THESIS

SMOKE+: A VIDEO DATASET FOR AUTOMATED FINE-GRAINED ASSESSMENT OF SMOKE OPACITY

Submitted by

Ethan Seefried

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2024

Master's Committee:

Advisor: Nathaniel Blanchard

Sarath Sreedharan
Jacob Roberts

ABSTRACT

SMOKE+: A VIDEO DATASET FOR AUTOMATED FINE-GRAINED ASSESSMENT OF SMOKE OPACITY

Computer vision has traditionally faced difficulties when applied to amorphous objects like smoke, owing to their ever-changing shape, texture, and dependence on background conditions. While recent advancements have enabled simple tasks such as smoke detection and basic classification (black or white), quantitative opacity estimation in line with the assessments made by certified professionals remains unexplored. To address this gap, I introduce the SMOKE+ dataset, which features opacity labels verified by three certified experts. My dataset encompasses five distinct testing days, two data collection sites in different regions, and a total of 13,632 labeled clips. Leveraging this data, we develop a state-of-the-art smoke opacity estimation method that employs a small number of Residual 3D blocks for efficient opacity estimation. Additionally I explore the use of MAMBA blocks in a video based architecture, exploiting their ability to handle spatial and temporal data in a linear fashion. Techniques developed during the SMOKE+ dataset creation were then refined and applied to a new dataset titled CSU101, designed for educational use in Computer Vision. In the future I intend to expand further into synthetic data, incorporating techniques into Unreal Engine or Unity to add accurate opacity labels.

ACKNOWLEDGEMENTS

I would like to thank Dr. Blanchard for bring me into his lab as an undergraduate, and helping me grow in my studies, leading to this thesis, without him I never would have ventured into graduate school. I would also like to especially thank Changsoo Jung, for his mentorship and contributions to not only my other papers, but this thesis. Without the support of professors Sarath Sreedharan, Jacob Roberts, and Nikhil Krishnawamy, I would never have made it this far. To everyone in the Vision lab who has been my friend or worked with me, I thank you. Finally, I would like to thank my friends and family for their love and support over the years.

DEDICATION

I would like to dedicate this thesis to all students who have taken a non-traditional path in school.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Visually determining the opacity of smoke presents a challenge with real-world implications, where accurate measurements inform and guide management practices in industrial settings. Since the Clean Air Act was enacted in 1970, the United States Environmental Protection Agency (EPA) has regulated the opacity of emissions from both stationary and mobile sources, including automobile tailpipes and industrial smokestacks.

While certified human observers are typically deployed and various sensors have been trialed for smoke opacity measurement, the use of computer vision for this purpose is novel. Building upon the need for enhanced measurement techniques, deep learning models offer a promising solution. Despite recent advancements in smoke detection through deep learning [1], the task of accurately predicting smoke opacity remains inadequately addressed. Accurate utilization of computer vision techniques for this task requires large amounts of data, which are often limited or impossible to obtain. To fill this research void, I introduce an innovative, soon to be public video dataset titled: Systematic Measurement Of Smoke Evaluations with Synthetic Data (SMOKE+). SMOKE+ has been meticulously annotated with independently verified ground-truth opacity readings. SMOKE+ presents a unique opportunity to apply advanced computer vision technologies, which will be explored in this thesis to enhance the accuracy and efficiency of smoke detection and opacity measurements. This dataset will be pivotal for developing and benchmarking deep learning models specifically designed for the precise prediction of smoke opacity, representing a significant step forward in opacity measurement technologies.

1.1 Challenges with Smoke

Detecting and characterizing smoke involves navigating a unique set of challenges due to its inherently amorphous nature. Unlike solid objects, smoke lacks a constant shape or structure, influenced heavily by its physical properties and environmental factors, leading to high variability

in form and appearance. This transience and rapid change in shape make consistent tracking and identification difficult for both human annotators, and Computer Vision models. Furthermore, the color and opacity of smoke can vary significantly, from light (white) to dark (black), depending on the combustion process, and are also affected by environmental conditions like sunlight, adding another layer of complexity. The task is again further complicated by the visual similarity between smoke and clouds, which share comparable textures. A successful algorithm will need to be robust to the presence of clouds, and other environmental factors.

1.2 Challenges of deep learning with smoke

Traditional methods, such as Convolutional Neural Networks (CNNs), when looking at images and videos, tend to pick out the general shape and structure of an object in the image. This proves useful when working objects definable by shape and/or texture, such as vehicles, buildings, and trees [2–5]. However, smoke is transient, constantly morphing, and its opacity is defined by the penetration of light, rather than texture — as its shape and texture are constantly changing from frame to frame. Pretrained CNNs are tuned to recognize objects in traditional datasets (e.g., ImageNet), and are specifically designed to filter out opacity features in layers prior to the latent embedding (after the penultimate layer, before classification). Traditional computer vision techniques for predicting opacity are dependent on prior information about the smoke's background and are not robust to temporal changes to the background such as shifting clouds. Thus, opacity estimation required the formulation of a novel solution centered around smoke's physical characteristics as manifested in video.

To address the challenges of visually characterizing smoke, in particular the classification of smoke opacity, we propose a custom model dubbed Opacity Measurement Engine with Graded Accuracy (OMEGA) that utilizes both the temporal and spatial characteristics of smoke. Our OMEGA model selects frames and feeds into multiple Residual 3D blocks to capture useful features of smoke from both the spatial and temporal dimensions. Furthermore, OMEGA introduces

a special weighted focal loss function that performs on the statistically skewed population of the SMOKE+ dataset.

