the results significantly across the RHI experiments. The trends of the graph lines in Figure 6.26 suggest that, as the DFS increases, the performance of the resulting generated SR images generally improves across all the evaluation metrics under investigation. It should also be noted that the four times resolution scaling dataset experiments had a distinct peak in their quantitative performance with a DFS of 16 while the quantitative performance of the

two times resolution scaling dataset experiments remained relatively consistent regardless of the changes to the DFS parameter. Through Chapters 6.1 - 6.4 and the DFS experiments' summary of results graphs, it has been observed that the evaluations for the RHI DFS experiments had the highest outcomes when the DFS was set to 16 and 128. These values performed well for both the quantitative and qualitative evaluations, performing in the top half throughout most all of the analyses. Since the visual quality of the generated SR images are vital to the radar engineers and researchers that would be analyzing the data, in addition to the observation that the interpolation dataset experiments were not as affected by changes in the DFS parameter, it can be hypothesized that a DFS parameter set to 128 would produce the highest performing SRGAN models for the RHI dataset in general. The DFS 128 SR images had the highest perceptibility for the physically representative dataset experiments and performed quite well on the evaluation metrics too. For future research, it would be beneficial to consider having a higher DFS parameter in order to improve the overall visual quality of the generated SR images, especially for higher resolution scales and both physically representative downsampling methods and real LR radar scans.

The graphs displayed in Figure 6.27 show the results for all the RHI GFS experiments summarized into three graphs, one for each of the evaluation metrics utilized. These also reveal that the interpolation dataset experiments had higher performances overall than the physically representative dataset experiments, except for the RHIx2_PhysRep experiments which was not substantially affected by changes in the GFS parameter. This was particularly evident when the GFS was set to its highest quantity of 128 and all the performances for the other experiments decreased drastically. These observations suggest that the GFS parameter has a significant affect on the RHI SRGAN models' performances in general, especially when set to higher quantities. Through Chapters 6.1 - 6.4 and the summary of results graphs, it has been observed that the evaluations for the RHI GFS experiments unanimously had the highest outcomes when the GFS parameter was set to 64 for most all of the quantitative and qualitative analyses conducted. The only exception to this was the RHIx4_PhysRep experiment which evaluated the GFS 8 trial as being the highest performing. But even then, the GFS 64 was observed as having the most perceptibility when compared to the

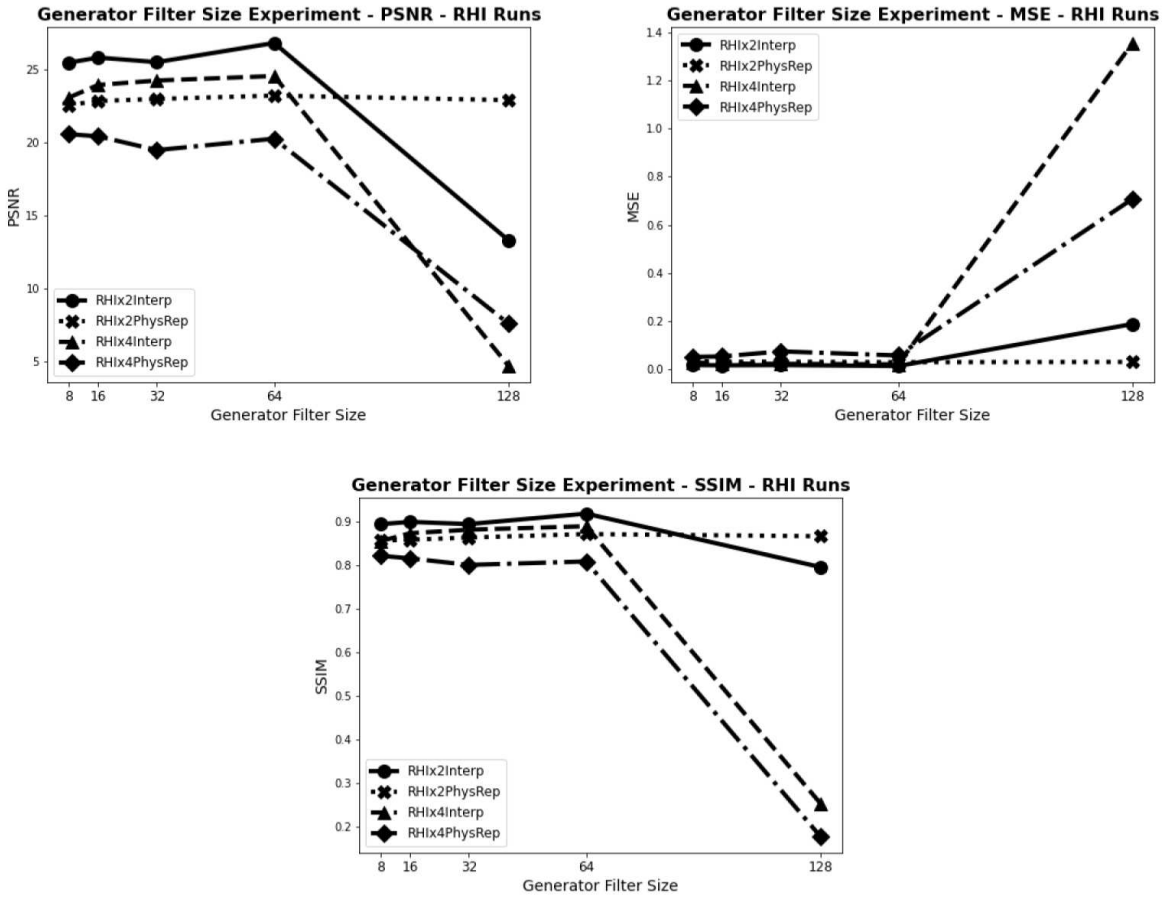

Figure 6.27: Summary of Results for the RHI GFS Experiments

other SR images. Thus, it can be hypothesized that a GFS parameter set to 64 will result in the highest visual quality and highest quantitatively performing SRGAN models for the RHI dataset in general, which will be especially important for further research endeavors.

A summary of the results for the RHI NRB experiments are illustrated in Figure 6.28. This set of graphs further shows how the interpolation dataset experiments were evaluated as generally having higher performances than the physically representative dataset experiments. The only exceptions were the RHIx4_Interp experiments conducted for the lowest NRB quantity of 4 and the highest NRB quantities of 64 and 128. However, the priority for future research efforts is to utilize actual LR radar scans when training the SRGAN models. Since the physically representative dataset experiments are more indicative of the actual LR radar scans than the interpolation-based

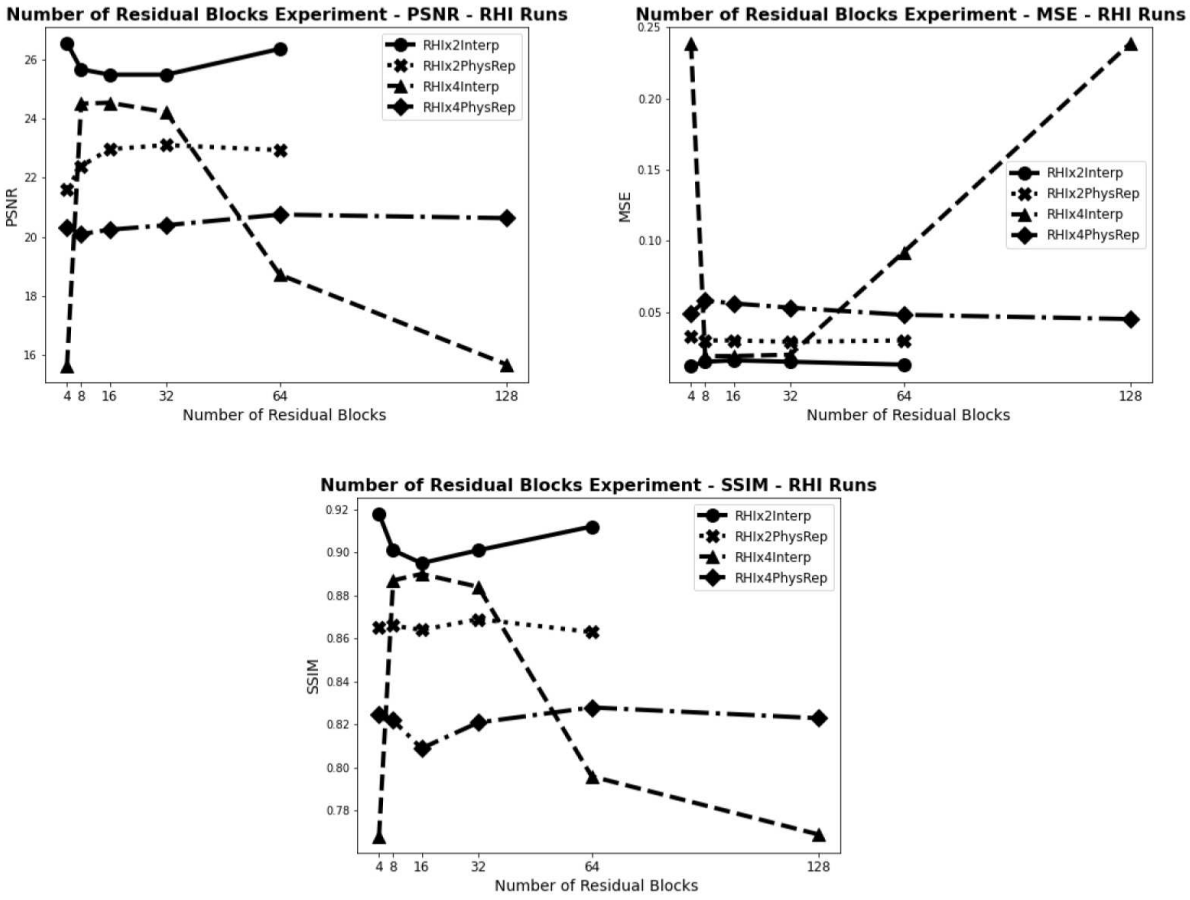

Figure 6.28: Summary of Results for the RHI NRB Experiments

dataset experiments, these observations are not as pertinent to the overall objective. Chapters 6.1 - 6.4, in addition to the summary of results graphs, have shown that the evaluations for the RHI NRB experiments had their overall highest performances when the NRB parameter was set to 64, especially the qualitative analyses for the physically representative dataset experiments that evaluated the generated SR images' visual quality. This was exemplified in the RHIx4_PhysRep experiment as the NRB 64 trial achieved the highest evaluations overall on the evaluation metrics as well as on its perceptibility. For future research efforts, it would be beneficial to consider having a NRB of 64 to improve the overall visual quality of the generated SR images, especially for higher resolution scale tests and both physically representative and actual LR radar scan datasets.

Figure 6.29 presents a summary of the results for all the PPI DFS experiments in order to better determine the nature of utilizing SRGAN models for weather radar scan super-resolution. The graphs illustrated in this figure exhibit further evidence for the previous assertion that the interpolation dataset experiments performed significantly higher overall than their corresponding physically representative dataset experiments. In addition, the variations in the DFS parameter do not affect the results significantly across the PPI experiments. These observations generally agree with the findings from the RHI summary of results graphs as well. Another observation gleaned from the graphs in Figure 6.29 is that the PPIx2 resolution scale experiments had a significant drop in their quantitative performances for the DFS 32 trial while the PPIx4 resolution scale experiments have little variation in their performances regardless of the changes in the DFS parameter. Through Chapters 6.5 - 6.8 and the DFS experiments' summary of results graphs, it has been observed that the evaluations for the PPI DFS experiments ranked the highest for when the DFS parameter was set to 16 and 64. The DFS 16 consistently outperformed the other DFS parameter quantities tested both quantitatively and qualitatively, being ranked as the most visually understandable SR image for three out of the four sets of dataset experiments. It even had the highest overall performance on the evaluation metrics for both of the PPI interpolation dataset experiments tested. For these reasons, it could be hypothesized that a DFS parameter set to 16 would result in the highest performing SRGAN models for the PPI dataset in general. However, when focusing primarily on the physically representative datasets, it could be argued that a DFS of 64 would be the recommended quantity because it was the second highest performing DFS parameter for both of the physically representative dataset experiments. This would be a particularly important consideration for future research studies focusing on physically representative downsampling methods and real LR radar scan inputs.

The graphs displayed in Figure 6.30 show the results for all the PPI GFS experiments. In general, the graphs further support previous assertions that the interpolation dataset experiments had higher performances in general than their physically representative dataset experiment counterparts. It is interesting to note that both of the PPIx2 resolution scale experiments had significant

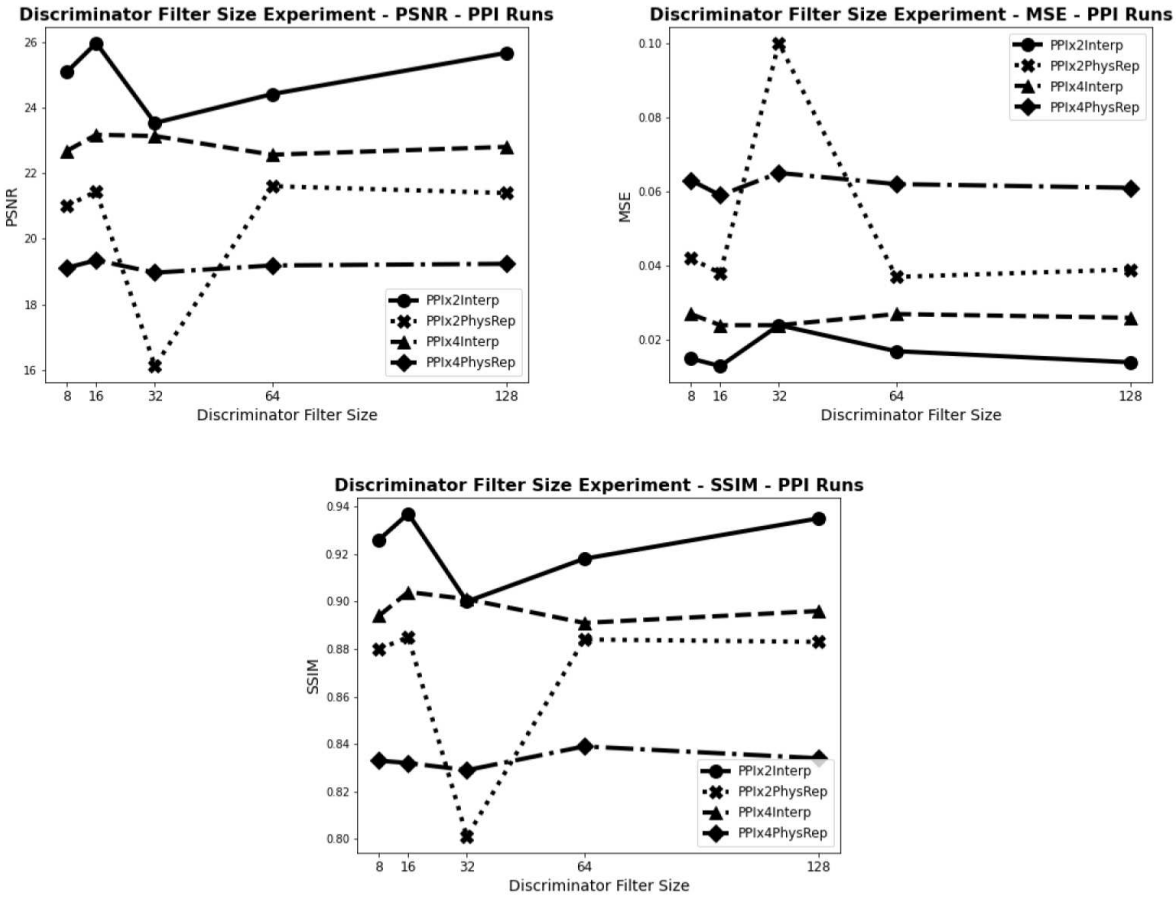

Figure 6.29: Summary of Results for the PPI DFS Experiments

decreases in their performances for GFS quantities higher than 16, especially when the GFS parameter was set to 64. From these graphs, it can be observed that the GFS parameter generally has a substantial impact on the performances of the PPI SRGAN models, especially when set to higher quantities. This deduction was also made for the RHI GFS experiments. Through Chapters 6.5 - 6.8 and the GFS experiments' summary of results graphs, it has been observed that the evaluations for the PPI GFS experiments had the highest outcomes when the GFS parameter was set to 16 for the PPIx2 resolution scale experiments and 64 for the PPIx4 resolution scale experiments. These quantities achieved the most of the highest results for their respective resolution scale experiments both quantitatively and qualitatively. If it is necessary to use a single GFS parameter setting for all experiments, a GFS of 32 should be considered. The SR images produced from the GFS 32

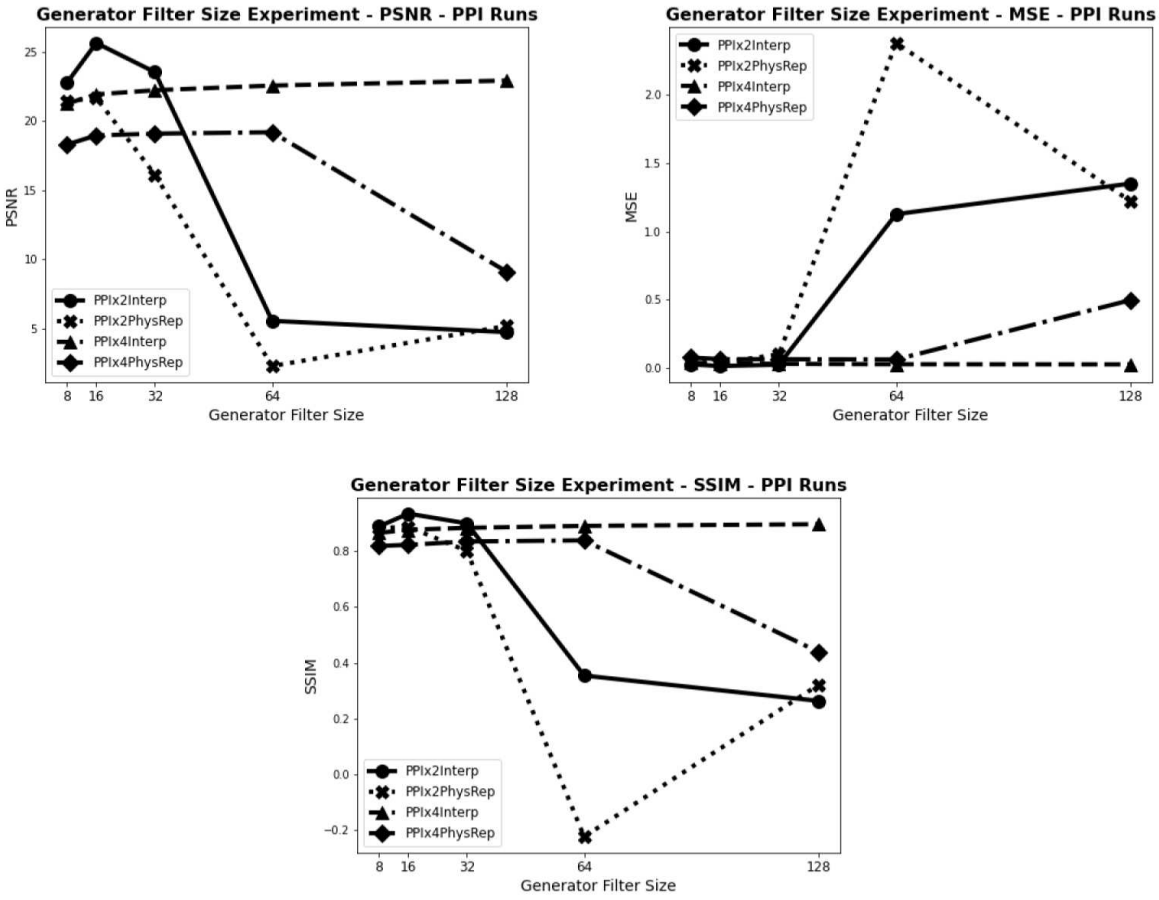

Figure 6.30: Summary of Results for the PPI GFS Experiments

experimental SRGAN models all received high rankings on their perceptibility for each of the dataset experiments. These proposed GFS quantities will be especially useful for further research endeavors.