Table 1.1: Comparison of the SMOKE+ dataset with similar existing datasets. The table presents the number of labeled clips, total frames, and average frames per clip. The 'Ratio of smoke frames' column indicates the percentage of frames in which smoke is present. Additionally, the presence of opacity values signifies whether opacity labels are provided for each clip in the respective dataset.

Dataset	Clips	Frames	Avg Frames per clip	Smoke Frames Ratio	Temporal Data	Opacity Labels
Xu *et al*. [6]	–	5,700	–	49%	✗	✗
Xu *et al*. [7]	–	3,578	–	100%	✗	✗
Xu *et al*. [8]	–	10,000	–	50%	✗	✗
Ba *et al*. [9]*†	–	6,225	–	16%	✗	✗
Lin *et al*. [10]*	–	16,647	–	29%	✗	✗
Yuan *et al*. [11]*	–	24,217	–	24%	✗	✗
Bugaric *et al*. [12]	10	213,909	21,391	100%	✓	✗
Ko *et al*. [13]	16	43,090	1,514	37%	✓	✗
Dimitropoulos *et al*. [14]	22	17,722	806	56%	✓	✗
Toreyin *et al*. [15]	21	18,031	820	98%	✓	✗
Filonenko *et al*. [16]*	396	100,968	255	61%	✓	✗
Hsu *et al*. [17]	12,567	452,412	36	41%	✓	✗
SMOKE+	**13,632**	**545,260**	**40**	**98%**	✓	✓

1.3 Main Contributions

- I have compiled a dataset containing 13,622 video clips of black and white smoke releases, and I have provided the true opacity labels of the clips in the dataset. The dataset contains a diversity of environments, weather conditions, and six separate view angles per smoke release.

- As far as I am aware this is the first publicly available dataset for the purpose of evaluating smoke opacity. Previous works have focused on only smoke detection.

- I incorporate a large amount of synthetic data into the dataset, allowing for pre-training on smoke detection, before opacity predictions.
- I present a novel architecture utilizing MAMBA blocks designed to exploit both the spatial and temporal features inherently present in smoke.
- A stereo-vision board was recorded at the beginning of each day making it possible to 3D model smoke using this dataset.

Chapter 2

Literature Review

2.1 Real-World Smoke

2.1.1 Traditional Techniques for Opacity Prediction

In the mid-1950s, the US government funded the Public Health Service (PHS) to research emissions from sources like industrial smokestacks and vehicles, with studies by Stahman *et al*. [18] and [19] noting public concerns about diesel engine smoke opacity. The introduction of regulations and visual smoke sensors led to the creation of tailpipe meters by the PHS in the late 1960s, yet concerns persisted regarding their design variations, as highlighted by Bascom *et al*. [20].

The US federal government recognizes human observation of smoke opacity, detailed in CFR Title 40 Part 60, Appendix A, Method 9, which includes requirements for observer positioning, prediction recording, and certification, which is similar to the usage of the 19th century Ringelmann chart [21]. Heinsohn et al. (1992) found that Method 9 for predicting smoke opacity was technically sound through the evaluation of trained and untrained human observers on various plumes.

An alternative to Method 9 is referred to as Alternative Method 082 (ALT-082) [22], which involves using conventional digital cameras and special software to analyze images of emissions to determine their opacities. The Digital Opacity Compliance System (DOCS) [23–25] and similar software, like those discussed in ALT-082, necessitate the manual selection of regions of interest (ROIs) by a human operator in the original images.

2.1.2 Deep Learning for Smoke Detection

Chaturvedi *et al*. [1] surveyed over 100 papers focused on smoke classification, segmentation, and bounding-box prediction, revealing that 58% utilized video-based datasets and 42% used

image-based datasets. Yin *et al*. [26] presented deep normalization and CNN architecture for smoke detection in still images, where normalized convolutional layers replaced simple convolutional layers to enhance the training process. Hu *et al*. [27] utilized spatial-temporal CNN features for real-time smoke detection. More recently, a hybrid RCNN and 3D-CNN-based smoke detection technique was proposed to recognize smoke in video frames [28].

2.1.3 Deep learning for Opacity Prediction

Pedrayes *et al*. [29] utilized deep learning for predicting opacity of fugitive emissions in industrial settings, employing semantic segmentation to identify emissions in images and the Sky and Building Percentiles in the Blue channel (SBPB) technique to measure opacity. In contrast to [29], my dataset is not labeled for semantic segmentation, and instead, I only label frames/videos with ground truth opacity readings. My dataset can be used to train deep learning models to perform smoke detection in addition to classification of smoke opacity.

2.2 Synthetic Data Generation

2.2.1 Synthetic Smoke

For the task of smoke detection, synthetic data has proven to be a useful resource to improve model performance [30–33]. Various methods mentioned above are applied to generate smoke. For example, [32] used two GANs to produce synthetic smoke images, and they found that images from the higher quality GAN resulted in better smoke detection. Similarly, [34] developed a pipeline to generate synthetic smoke images that allowed for adjustable parameters to yield desired smoke components. Several previous works used Blender for manually generating synthetic smoke data [30, 31, 33]. However, generating a variety of quality smoke images can be difficult and some of the input can be automated [31].