A summary of the results for the PPI NRB experiments are illustrated in Figure 6.31. This set of graphs provides evidence that the interpolation dataset experiments were evaluated as generally having higher performances than the physically representative dataset experiments. The only exception was the trial where the NRB was set to 64 in which the performance of the PPIx2_Interp dataset experiment declined significantly. In general, however, the NRB parameter appears to not significantly affect the quantitative evaluation results, especially for the PPIx4 resolution scale dataset experiments, except for a couple outliers. Chapters 6.5 - 6.8, in addition to the summary

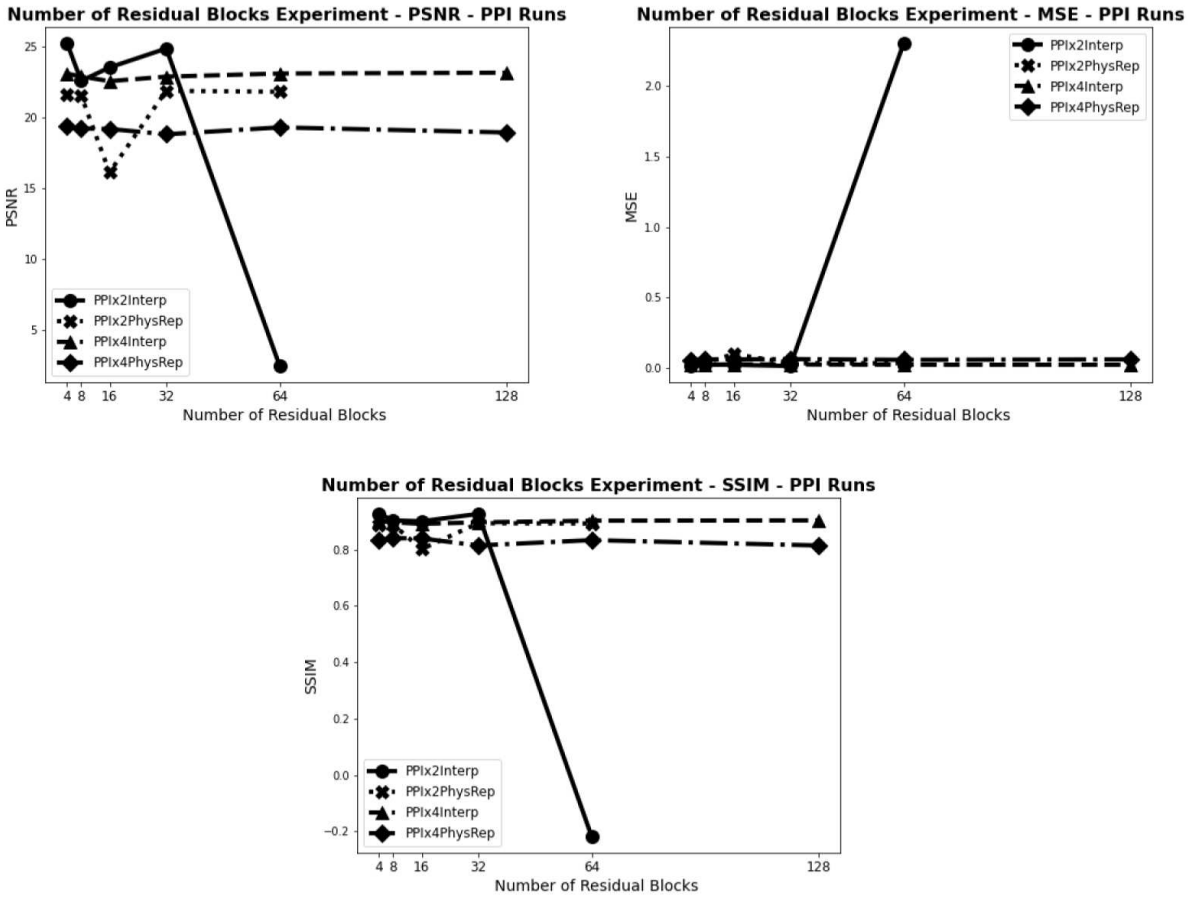

Figure 6.31: Summary of Results for the PPI NRB Experiments

of results graphs, have shown that the evaluations for the PPI NRB experiments had their overall highest performances when the NRB parameter was set to 4 for the PPIx2 resolution scale dataset experiments and 128 for the PPIx4 resolution scale dataset experiments, especially with regards to their visual perceptibility. If it is necessary to use a single NRB parameter setting for all experiments, a NRB of 4 is suggested as it had the highest performance, visually, for the PPIx2_PhysRep experiment and achieved comparable visual performances on both of the interpolation dataset experiments as well. For future research efforts, it would be beneficial to consider these recommended NRB quantities.

6.10 Baseline Comparisons

Both the quantitative and qualitative evaluation results of the experimental SRGAN models have been thoroughly examined throughout Chapters 6.1 - 6.9. So far, these analyses have focused on comparing the performances of the experimental models against one another in order to study the difference in their behaviors when subject to changes in their architectural composition. The analyses were also used to determine the highest performing sets of parameters amongst them. This chapter aims to evaluate the highest performing SRGAN models against the baseline interpolation methods discussed in Chapter 5.2. In order to accomplish this task, Table 6.9 presents the quantitative evaluation results of the experimental SRGAN models alongside the baseline interpolation methods. Only the evaluation results from the highest ranking experimental SRGAN models that were determined through Chapters 6.1 - 6.8 are shown as they will be the focus of discussion from this point onward. Figures 6.32 and 6.33 depict the sample super-resolved images generated by the highest ranking experimental SRGAN models as well as the HR image output by the baseline interpolation methods for the RHI and PPI dataset experiments, respectively. From the table and the figures, comparisons and analyses are drawn as to the efficacy of utilizing SRGAN models for super-resolving weather radar images.

Table 6.9 shows the evaluation results of baseline models compared to the highest performing experimental SRGAN models described throughout Chapters 6.1 - 6.9. For both of the RHIx2 resolution scale dataset experiments, the baseline interpolation methods have notably higher performances when compared to the experimental SRGAN models; all except for the Nearest Neighbors method when compared to the RHIx2_PhysRep SRGAN model. This is especially prominent in the PSNR and SSIM evaluations in which the evaluation results between the baseline methods and the experimental SRGAN models differ more considerably than in the MSE evaluations. In both of the RHIx4 resolution scale dataset experiments, on the other hand, the experimental SRGAN models generally outperform the baseline interpolation methods. The only exception to this statement is the bicubic interpolation method when compared to the RHIx4_Interp SRGAN model. They have a comparable performance as they are evaluated as having the same MSE evaluation, with

Table 6.9: Result Assessment: SRGAN vs Baseline Models

Dataset	Model	PSNR	MSE	SSIM	Dataset	Model	PSNR	MSE	SSIM
RHIx2 Interp	[32, 64, 16]	26.79	0.012	0.919	PPIx2 Interp	[16, 32, 16]	25.98	0.013	0.937
	Bicubic	27.85	0.009	0.946		Bicubic	27.64	0.008	0.961
	Nearest Neighbors	26.83	0.011	0.944		Nearest Neighbors	25.76	0.013	0.954
	Lanczos	27.91	0.009	0.945		Lanczos	27.79	0.008	0.960
RHIx2 PhysRep	[32, 64, 16]	23.19	0.028	0.872	PPIx2 PhysRep	[32, 32, 32]	21.89	0.035	0.893
	Bicubic	23.34	0.028	0.892		Bicubic	22.42	0.032	0.906
	Nearest Neighbors	22.90	0.030	0.891		Nearest Neighbors	21.66	0.037	0.901
	Lanczos	23.32	0.028	0.890		Lanczos	22.40	0.032	0.905
RHIx4 Interp	[16, 64, 16]	24.70	0.018	0.890	PPIx4 Interp	[16, 64, 16]	23.18	0.024	0.904
	Bicubic	24.61	0.018	0.892		Bicubic	23.15	0.024	0.890
	Nearest Neighbors	23.50	0.024	0.890		Nearest Neighbors	21.86	0.032	0.884
	Lanczos	24.66	0.018	0.890		Lanczos	23.24	0.023	0.885
RHIx4 PhysRep	[64, 64, 64]	20.75	0.048	0.828	PPIx4 PhysRep	[64, 64, 4]	19.38	0.058	0.834
	Bicubic	20.51	0.054	0.835		Bicubic	19.89	0.053	0.844
	Nearest Neighbors	20.05	0.059	0.836		Nearest Neighbors	19.21	0.061	0.841
	Lanczos	20.50	0.054	0.831		Lanczos	19.88	0.053	0.840

the bicubic interpolation method having a marginally higher SSIM and the RHIx4_Interp SRGAN model as having a higher PSNR. Thus, generally, the SRGAN models tend to outperform the baseline methods for the RHI datasets at higher resolution scales. It is hypothesized that a possible reason for this behavior is because the lower resolution scaling does not significantly affect the

appearance of the radar scan from the HR image to the LR image – even when using the physically representative downsampling method – especially for the regions within the closer ranges of the radar scan, giving the interpolation methods an advantage. However, when the structure and features of the LR input image are significantly changed from the HR image due to the downsampling process, as in the RHIx4 resolution scale experiments, it is observed that the SRGAN models are more suitable for super-resolving the RHI LR images.

This hypothesis is noted as being consistent with the results from the RHI downsampling method experiments as well. In general, the RHI_Interp evaluation results are significantly higher than their corresponding RHI_PhysRep evaluations. A prospective explanation for this behavior is that the interpolation downsampling method does not significantly affect the structure or features of the input HR image, resulting in a LR image that appears slightly blurred in comparison, albeit very similar to its respective HR image. Furthermore, as described in Chapter 4.3, the LR images are created by applying one of the downsampling methods to it and then reducing the size of the image by a factor of the respective resolution scale using an interpolation kernel. Since the resulting LR image's features and structure were generated by an interpolation kernel, it stands to reason that the RHI_Interp models would have an advantage when super-resolving the LR images than their RHI_PhysRep counterparts. Nevertheless, the RHI_PhysRep SRGAN models perform better relative to the baseline interpolation methods than the RHI_Interp SRGAN models. The RHIx2_Interp SRGAN model is outperformed by every baseline interpolation method across all the evaluation metrics tested. Meanwhile, the RHIx2_PhysRep SRGAN model outperforms the corresponding nearest neighbors method's evaluation results and has a much closer MSE evaluation to the other baseline methods, comparatively. The RHIx4 SRGAN models both have higher performances than their respective baseline methods evaluation results. A notable difference between these performances, however, is that the RHIx4_Interp SRGAN model's MSE and SSIM evaluation results are quite similar to the baseline methods' evaluation results. The RHIx4_PhysRep SRGAN model's evaluation results, in contrast, exhibit considerably higher performances than the baseline methods across all of the evaluation metrics investigated. The physically representative downsampling

method affects the actual structure of the HR image in way that better reflects a real low-resolution radar scan. This is particularly apparent in the RHIx4_PhysRep LR image shown in Figure 6.32. The physically representative downsampling method introduces a segmented characteristic into the resulting LR images of the RHI_PhysRep experiments. This segmented nature is propagated through in the baseline interpolation methods and is shown in their resulting SR images. The SRGAN SR images, on the other hand, do not exhibit this characteristic and are observed as reconstructing the natural shape of the HR image more closely. Overall, this observation is paramount in determining the SRGAN model's efficacy as it demonstrates that the SRGAN can outperform the baseline methods in super-resolving weather radar images, in terms of their visual quality, especially with the more physically representative LR image input datasets. This suggests that the SRGAN model would be a more effective SR technique in real-world implementations. This provides further support for previous assertions that the evaluation metrics do not reflect the visual quality of the SR images that they are evaluating.

Figure 6.32 aptly illustrates the differences between the baseline interpolation methods' output HR images and the SRGAN model's generated SR images for the RHI dataset experiment comparisons. All of the baseline interpolation methods produce very similar HR images compared to each other for each of their respective dataset experiments. The main difference is revealed within the four times resolution scale dataset SR images. The nearest neighbors output image appears more pixelated than the other baseline interpolation methods which appear more blurred. All of the SRGAN SR images are observed as slightly underestimating the reflectivity within the updraft regions and the low-frequency region of the second system. The RHIx2_Interp SR image is visually similar to the baseline interpolation methods' output images in general. Some of the high-frequency details are, however, obscured within the upper levels of the first storm system. Overall, this does not negatively affect the perceptibility of the SR image. The RHIx2_PhysRep SRGAN SR image also exhibits less distinct high-frequency details. However, while the segmented characteristic of the RHIx2_PhysRep LR image is still observed within the baseline interpolation methods' output HR images, the affect of the physically representative downsampling method is mitigated in

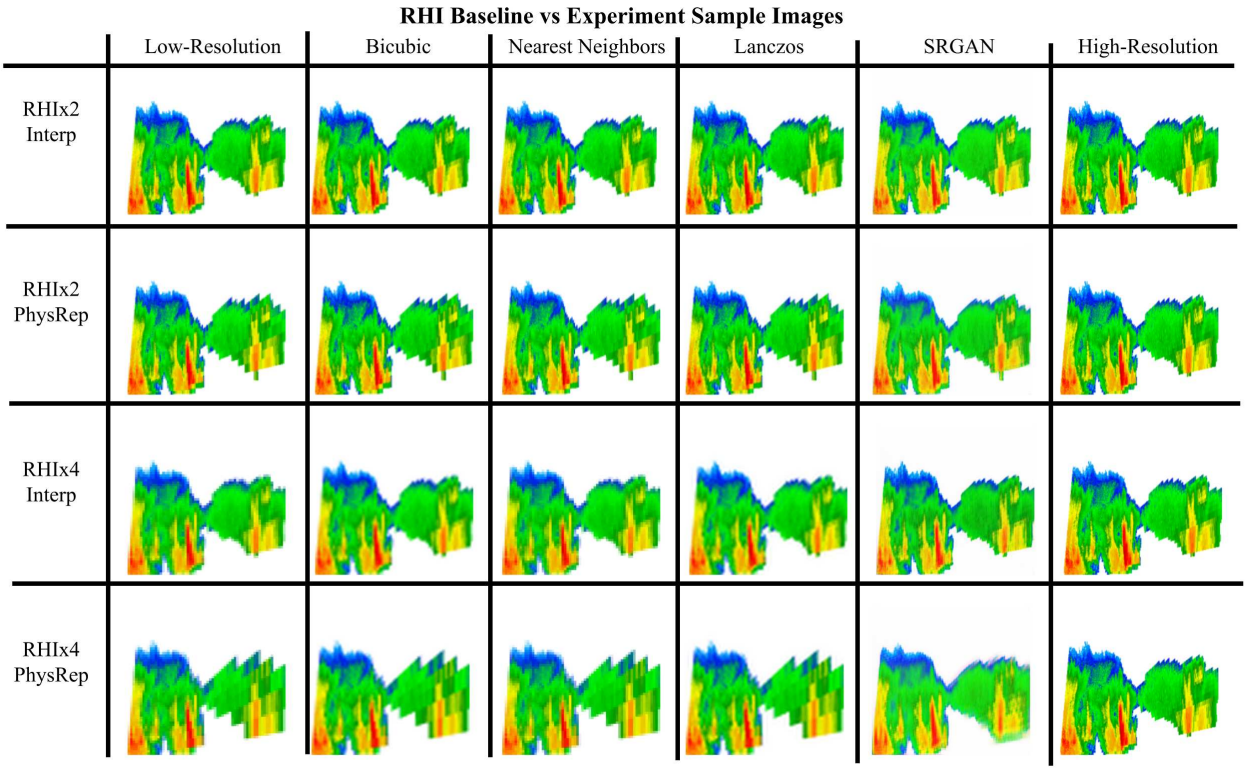

Figure 6.32: RHI Baseline vs Experiment Examples

the RHIx2_PhysRep SRGAN SR image. The RHIx2_PhysRep SRGAN model appears to have achieved a more natural representation of the HR image than the baseline interpolation methods as the color boundaries and object shapes within the second system are significantly more distinct in the RHIx2_PhysRep SRGAN SR image. The HR images from the RHIx4_Interp baseline interpolation methods are considerably blurred making the high-frequency details and object shapes less defined. These characteristic features are better defined in the RHIx4_Interp SRGAN SR image, despite the pixelation within the second half of the second system of its SR image. The blurred properties and segmented features within the RHIx4_PhysRep baseline interpolation method HR images are quite significant. The RHIx4_PhysRep SRGAN SR image has better perceptibility, in general, due to its more distinct object shapes and color boundaries. Nevertheless, it contains a considerable amount of pixelation within the second system and along the storm edges. It should also be noted that the RHIx4_PhysRep SRGAN SR image has a shadow artifact along the second system's lower levels and significantly underestimates the reflectivity throughout the low-frequency

regions of the radar scan. That being said, the images exhibit the SRGAN model's ability to rectify the segmentation from the physically representative downsampling method. Altogether, these analyses suggest that the SRGAN model could be an effective tool in conducting SR for actual LR radar scans.

The baseline interpolation methods have significantly higher performances when compared to the SRGAN models for both of the PPIx2 resolution scale dataset experiments, as shown in Table 6.9. The only exceptions to this are the nearest neighbors method evaluations in which the SRGAN model has a comparable performance for the PPIx2_Interp dataset experiment and slightly outperforms the nearest neighbors baseline method for the PPIx2_PhysRep dataset experiment. This analysis suggests that the experimental SRGAN models are not adequate at super-resolving PPI images with a low resolution scale factor. In the PPIx4_Interp resolution scale dataset experiment, the experimental SRGAN model performs better than most of the baseline interpolation methods tested except the Lanczos method with which it had a comparable performance. They both performed similarly on the MSE evaluation metric while the Lanczos baseline method was evaluated as having a higher PSNR and the PPIx4_Interp SRGAN model was evaluated as having a higher SSIM. It is interesting to note that the experimental SRGAN model for the PPIx4_Interp experiment had the highest SSIM overall when compared to its counterpart baseline interpolation methods. The PPIx4_PhysRep SRGAN model, on the other hand, outperformed only the Nearest Neighbors baseline interpolation method. This was not consistent across all evaluation metrics. The nearest neighbors baseline method was evaluated as having a higher SSIM than the PPIx4_PhysRep SRGAN model. Therefore, though this analysis, it can be said that the SRGAN models generally tend to have greater performances when compared to the evaluation results for their respective baseline methods when using the PPI datasets at higher resolution scales. It is hypothesized that a possible reason for this behavior is due to the observation that the lower resolution scaling technique does not significantly affect the overall perceptibility of the radar scan. The HR and LR images are quite comparable – even for the physically representative downsampling LR images – especially the areas of the radar scan that are within the closer ranges. These

analyses suggest that the baseline interpolation methods have an advantage in generating well-performing HR images while under these conditions. Furthermore, although the SRGAN model is evaluated as having an improved performance for the higher resolution scale, PPI dataset experiments, the SRGAN models do not consistently outperform the baseline interpolation methods. The experimental SRGAN model performs higher on the evaluation results for the PPIx4_Interp dataset experiment, whereas the PPIx4_PhysRep SRGAN model only has a higher performance than the nearest neighbors baseline method. This analysis does not agree with the RHI dataset experiment findings which found the SRGAN models to be more suitable for super-resolving the RHI LR images in the higher resolution scale experiments.