2.2.2 Quality of Synthetic Data

Determining measures of quality in synthetic data is important for understanding its impact on model performance, and much work has been done to create such metrics [35–38]. The Fréchet Inception Distance [39] was utilized in [40] to determine the quality of synthetic data generated by a GAN. In [41], Peak Signal-to-Noise ratio (PSNR) and Structural Similarity Index Measure (SSIM) [42] are introduced into the loss function of a GAN with the aim of reducing noise and thus improving quality. The impact of synthetic data quality varies by task and field. For example, in a review of synthetic data, [43] found that studies on photorealistic synthetic data presented different results, and that the impact depended on the task. Previous work has found that object detection improved with photorealistic synthetic data [44, 45]. Still, synthetic data created using domain randomization yields better results than using only real data [45]. It is important to consider that photorealistic synthetic data has a higher computational cost to produce, and unrealistic data does still show improvements in model performance [43]. The trade-off between computational cost and model improvement is an unanswered question which will likely vary by task. Here, I dive deeper into this problem for the task of smoke detection in industrial settings.

Chapter 3

Dataset

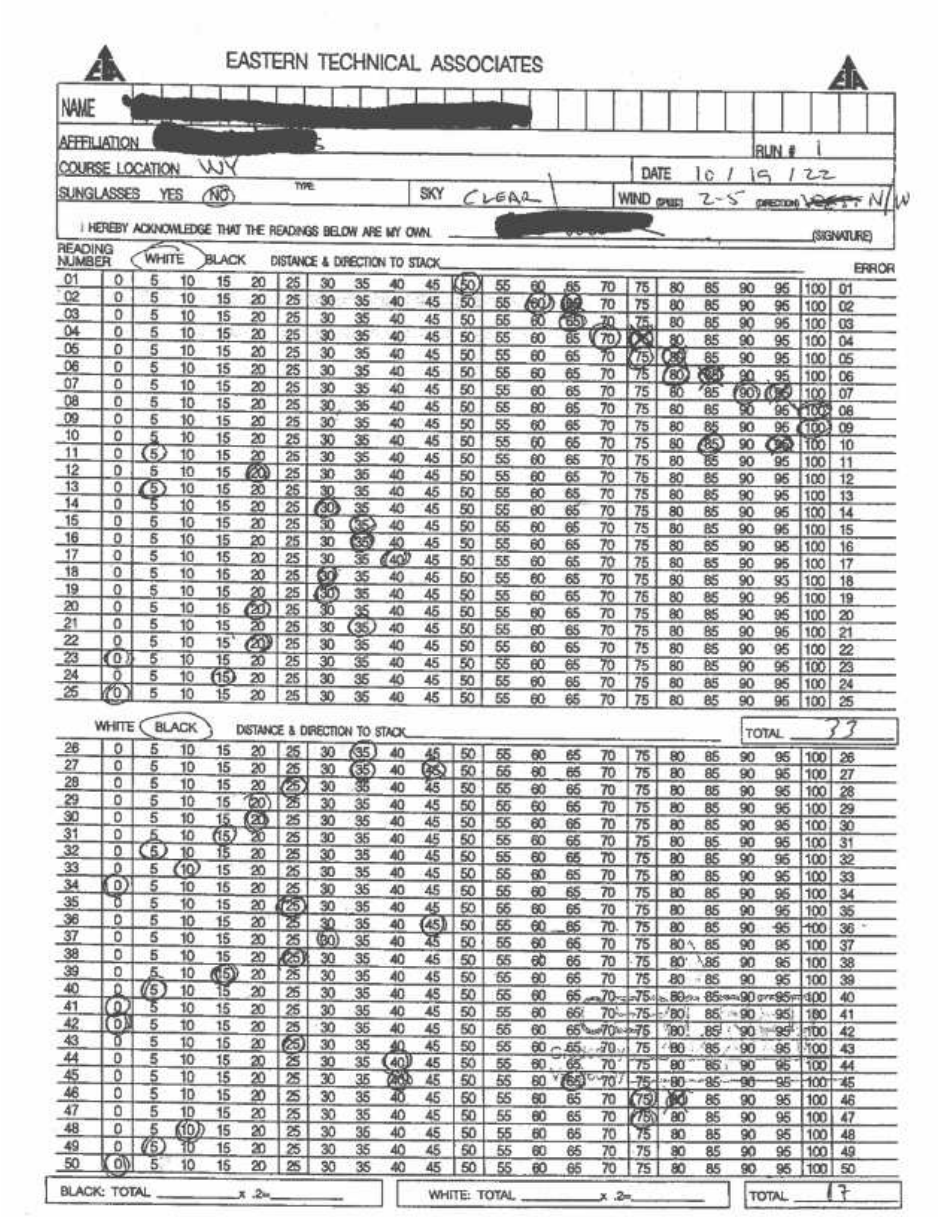
This chapter provides a comprehensive overview of the dataset utilized in this study. I begin by detailing the methods employed for data collection and the procedures followed to ensure the accuracy and relevance of the data. The annotations accompanying the dataset are described, highlighting how they facilitate various machine learning tasks. Additionally, I present an analysis of the dataset's statistical properties, offering insights into its composition and diversity. A significant portion of the dataset comprises real-world data, complemented by a smaller segment of synthetic smoke data, specifically designed to enhance the robustness of the training process.

3.1 Data Collection

The dataset was amassed over five days of testing at two distinct locations: the first being in Cheyenne, Wyoming, during an expert re-certification event, and the second, METEC, Fort Collins, CO, hosting a private session. In total, the collection of real world data resulted in 716 GB of video data, equivalent to 19.25 hours of footage, each recorded at 1080p resolution and 24 frames per second (FPS). Table 1.1 displays the statistics of the dataset in comparison with previous works.