The PPI_Interp evaluation results are observed as being significantly higher than their corresponding PPI_PhysRep evaluations. This could be attributed to the observation that the interpolation downsampling method does not affect the overall visual quality to the same extent as the physically representative downsampling method. Downsampling by interpolation results in a LR image that appears slightly blurred but, otherwise, is quite similar to its respective HR image in terms of the characteristic features. In addition, the LR images are created by applying the downsampling method and then using an interpolation kernel to reduce the size of the image. Due to the interpolation method being utilized to generate the resulting LR image's features and overall structure, it stands to reason that the PPI_Interp models would have an advantage when super-resolving the LR images when compared to their respective PPI_PhysRep counterparts. That being said, the RHIx2_PhysRep and PPIx2_PhysRep SRGAN models had higher performances compared to the RHIx2_Interp and PPIx2_Interp SRGAN models relative to their respective baseline interpolation method evaluations. Essentially, the x2_PhysRep SRGAN models were closer in performance to the baseline interpolation methods than the x2_Interp SRGAN models. This is significant as it demonstrates that the SRGAN model can be more suitable for super-resolving physically representative LR images. This can be further extrapolated to suggest that the SRGAN model would be able to perform SR on actual LR radar scans to an improved capacity as well. However, this logic primarily applies to the lower resolution scale dataset. The PPIx4_Interp SRGAN model

had a higher performance relative to its baseline interpolation methods' performances than the PPIx4_PhysRep SRGAN model. This also differs with the analysis from the RHI data experiment which found that the SRGAN model would be a more effective SR technique for physically representative LR images. Although different, the result assessments discussed exhibit the importance for conducting experiments on the SRGAN model with both types of radar scans.

Figure 6.33 shows the sample SR images generated by the experimental PPI SRGAN models and the baseline interpolation methods. When compared to one another, the baseline interpolation methods generate very similar HR images for each of their respective dataset experiments, especially for the two times resolution scale datasets. In the four times resolution scale datasets' sample output HR images, the nearest neighbors SR image appears more pixelated than the other baseline interpolation methods that appear more blurred. It is also noted that each of the SRGAN SR images contain artifacts that notably affect their visual quality. In addition, most all of the SRGAN SR images underestimate the reflectivity throughout the storm especially within the lower-frequency areas around the squall line features, except for the PPIx2_Interp SR image. The PPIx2_Interp SR image has lower values in general which is shown within the primary squall line as an overestimation of the reflectivity represented. A fair number of black dot artifacts are observed to reside within this area of the PPIx2_Interp SR image as well. In general, the pertinent radar information is retained in the SRGAN generated SR image. The PPIx2_PhysRep SR image, on the other hand, is significantly distorted due to the multiple artifacts that are affecting it. Multiple black spot artifacts, pixelation and a substantial blur that alters a majority of the primary squall line feature's object shape all negatively impact the visual quality of the SR image. Black spot artifacts are also observed within the PPIx4_Interp SRGAN model's SR image; however, the overall reflectivity representation and likeness to the HR image in terms of object shapes and high-frequency details is considerably improved in the SRGAN SR image when compared to the baseline interpolation method SR images. The baseline interpolation method SR images for the four times resolution scale PPI dataset either have substantial blurring or pixelation that affects their perceptibility. The SR image from the PPIx4_PhysRep SRGAN displays better defined object

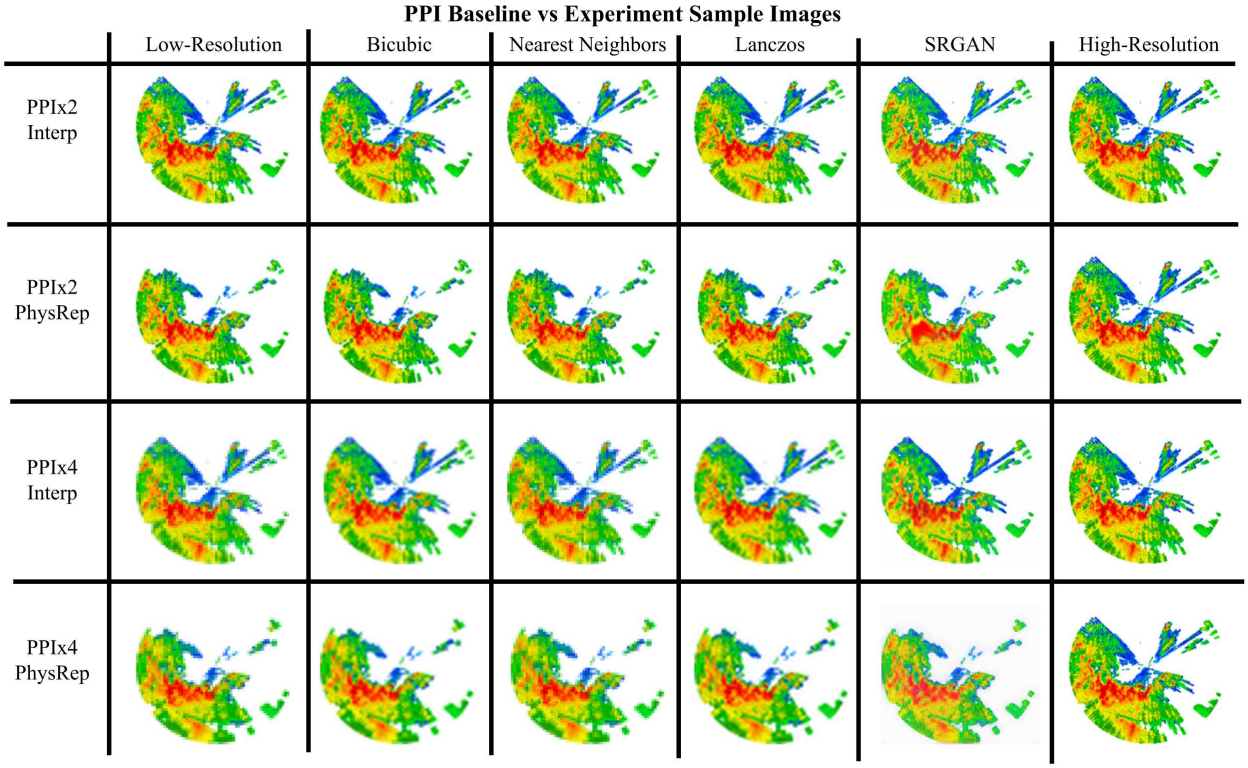

Figure 6.33: PPI Baseline vs Experiment Examples

shapes within the primary squall line than the baseline interpolation methods. Nevertheless, it is still subject to pixelation, overestimation of the reflectivity within the primary squall line, black dot artifacts, and less distinct high-frequency details when compared to the HR image. The baseline interpolation methods' SR images, on the other hand, contain more consistent visual information with the HR image, albeit their visual quality is reduced due to blurring and pixelation. The analysis from these results indicate that the SRGAN model would require further development and testing in order to properly generate physically representative SR images for PPI radar scans.

When comparing the results from the different radar scan type datasets – RHI and PPI – the SRGAN model is determined to be more effective at conducting SR for the RHI scans. Overall, the RHI dataset experiments have notably higher performances on their evaluations than their PPI dataset experiment counterparts for all dataset configurations. It is interesting to note that this analysis applies for both the PSNR and MSE evaluation metrics. The SSIM results, however, primarily favor the PPI dataset experiments. Furthermore, the RHI SRGAN models are able to outperform

the baseline interpolation methods for more of the dataset configurations tested than the PPI SRGAN models. The SRGAN models tested, when compared across radar scan types, have similar performances for the two times resolution scale dataset experiments. Assessment of the four times resolution experimental results, however, shows that the RHI radar scan type provides an advantage in the SRGAN model's ability to conduct super-resolution. The PPIx4_Interp SRGAN model outperforms two out of the three baseline interpolation methods. The PPIx4_PhysRep SRGAN model only outperforms the nearest neighbors method. Meanwhile, both of the RHI four times resolution scale dataset experiments are found to outperform all of the baseline interpolation methods. This difference in performance could be attributed to the nature of the different radar scan types' ranges and resulting resolutions. The PPI scans have a much higher effective resolution when compared to the RHI scans, resulting in finer details that need to be reconstructed during the SR images' generation. The RHI scans range from 0 - 140 km on the x axis and 0 - 20 km on the y axis. Conceptually, for a RHI SR image of size 256x256, each pixel is representing $\approx 0.55\ km$ in the x dimension and $\approx 0.08\ km$ in the y dimension. The PPI scans range from -150 - 150 km on both the x and y axes. Conceptually, this means that the PPI SR image pixels are representing $\approx 1.2km$ in both the x and y dimensions; nearly twice the amount of information is being represented by a single pixel in the x dimension and nearly fifteen times in the y dimension for the PPI scans when compared to the RHI scans. Thus, the RHI SRGAN models can observe more information during training while having fewer high-frequency details to generate in their SR images than the PPI SRGAN models, which would explain this significant difference in performance.