Table 3.1: Descriptive statistics of individual sub-datasets used in the study. The table highlights the number of runs, total clips, and specific weather conditions encountered on different days. Data was captured across two distinct filming locations — Cheyenne, WY, and Fort Collins, CO, providing a diverse range of environmental conditions for the analysis.

Dataset	Total Clips	Hours of Film	Cloudy	Overcast	Rainy	Clear
Wyoming D1	1553	4.25	✗	✗	✗	✓
Wyoming D2	1548	3.50	✗	✗	✗	✓
METEC D1	4058	6.25	✓	✗	✗	✗
METEC D2	4488	3.50	✓	✓	✓	✗
METEC D3	1985	1.75	✗	✓	✗	✗



EASTERN TECHNICAL ASSOCIATES

NAME [redacted]

AFFILIATION [redacted]

COURSE LOCATION WY

RUN # 1

DATE 10 / 19 / 22

SUNGLASSES YES (NO) TYPE:

SKY CLEAR

WIND (SPEED) 2-5 (DIRECTION) ~~VARY~~ N/W

I HEREBY ACKNOWLEDGE THAT THE READINGS BELOW ARE MY OWN. [redacted] (SIGNATURE)

READING NUMBER (WHITE) BLACK DISTANCE & DIRECTION TO STACK

READING NUMBER	WHITE	BLACK	ERROR
01	0	50	01
02	0	60	02
03	0	65	03
04	0	70	04
05	0	75	05
06	0	80	06
07	0	90	07
08	0	100	08
09	0	100	09
10	0	95	10
11	5	100	11
12	20	100	12
13	5	100	13
14	30	100	14
15	30	100	15
16	35	100	16
17	40	100	17
18	30	100	18
19	30	100	19
20	20	100	20
21	30	100	21
22	20	100	22
23	0	100	23
24	15	100	24
25	0	100	25

WHITE (BLACK) DISTANCE & DIRECTION TO STACK

TOTAL 33

READING NUMBER	WHITE	BLACK	ERROR
26	0	35	26
27	0	35	27
28	0	25	28
29	0	20	29
30	0	20	30
31	0	15	31
32	0	5	32
33	0	10	33
34	0	0	34
35	0	25	35
36	0	45	36
37	0	30	37
38	0	25	38
39	0	15	39
40	0	5	40
41	0	0	41
42	0	0	42
43	0	25	43
44	0	40	44
45	0	40	45
46	0	40	46
47	0	30	47
48	0	10	48
49	0	5	49
50	0	0	50

BLACK: TOTAL _____ x .2= _____

WHITE: TOTAL _____ x .2= _____

TOTAL 17

Figure 3.1: An example score-sheet used by a human expert during the opacity measurement tests in Wyoming. This score-sheet illustrates the point system for scoring accuracy, showing deductions for specific percentage deviations from the target values. This format was consistently used to evaluate expert performance and train models to replicate human assessment accuracy.

3.2 Ground Truth Labels

Ground truth labels for smoke opacity were determined using an apparatus with a light source on one side of the smokestack and a light receiver directly opposite. The receiver measured the percentage of light that passed through the smoke, thereby quantifying opacity. For example, an

opacity reading of 50% was recorded if half the light penetrated the smoke. Figure 3.3 visualizes the setup of light passing through the smokestack, before being received. These measurements were systematically rounded to the nearest 5% to align with industry standards and enhance the reliability of human assessments.

Figure 3.2: A demonstration of how to properly read opacity's from a smokestack. A common misconception is to read the opacity from the cloud, however, it's imperative that it is read at the top of the smoke stack, immediately after emission.

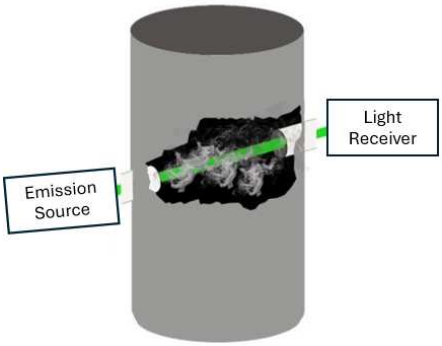


Figure 3.3: An illustration of ground label collection using a light emission source: Light is emitted through smoke, with the amount received by a light receiver determining the opacity. If 50% of the emitted light is received, the smoke is labeled as 50% opaque.

3.3 Human Labels

To supplement our dataset with robust ground truth labels, we enlisted 40 experts for evaluations on the initial day of data collection in Wyoming. These experts, on site to achieve their certification in environmental monitoring, evaluated smoke opacity using a rigorous scoring system. Missing the target opacity by 5% led to a deduction of one point, by 10% two points, and by 15% three points. A deviation of 20% or more in any single measurement, or accruing more than 37 points in either half of the test, resulted in automatic failure. These stringent criteria ensure that our dataset reflects high standards of accuracy. The evaluations not only benchmark against established regulatory standards but also help in training our model to emulate nuanced human expert decision-making in real-world conditions. Figure 3.1 provides an example of an expert test setup, and Figure 3.2 reiterates the crucial technique for accurate smoke opacity measurement at the emission point, reinforcing the necessity of precise measurement locations to avoid common observational errors.