The SR images shown in Figures 6.32-6.33 further support the assertion that the SRGAN model is more effective at conducting super-resolution for the RHI radar scans. By observing the SR images, it is evident that an increased amount of artifacts affect the prominent features of the PPI SRGAN SR images when compared to the RHI SRGAN SR images. These artifacts significantly reduce the visual quality of the SRGAN SR images for the PPI dataset experiments. The RHI SRGAN SR images do not contain many artifacts and the artifacts that do affect the SR images do not alter the characteristics of the primary storm features. In addition, it is difficult to discern whether

or not the PPI SRGAN models mitigate the segmentation affect of the physically representative downsampling model. When comparing the PhysRep dataset LR images across the radar scan types, the segmentation from the downsampling method is quite apparent in the RHI_PhysRep dataset LR images. The segmentation affect is then propagated through the RHI dataset baseline interpolation methods' SR images. Meanwhile, the RHI dataset SRGAN models are observed as more closely reconstructing the characteristic features since their generated SR images adequately represent the storm feature information of the respective HR images. However, this is not as apparent in the PPI_Physrep dataset's LR images. It is more difficult to perceive the affect of the physically representative downsampling method on the PPI LR images. Therefore, it is also more difficult to observe the extent to which the PPI dataset SRGAN models can mitigate the segmentation affect of the physically representative downsampling method. It is plausible that this is also caused by the difference in the radar scan types' ranges and resulting resolutions. This gives the PPI dataset SRGAN models a disadvantage when conducting SR as they are required to reconstruct finer details from the PPI radar scans. Through these results and analyses, it is determined that the SRGAN models developed are generally more compatible with RHI radar scan type datasets than the PPI radar scan type datasets.

6.11 Application

In order to exemplify the strengths of the experimental SRGAN models, Chapter 6.11 presents applications in which the SRGAN model sufficiently super-resolves a subset of the LR image dataset. The images for the subset of LR data were manually constructed from the original LR dataset. The experimental SRGAN model trained on the original LR dataset was evaluated on the LR subset dataset. Training was not re-conducted for the LR subset dataset. Specifically, Chapter 6.11 demonstrates the effectiveness of the SRGAN model in super-resolving the LR dataset for RHI weather radar scans so that the definition of the melting layer within the storm is maintained.

RHI weather radar scans enable meteorological scientists and atmospheric researchers to observe vertical sections of the weather event of interest. From this, vertical profiles of the different

layers that comprise the volumetric target can be distinctly analyzed. One of these layers is called the melting layer. It is defined as the area of the storm in which the temperature is high enough to start melting the ice particulates, which are in a frozen state higher up in the atmosphere, as they fall down to the earth. In a RHI scan, this can be observed as a brightband section of high reflectivity values maintained across a fairly constant altitude. A brightband section within a weather radar scan typically refers to an area of high reflectivity as many of the standard color definitions used across weather radar agencies use warm colors to represent higher reflectivity values. As the ice particulates start to melt, they become surrounded by a layer of water which makes their back scattered energy signature appear to read similarly to large raindrops to the reflectivity product. The altitude of the melting layer is an important variable in determining weather analysis products such as hydrometeor classification. It is also crucial for short-term forecasting applications as it can give insight into the probability of a weather event producing hail as well as the precipitation type at the ground-level. Since this layer is typically higher up in the atmosphere, the melting layer is commonly observed using the higher elevation angles of the radar data and are readily observed within RHI weather radar scans that cover multiple elevation angles within a single scan.

The highest performing RHI SRGAN model is used to demonstrate the effectiveness of the SRGAN model in maintaining the melting layer as opposed to the baseline interpolation methods. Table 6.10 presents the quantitative evaluation results of the RHIx4_PhysRep SRGAN model compared against the baseline interpolation methods for the brightband subset LR dataset. From these, it is evident that the RHIx4_PhysRep SRGAN model outperforms the baseline interpolation methods in both PSNR and MSE. A sufficient explanation for this is that the baseline interpolation methods are inadequate at performing well when transforming images with a significant amount of distortion from the target HR image. The RHIx4_PhysRep LR images contain considerable distortion due to the segmentation affect of the physically representative downsampling method that is more indicative of actual LR weather radar scans. It is also interesting to note that the RHIx4_PhysRep SRGAN model was outperformed by all baseline interpolation methods in the SSIM evaluation metric results. However, these results are not consistent with the visual quality

Table 6.10: Result Assessment: Brightband Application

Dataset	Model	PSNR	MSE	SSIM
RHIx4 PhysRep	[64, 64, 64]	17.63	0.072	0.706
	Bicubic	17.14	0.082	0.731
	Nearest Neighbors	16.77	0.088	0.730
	Lanczos	17.13	0.082	0.725

of the SR images as shown in Figure 6.34 below. Figure 6.34 portrays the SR image output from the RHIx4_PhysRep SRGAN model alongside its corresponding LR, HR, and baseline interpolation methods' outputs. The physically representative downsampling method's segmentation affect is clearly seen in the LR image. The radar rays become wider, blocked segments that are more representative of actual LR weather radar scans. This affect is propagated through all of the baseline interpolation methods causing increased distortion due to segmentation, overestimations in the reflectivity being represented and undefined object shapes and color boundaries as a result of increased blurriness in general. Crucially, the baseline interpolation methods are unable to reconstruct the melting layer after the 20 km range. The segmentation and overestimations combined cause a significant level of distortion in the melting layer leaving it illegible. While the baseline interpolation methods are found to be ineffective, the RHIx4_PhysRep SRGAN model is found to amply perform SR on the physically representative brightband subset dataset. The segmentation affect is nullified in the RHIx4_PhysRep SRGAN model's generated SR image. The object shapes and color boundaries better reflect those found in the ground-truth HR image. Most importantly, the melting layer region is distinctly defined and clearly interpretable. This suggests that the SRGAN model is capable of generating high-quality SR images for the RHI radar scan type at higher resolution scale factors. Furthermore, the SRGAN model is predicted to outperform baseline interpolation methods when super-resolving actual LR weather radar scans.

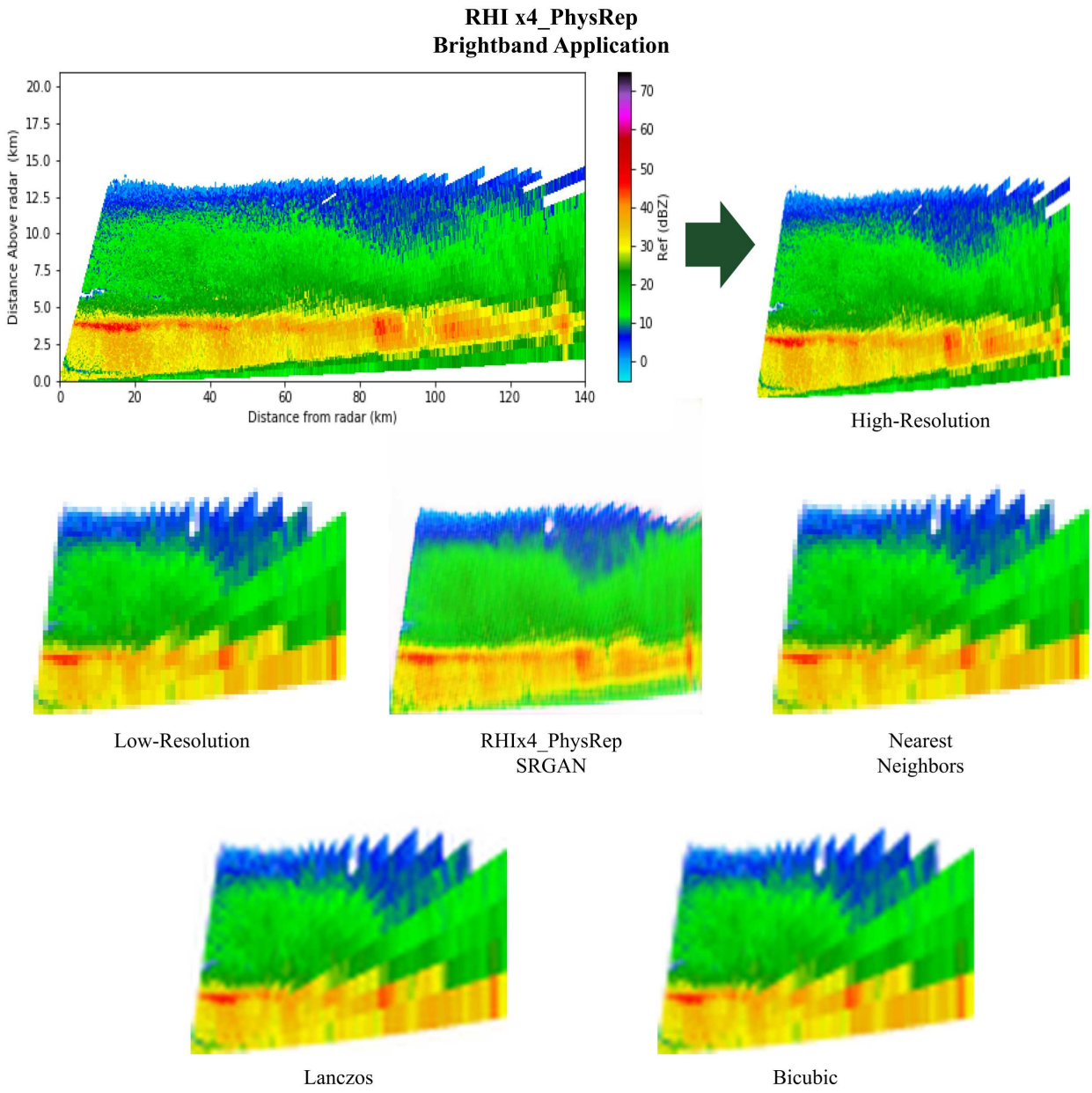

Figure 6.34: Brightband Application SRGAN model vs Baseline Comparison Example

Chapter 7

Summary

Weather radar is crucial in helping ensure preparedness before and during severe weather events for reservoir management, flood control, sewer-stormwater agencies and emergency response operations. Weather radar can collect a wide variety of data containing insightful information about the hydrometeor volume of interest, enabling researchers and scientists to predict precise details about the storm's speed, trajectory, development, precipitation type and intensity. Together, these provide a comprehensive outlook of the storm and how to best prepare for it. Because of this, weather radar scans provide longer lead times on impending weather events which is vital especially when forecasting severe storms or natural disasters such as hurricanes, tornadoes and blizzards. With the importance of weather radar data growing due to the effects of climate change exacerbating the intensity of these severe weather conditions, the need for more data being collected at faster rates while maintaining the quality of higher resolution scans is growing as well.

This thesis' research focuses on the effectiveness of utilizing a SRGAN model in order to generate super-resolved weather radar scans from LR scans. The motivation behind this research is in the efforts that using a DL SR model to super-resolve LR weather radar scans into quality HR scans would enable the weather radars to operate at increased scanning rates. Therefore, the weather radars would conduct LR scans and collect LR data while maintaining the same level of quality in the output data, by super-resolving the LR data, as if HR scans were conducted. If fully implemented, the weather radar scanning paradigm could be significantly enhanced. Weather radars would be able to scan at faster rates, two to four times faster, collecting significantly more data in a shorter amount of time for every weather event while still having high-resolution, quality data outputs available.

In order to test the effectiveness of the SRGAN model in super-resolving weather radar scans, experiments were designed to test different aspects of the SRGAN's generative capabilities. To begin, an optimization process of the SRGAN models' hyperparameters, referred to as HPO, was

carried out in order to fine-tune the SRGAN models and provide them with a well-curated environment in which to train. Three dataset-type experiments were set up and eight different training datasets were built for each combination of these experiments. The first dataset-type experiment tested the SRGAN models' capabilities in super-resolving different radar scan types, specifically RHI and PPI. The second dataset-type experiment tested different resolution scale factors used for super-resolving the input LR images, specifically x2 and x4 resolution scaling. Finally, the third dataset-type experiment tested different methods used during the pre-processing downsampling stage to build the LR datasets. The standard bicubic interpolation kernel was used for creating half of the LR images in order to stay consistent with the rest of the literature. However, the other half of the LR images were produced using a physically representative downsampling method that creates LR images that are more akin to LR radar scans in terms of their characteristic features. With the dataset-type experiments defined, additional experiments were carried out to determine what combination of the SRGAN models' architectural variables had the highest performance in generating SR images. Three architectural variables were investigated, namely: the discriminator filter size, the generator filter size and the number of residual blocks used within the generator neural network. A set of experimental values were tested for each of these experimental variables. With each combination of values for these experimental variables, both the dataset-type and architectural variables, being investigated, a total of 108 individual experimental SRGAN models were trained, evaluated and analyzed through this thesis work.

After training, the experimental SRGAN models' results were compared against one another in order to determine the highest performing models for each of the eight different dataset-type SRGAN models. The ranking of the evaluation results was determined using a combined analysis based on three evaluation metrics: PSNR, MSE and SSIM. In addition, qualitative analyses were provided as well since the visual perceptibility and interpretability of the SR images is of paramount importance. A comprehensive overview of the experimental SRGAN models' evaluation results was provided that analyzed trends and patterns in performance between the experimental architectural variables being investigated. This was to give insight into the nature of utilizing

SRGAN models for weather radar scan super-resolution for different architectural configurations. The highest performing sets of variables for each of the eight experimental, dataset-type SRGAN models were then compared against the baseline interpolation methods. Then, a subset of the testing dataset was created that solely consisted of brightband RHI scans. This was used to exemplify the strengths of the experimental SRGAN models for real-world applications, such as preserving the melting layer in RHI scans for implementations such as hydrometeor classification.

7.1 Conclusion

The qualitative and quantitative analyses provided in Chapter 6 will drive the following conclusions that are made as follows throughout Chapter 7.1. Firstly, the results from this thesis concurs with other works in the literature that state that the accepted evaluation metrics – considered standard throughout the ML and computer vision communities – do not adequately measure the perceptibility of the input images being investigated. This was epitomized through the qualitative analyses made throughout Chapter 6 in which multiple generated SR images that performed higher on the evaluation metrics proved to have a greater extent of distortion and lower perceptibility in the primary, characteristic features, when observed visually. At the same time, there were also SRGAN models that generated low-performing SR images even though their visual representation of the characteristic features were observed to be of a higher quality than other SR images that had higher performances on the evaluation metrics. This is a long-established challenge within the ML and computer vision communities, particularly when developing SR techniques. The need for perceptually relevant evaluation metrics is evident.

Secondly, the original SRGAN architecture designed in [14] needed to be slightly reconfigured in order to perform well in conducting SR on each of the different types of weather radar datasets, within the context of the training process defined. The set of reference values utilized did not consistently have higher performances than the other experimental values tested. This is shown in Table 6.9 as not a single one of the highest performing SRGAN models utilized the complete set of reference values. This demonstrates the importance of this thesis' experimenta-

tion in training individual SRGAN models with different sets of architectural parameter values. In addition, the experimental SRGAN models that were trained on the interpolation-based datasets had consistently higher performances than those trained on the physically representative datasets. A suitable explanation for this behavior is that the original SRGAN architecture was designed with the consideration that bicubic interpolation would be used as the downsampling method in order to create the LR input dataset. With this in mind, it is understandable why the interpolation dataset experiments were generally favored in having higher performances than the physically representative dataset experiments for both the RHI and PPI radar scan types. In addition, the interpolation downsampling method generates a slightly blurred LR image that is very similar in terms of its characteristic features when compared to its corresponding HR image. This gives the interpolation-based SRGAN models an initial advantage as the LR images that they learn from contain more information that is, at the same time, more similar to the HR target images. Thus, the interpolation-based SRGAN models are given a more suitable environment for training in general than the physically representative SRGAN models. These results suggest that the either the original SRGAN model architecture would need to be re-designed and specifically configured for weather radar scans or that the training procedure utilized would need to be improved. This way, the experimental SRGAN models would have higher performances on the physically representative dataset which is more representative of actual LR weather radar scans. However, this conclusion primarily considers the evaluation metric results which, based off of the previous conclusion, is lacking as the evaluation metrics do not consistently reflect the perceptual interpretability of the generated SR images. The visual quality of the SR images should also be taken into account. In addition, it should also be noted that the physically representative SRGAN models – trained on the dataset that better represents the characteristic features of actual LR weather radar scans – had higher performances relative to their baseline interpolation methods than the interpolation-based SRGAN models. This is a significant result as the SRGAN model is able to conduct SR to a higher degree within the context of super-resolving physically representative LR images when compared to the baseline methods. This demonstrates the efficacy of utilizing the SRGAN model for super-

resolving the physically representative LR images that are more similar to weather radar scans in real-world applications.

Thirdly, the SRGAN models tested were able to outperform some of the baseline interpolation methods, specifically for the higher resolution scale factor experiments. The x4 PPI SRGAN models outperformed some of the baseline interpolation methods, although these results were not consistent across all evaluation metrics for both the interpolation and physically representative datasets. The x4 RHI SRGAN models, however, consistently outperformed the baseline interpolation methods for both the interpolation-based and the physically representative datasets when considering the results of all the evaluation metrics. Therefore, the SRGAN model is more effective at conducting SR for RHI radar scan types and for higher resolution scaling factors. This result is significant as it suggests that the SRGAN model is able to generate SR images that perform to a higher degree when the LR input images are significantly more distorted in terms of their structure and characteristic features due to the increased resolution scaling conducted during the downsampling process. From these, it is concluded that the SRGAN model is more suitable for super-resolving RHI LR weather radar scans using higher resolution scaling factors. From the analysis of the results prior, the proposed optimal set of architectural parameters for the different radar scan type datasets would be a DFS, GFS and NRB of [128, 64, 64], respectively, for the RHI dataset and [64, 32, 4] for the PPI dataset. These values were determined taking into consideration the evaluation results of the highest performing SRGAN models versus their respective baseline interpolation methods' performances. For these optimal sets, priority was given to the physically representative SRGAN models' results as they are more representative of the SRGAN's performance in super-resolving actual LR weather radar scans. The overall affects of each of the architectural parameters should be reiterated as well. It was found that the DFS parameter does not significantly affect the evaluation results of the experimental SRGAN models for neither the PPI nor the RHI datasets. Conversely, the GFS parameter has a substantial impact on the performances of all of the experimental SRGAN models, especially when set to quantities greater than 64 after which the performance significantly degrades. Meanwhile, the NRB paramter had a greater affect

on the performance of the RHI experimental SRGAN models when compared to the PPI experimental SRGAN models whose results were relatively consistent even when changing the NRB parameter values.

Finally, the SRGAN models' SR images did not exhibit the segmentation affect caused by the physically representative downsampling method and, instead, more closely reconstructed the natural shape of the ground-truth, HR images. This strongly demonstrates the SRGAN model's efficacy for super-resolving actual weather radar scans as it can outperform the baseline methods, in terms of the visual quality, by generating a more natural, physically representative SR image that mitigates the segmentation affect found in actual LR weather radar scans. This suggests that the SRGAN model would be a more efficacious SR technique when applied to real-world implementations. Although further experimentation would be necessary for super-resolving PPI weather radar scans in particular, the SRGAN model has been shown to be effective at super-resolving RHI weather radar scans, especially at higher resolution scaling factors.