3.4 Hardware Configuration

Data capture was conducted using six GoPro Hero Black cameras in two distinct arrangements to comprehensively document the smoke plume dynamics from multiple perspectives. Initially, the cameras were positioned equidistantly around the smoke source to capture six divergent viewpoints. This arrangement was later modified, grouping the cameras into pairs to form a triangular configuration around the emission point. This setup was particularly aimed at supporting stereovision applications and enhancing the generation of 3D smoke dispersion models. Figure 3.4 illustrates these configurations, showcasing the systematic approach to data acquisition. Synchronization and control of the six GoPros were achieved through a GoPro iPhone app, which also facilitated video tagging for efficient data management.

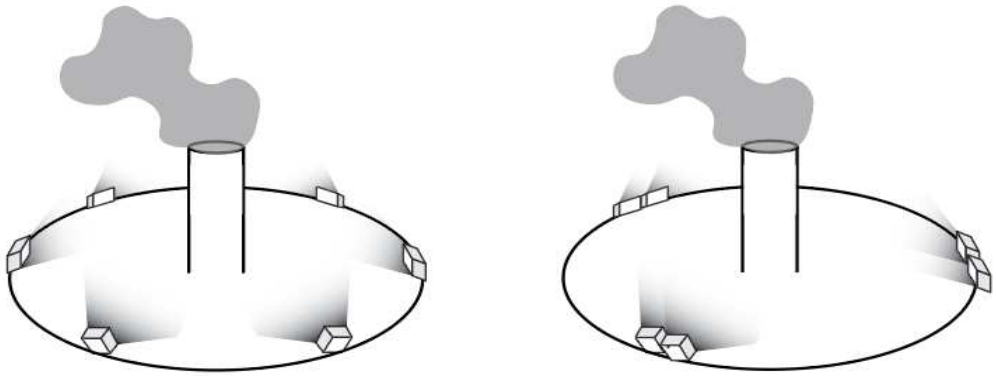


Figure 3.4: Data collection setups: Left - Primary configuration with six GoPros positioned in a circle for capturing smoke from multiple angles and backdrops, enhancing dataset diversity. Right - Setup for 3D smoke modeling using paired cameras and a stereo-vision system for accurate spatial localization and detailed 3D rendering

3.5 Stereo-Vision Board

Utilizing the camera setups described in Chapter 3.4, I initially recorded a stereo-vision board at the start of each day's data collection, immediately after setting up the cameras. In computer vision, a stereo-vision board is crucial for calibrating stereo camera systems. It enables the cameras—typically positioned opposite each other—to accurately determine the depth and distance of objects based on the known dimensions of the board's squares. With known locations of both the smokestack, and adjacent cameras, I can utilize techniques such as NERF, to 3D model the smoke. Figure 3.4 illustrates the method used to record the stereo-vision board.

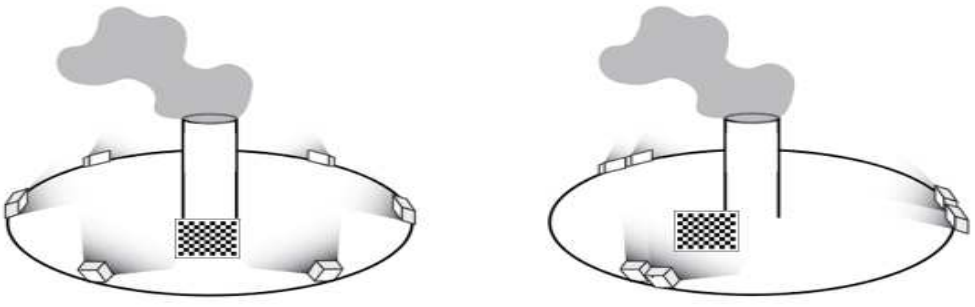


Figure 3.5: Stereo-Vision Board Setup: On the left, the primary configuration with the cameras spaced apart, the stereo board centrally positioned between them near the smoke stack. On the right, the cameras are grouped together, positioning the stereo-vision board directly in front of them. This arrangement was primarily used for 3D modeling.

3.6 Data Processing

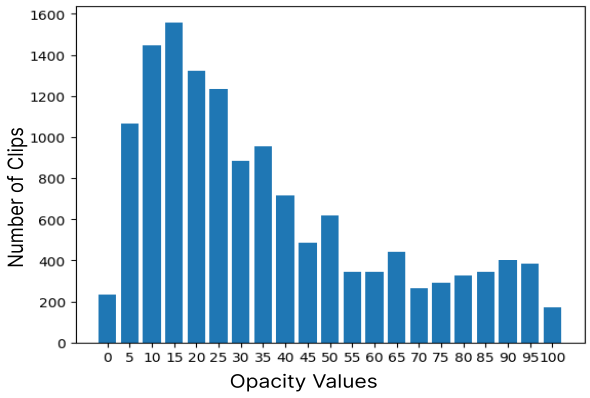
During each testing day, we conducted multiple 'runs.' A 'run' involves a certified expert measuring the opacity of smoke emitted from a small test smoke stack at 25 distinct moments. At each of the 25 points, the instructor would signal a fixed opacity level by announcing 'mark,' prompting the operators to note their opacity predictions. Simultaneously, this verbal cue triggered the addition of a metadata tag to the recordings on all six cameras, enabling precise later annotation of the videos with ground truth opacity values. To facilitate dataset usability, I segmented these marked instances into separate video clips, each comprising 40 frames at 24 FPS, and annotated them with ground truth opacity values, rounded to the nearest 5% increment.

3.7 Dataset Analysis

3.7.1 Dataset Statistics

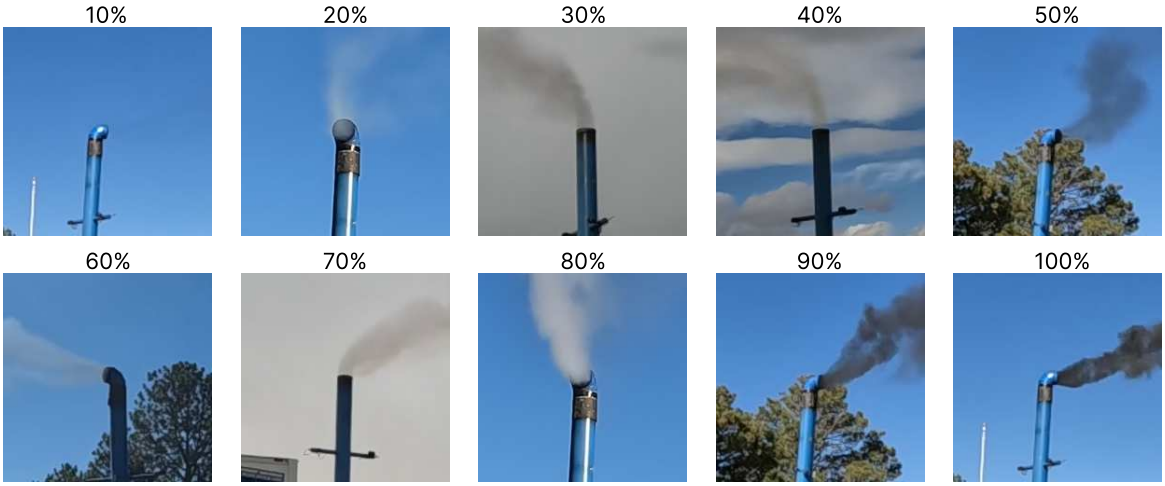
The dataset consists of 13,839 labeled clips, with an approximately equal distribution between instances of white and black smoke. For experiments, I separated the dataset into a 80-15-5 split, for training, validation, and testing respectively. Chapter 3.6c illustrates examples of both black and white smoke across a spectrum of opacities ranging from 10% to 100%. For the purpose of opacity prediction analysis, I divided the dataset into five sub-sets: two corresponding to individual days of testing in Wyoming, and three for each day at the METEC facility, to account for the considerable variability in weather conditions experienced on different days. As depicted in Chapter 3.1, this segmentation approach utilizes the consistent, sunny, and cloudless weather conditions at Wyoming as a foundational baseline for testing. Conversely, the unique weather scenarios encountered on each day at METEC serve to rigorously test the model's accuracy in varying and sub-optimal environmental conditions.

Furthermore, the presence of human experts for re-certification during the Wyoming runs contributed to a more normalized distribution of opacity values. Conversely, for METEC, we aimed for a right-tailed distribution, as depicted in Chapter 3.6a, to test model performance in accurately predicting lower opacity values—a range particularly relevant for environmental permit and



(a) The range of opacity values collected over the five-day study. As smoke opacity is often more difficult for humans to predict at the lower end of the scale, the dataset is right-tailed, primarily focusing on the range of opacities between 5-50.

(b) A sample distribution of weather conditions in the dataset, showcasing clear, cloudy, and overcast days. These images represent part of the diverse environmental settings captured in the dataset.



(c) Images showcasing smoke opacity at every 10% increment (excluding 0%), from 10% to 100% opacity. These illustrate the diversity within the dataset, including both black and white smoke.

Figure 3.6: Comprehensive Analysis of Dataset Characteristics: This figure presents an in-depth look at the dataset's composition, including the distribution of video clips, the variety of weather conditions encountered, and examples of smoke opacity.

licensing compliance, and one that typically poses greater difficulty for human observers to accurately characterize. This approach not only aligns with regulatory interests but also sets a stringent benchmark for model precision in conditions that closely mimic real-world observation challenges.

3.7.2 Interesting Cases

Figure 3.6b visualizes some of the conditions present in the dataset. Weather conditions such as snow, rain, and cloud coverage pose challenges, but also offer an opportunity for enhancing model robustness. Each smoke event was recorded by six different cameras, resulting in unique lighting and background scenarios for every smoke release. In Wyoming, certified experts provided opacity readings as part of their re-certification process in accordance with US EPA Method 9, which serves as a valuable baseline for model evaluation and offers potential directions for future work.

3.8 Synthetic Data

3.8.1 Game and Physics Engines

To complement our real-world dataset, we investigated various methods of synthetic data generation. Ultimately, we selected Unreal Engine 5 for its comprehensive game simulation capabilities, and NVIDIA Omniverse for advanced physics modeling. These engines offer extensive control over variables critical to model training, such as time of day and wind conditions, thereby enhancing the robustness of our models. We discuss the integration and specific uses of these engines in generating synthetic datasets in more detail in [46].