7.2 Future Work

A discussion of the aspects of this thesis research that could be improved upon is provided in the comments following. The primary improvements that could be made to this thesis work would be to improve the architectural configuration of the SRGAN model being investigated for use within the weather radar regime, train on a larger dataset with larger batch sizes for each epoch and to test the SRGAN's generative capabilities using the base radar data itself as the inputs instead of plotted images of the weather radar scans. It would also be interesting to develop a segmented training SRGAN in which the SRGAN is trained on smaller sections of the input images and generating SR images based on those. Then, during evaluation, the SRGAN is run multiple times to generate every section of the input image until a full HR image is constructed. This could improve the SRGAN's learning development by training on smaller sections and potentially increase the amount of SR conducted allowing for higher resolution regimes to be reached. In addition, validation of the research and results conducted through this thesis study would be quite useful, particularly

studying with actual LR weather radar scans. Scanning strategies could be developed for a single radar that alternates between HR and LR scans that could function as the input data pairs for training the SRGAN model on actual weather radar scan data. In addition, scanning strategies could be coordinated between multiple radars – or radars that have dual-frequency capabilities – that are alternating between HR and LR scan sequences asynchronously so that HR and LR pairs could be collected for each scan for each radar. The paramount research endeavor to pursue would be to demonstrate the validity of utilizing a SRGAN model for super-resolving actual LR weather radar scans by implementing its generative capabilities in a real-time, operational weather radar system. This would allow the weather radar to scan at a faster rate and double or quadruple the amount of weather radar data collected at a lower resolution while maintaining the data quality of the higher resolution scans by utilizing the generative SRGAN model to generate SR weather radar scans in real-time.

Bibliography

[1] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[2] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine Vision and Applications*, vol. 25, p. 1423–1468, 2014.

[3] R. Y. Tsai and T. S. Huang, "Multiframe image restoration and registration," in *Advances in Computer Vision and Image Processing*, vol. 1, pp. 317–339, 1984.

[4] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical Models and Image Processing*, vol. 53, no. 3, pp. 231–239, 1991.

[5] A. Tekalp, M. Ozkan, and M. Sezan, "High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 169–172 vol.3, 1992.

[6] M.-C. Chiang and T. Boult, "Efficient image warping and super-resolution," in *Proceedings Third IEEE Workshop on Applications of Computer Vision. WACV'96*, pp. 56–61, 1996.

[7] P. Cheeseman, B. Kanefsky, R. Kraft, J. Stutz, and R. Hanson, "Super-resolved surface reconstruction from multiple images," in *Maximum Entropy and Bayesian Methods: Santa Barbara, California, U.S.A., 1993* (G. R. Heidbreder, ed.), (Dordrecht), pp. 293–308, Springer Netherlands, 1996.

[8] R. Hardie, K. Barnard, and E. Armstrong, "Joint map registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1621–1633, 1997.

[9] M. Elad and A. Feuer, "Superresolution restoration of an image sequence: adaptive filtering approach," *IEEE Transactions on Image Processing*, vol. 8, no. 3, pp. 387–395, 1999.

[10] S. Son and K. M. Lee, *Image Super-Resolution*, pp. 646–650. Cham: Springer International Publishing, 2021.

[11] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision – ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), (Cham), pp. 184–199, Springer International Publishing, 2014.

[12] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[13] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1646–1654, 2016.

[14] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2017.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[17] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1132–1140, 2017.

[18] B. Wu, H. Duan, Z. Liu, and G. Sun, "Srpgan: Perceptual generative adversarial network for single image super resolution," 2017.

[19] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," 2018.

[20] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Computer Vision – ECCV 2018 Workshops* (L. Leal-Taixé and S. Roth, eds.), (Cham), pp. 63–79, Springer International Publishing, 2019.

[21] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," 2018.

[22] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia*, vol. 21, pp. 3106–3121, dec 2019.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[25] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Processing*, vol. 128, pp. 389–408, 2016.

[26] M. Alshaye, F. Alawwad, and I. Elshafiey, "Hurricane tracking using multi-gnss-r and deep learning," in *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*, pp. 1–4, 2020.

[27] J. M. Cuomo, "Machine learning models applied to storm nowcasting," Master's thesis, Colorado State University, 2020.

[28] L. Chen, Y. Cao, L. Ma, and J. Zhang, "A deep learning-based methodology for precipitation nowcasting with radar," *Earth and Space Science*, vol. 7, no. 2, p. e2019EA000812, 2020. e2019EA000812 10.1029/2019EA000812.

[29] S. R. Gooch, *Transfer Learning with Weather Radar*. PhD thesis, Colorado State University, 2020.

[30] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community," *Journal of Applied Remote Sensing*, vol. 11, p. 1, sep 2017.

[31] X. Zhang, J. He, Q. Zeng, and Z. Shi, "Weather radar echo super-resolution reconstruction based on nonlocal self-similarity sparse representation," *Atmosphere*, vol. 10, no. 5, pp. 254–, 2019.

[32] H. Yuan, Q. Zeng, and J. He, "Adaptive sparse domain selection for weather radar superresolution," *Journal of mathematics (Hidawi)*, vol. 2021, pp. 1–11, 2021.

[33] Q. Yu, M. Zhu, Q. Zeng, H. Wang, Q. Chen, X. Fu, and Z. Qing, "Weather radar super-resolution reconstruction based on residual attention back-projection network," *Remote Sensing*, vol. 15, no. 8, 2023.

[34] A. Geiss and J. C. Hardin, "Radar super resolution using a deep convolutional neural network," *Journal of Atmospheric and Oceanic Technology*, vol. 37, no. 12, pp. 2197 – 2207, 2020.

[35] M. Rifat Arefin, V. Michalski, P.-L. St-Charles, A. Kalaitzis, S. Kim, S. E. Kahou, and Y. Bengio, "Multi-image super-resolution for remote sensing using deep recurrent networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 816–825, 2020.

[36] M. Sit, B.-C. Seo, and I. Demir, "Tempnet – temporal super resolution of radar rainfall products with residual cnns," 2022.

[37] H. Chen, X. Zhang, Y. Liu, and Q. Zeng, "Generative adversarial networks capabilities for super-resolution reconstruction of weather radar echo images," *Atmosphere*, 2019.

[38] R. J. Doviak and D. S. Zrnić, *Doppler Radar and Weather Observations (Second Edition)*. San Diego: Academic Press, second edition ed., 1993.

[39] P. Meischner, *Weather Radar Principles and Advanced Applications*. Springer Berlin, Heidelberg, 2004.

[40] V. N. Bringi and V. Chandrasekar, *Polarimetric Doppler Weather Radar: Principles and Applications*. Cambridge University Press, 2001.

[41] S. M. Torres and C. D. Curtis, "5b.10 initial implementation of super-resolution data on the nexrad network," in *23rd Int. Conf. on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, 2007.

[42] M. Yeary, B. L. Cheong, J. M. Kurdzo, T.-y. Yu, and R. Palmer, "A brief overview of weather radar technologies and instrumentation," *IEEE Instrumentation & Measurement Magazine*, vol. 17, no. 5, pp. 10–15, 2014.

[43] S. M. Torres and C. D. Curtis, "The impact of range-oversampling processing on tornado velocity signatures obtained from wsr-88d superresolution data," *Journal of Atmospheric and Oceanic Technology*, vol. 32, no. 9, pp. 1581 – 1592, 2015.

[44] N. Takano and G. Alaghband, "Srgan: Training dataset matters," 2019.

[45] I. Arias, V. Chandrasekar, and S. S. Joshil, "Cross-validation of csu-chivo radar and gpm during relampago," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 7586–7589, 2019.

[46] S. Nesbitt, "Proyecto relampago-cacti argentina 2018-9," 2019.

[47] S. Nesbitt, P. Salio, E. Ávila, P. Bitzer, L. Carey, V. Chandrasekar, W. Deierling, F. Dominguez, M. Dillon, C. Garcia, D. Gochis, S. Goodman, D. Hence, K. Kosiba,

M. Kumjian, T. Lang, L. Luna, J. Marquis, R. Marshall, L. A. McMurdie, E. de Lima Nascimento, K. L. Rasmussen, R. Roberts, A. K. Rowe, J. J. Ruiz, E. F. S. Sabbas, A. C. Saulo, R. S. Schumacher, Y. G. Skabar, L. A. T. Machado, R. J. Trapp, A. C. Varble, J. Wilson, J. Wurman, E. J. Zipser, I. Arias, H. Bechis, and M. A. Grover, "A storm safari in subtropical south america: Proyecto relampago," in *Bulletin of the American Meteorological Society*, vol. 102, pp. E1621–E1644, 2021.

[48] S. W. Nesbitt, P. Salio, R. J. Trapp, R. D. Roberts, A. C. Varble, F. Dominguez, L. A. T. Machado, and C. Saulo, "Understanding processes and improving predictions of hydrometeorological extremes in subtropical south america: Proyecto relampago-cacti," 2018.

[49] S. K. Biswas, V. Chandrasekar, S. Sahoo, and A. K. Lakshmi, "Study of a convective event during the relampago field experiment using spectral polarimetry," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6534–6537, 2022.

[50] H. Chen, V. Chandrasekar, and R. Bechini, "An improved dual-polarization radar rainfall algorithm (drops2.0): Application in nasa ifloods field campaign," *Journal of Hydrometeorology*, vol. 18, no. 4, pp. 917 – 937, 2017.

[51] J. J. Helmus and S. M. Collis, "The python arm radar toolkit (py-art), a library for working with weather radar data in the python programming language," *Journal of Open Research Software*, Jul 2016.

[52] M. Bevilacqua, A. Roumy, C. M. Guillemot, and M.-L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *British Machine Vision Conference*, 2012.

[53] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces* (J.-D. Boissonnat, P. Chenin, A. Cohen, C. Gout, T. Lyche, M.-L. Mazure, and L. Schumaker, eds.), (Berlin, Heidelberg), pp. 711–730, Springer Berlin Heidelberg, 2012.

[54] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5197–5206, 2015.

[55] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1122–1131, 2017.

[56] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," in *Multimedia Tools and Applications*, vol. 76, pp. 21811–21838, 2017.

[57] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are gans created equal? a large-scale study," 2018.

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[59] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[60] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.