3.8.2 Unreal Engine 5

Unreal Engine 5 emerges as a promising tool in synthetic data generation, offering an accessible platform that is both easy to learn and use. Its abundant availability of free or affordable assets enables rapid scene creation, allowing researchers to swiftly commence data generation. Furthermore, its lower computational power requirements make it an attractive option for research institutions worldwide, making the ability to perform synthetic data generation more accessible. However, while Unreal Engine facilitates quick setup and initial data generation, the fidelity of data for amorphous objects like smoke may be compromised. Achieving high-quality representations of such complex phenomena often necessitates access to advanced simulation tools specifically designed for use within these game engines. Although this lower quality might not significantly

impact training for general datasets, the nuanced and fine-grained details essential for accurately detecting smoke demand a superior level of data realism. Figure 3.7 showcases the qualitative differences with real-world and Omniverse data, illustrating the variance in backgrounds and smoke generated using Unreal Engine. For our study, we integrated up to a total of 280 clips generated via Unreal Engine into our training set, evenly split between 140 clips depicting smoke and 140 clips without, to evaluate the engine's efficacy in supporting smoke detection research.

3.8.3 NVIDIA Omniverse

To achieve more realistic smoke simulations, we opted for NVIDIA Omniverse, a physics simulator known for its high-fidelity outputs. While Omniverse offers unparalleled detail and realism, it introduces specific challenges, including a limited selection of publicly available assets and the need for substantial computational resources. Our experience revealed that running the engine optimally requires at least a 2080 GPU, yet we encountered notable performance issues on a single 3090 GPU. Performance markedly improved when we enabled multi-GPU mode (two 3090's), leading to more efficient data generation. The quality of smoke simulations generated by Omniverse was notably higher to that produced by Unreal Engine. Smoke visualizations in Omniverse were almost indistinguishable from real-world smoke to the human eye, in stark contrast to the more artificial appearance of smoke from Unreal Engine. This visual distinction raised an intriguing question for our research: How significant is the impact of such high-quality synthetic data on model performance? Given our focus on smoke detection, and opacity predictions, it was essential to explore whether the enhanced realism of Omniverse-generated smoke would translate into measurable improvements in model accuracy. Preliminary findings suggest that while the visual quality difference is apparent to the human eye, the incremental benefit for model training, especially in distinguishing smoke from no-smoke scenarios, might be nuanced.

Figure 3.7: Comparison of image quality across simulated and real-world Data: This figure illustrates a side-by-side quality comparison of images from Unreal Engine, NVIDIA Omniverse, and real-world environments, with and without the presence of smoke. These comparisons highlight the simulated data's ability to mimic real-world conditions and demonstrates the effects of environmental elements like smoke on image quality.

Chapter 4

Methods

Smoke features are challenging to extract since smoke has unstructured shapes and unpredictable multidirectional movements [47]. To resolve these challenges, we proposed to employ a black box method to predict accurate opacity (Chapter 4.1). In this chapter, I discuss the OMEGA architecture, my SmokeMamba architecture, Selected Input Data (SID), and the loss function.

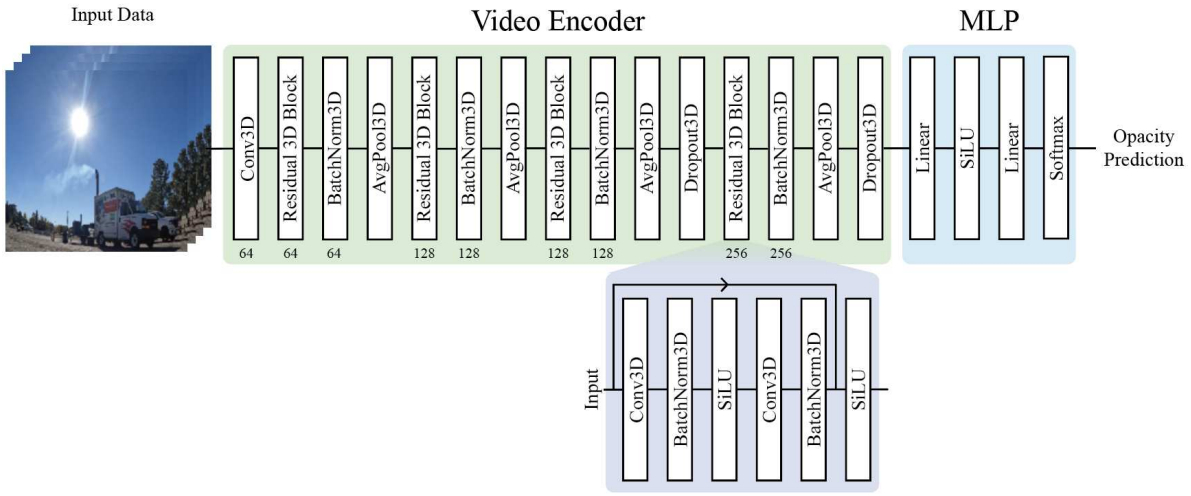


Figure 4.1: Overall architecture of OMEGA. It takes a small number of frames from a video input. The selected frames are fed into Video Encoder that is built with Residual 3D blocks to extract smoke features. In the end, it estimates the opacity class of smoke by MLP.

4.1 OMEGA

In this work, we introduce Opacity Measurement Engine with Graded Accuracy (OMEGA), a novel framework designed to accurately extract and analyze sequential smoke data features from video input. At the heart of OMEGA is a CNN-based Video Encoder, specifically engineered to address the unique challenges of smoke detection in dynamic environments. Unlike traditional approaches, our Video Encoder leverages custom-designed Residual 3D Blocks [48], which are

pivotal for their efficiency and efficacy in capturing temporal smoke patterns without succumbing to overfitting.

To optimize our model for both accuracy and computational efficiency, we introduce a light-weight variant of the Residual 3D Block. To prevent losing temporal information, the kernel and stride sizes are reduced to smaller dimensions. This design choice not only significantly reduces the model's parameter count but also enhances inference speed, making OMEGA well-suited for real-time applications. Furthermore, by selectively processing only four frames from each 40-frame video segment, we maintain high detection performance while minimizing computational load. This selective frame processing is facilitated by our innovative filtering approach, where the video input $\boldsymbol{I} \in \mathbb{R}^{40\times3\times224\times224}$ is effectively condensed into $\hat{\boldsymbol{I}} \in \mathbb{R}^{4\times3\times224\times224}$, ensuring that the Video Encoder focuses on the most informative segments of the input data.

The features extracted by the Video Encoder are subsequently transformed into opacity predictions $\boldsymbol{O} \in \mathbb{R}^{21}$ through a Multi-Layer Perceptron (MLP). This step not only underscores the precision of our feature extraction process but also demonstrates the effectiveness of our approach in generating reliable smoke opacity measurements from video data.

4.1.1 Selected Input Data (SID)

We assumed that using all frames of input frames is inefficient and increases the difficulty of extracting smoke features. Since smoke movement is slow and doesn't have dramatic changes between the next frames, we selected frames by two variables, $\boldsymbol{N_f}$ and $\boldsymbol{N_t}$. The $\boldsymbol{N_f}$ refers a number of frames for our method and $\boldsymbol{N_t}$ is a number of skipped frames between the next frames (Chapter 4.2). Figure N shows the significant changes in smoke when $\boldsymbol{N_t}$ is four instead of one. We had experiments to figure out the best $\boldsymbol{N_f}$ and $\boldsymbol{N_t}$ (Chapter 5.3). Since our approach to preserving temporal information within a small number of frames, we proposed to take smaller kernel and stride sizes to ensure that crucial temporal details are retained within the number of frames. This technique enables the extraction of spatial features while managing temporal features within the MLP layers.

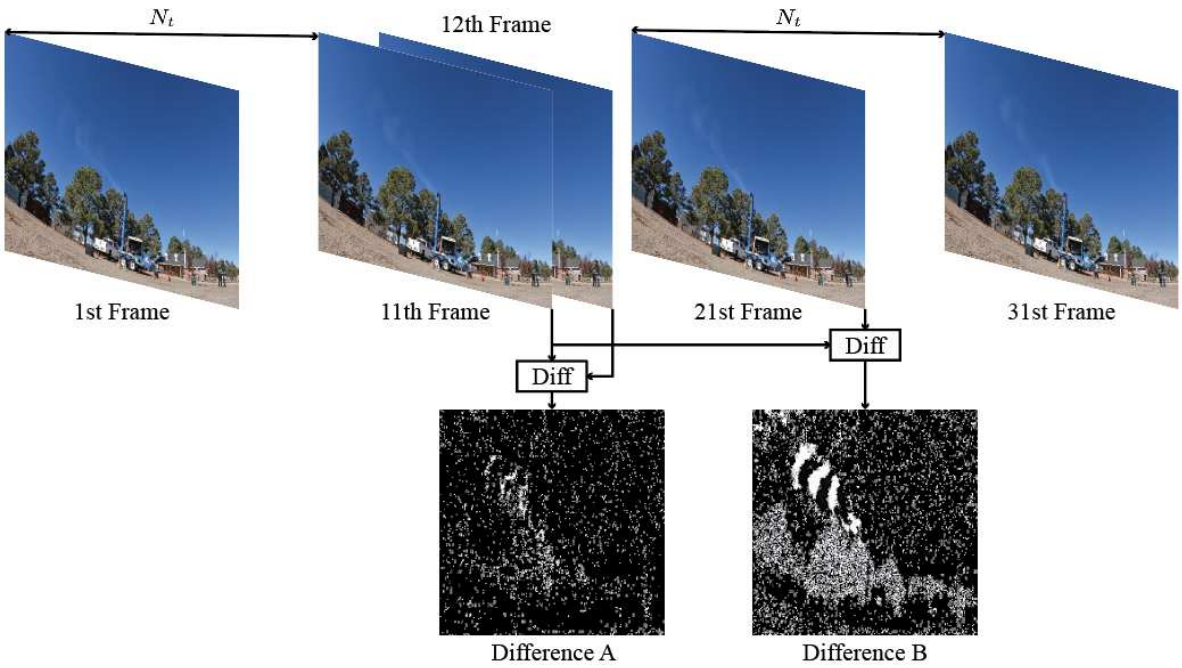


Figure 4.2: Effect of Frame Interval (N_f) on Smoke Detection. This comparison utilizes selected frames (1st, 11th, 21st, 31st) from 40-frame sequences with $N_f = 10$. The 'Difference A' and 'Difference B' images, derived from normalizing absolute frame differences, illustrate that increasing N_f reveals more pronounced smoke changes, highlighting the critical role of frame selection in capturing smoke dynamics.

4.1.2 Loss Function

Given the imbalanced nature of our dataset, as illustrated in Chapter 3.6a, we proposed to employ a novel focal loss to concentrate target opacity classes [49]. For multiple classes, we applied constant values for α and γ, 0.6 and 2.5 respectively (Chapter 4.1) instead of weighted α and γ values for each class [50–53]. This decision to exceed the conventional parameters of $\alpha = 0.5$ and $\gamma = 2.0$ was driven by the aim to more aggressively penalize class misclassifications, thereby mitigating the effects of dataset imbalance. The focal loss function we employ is defined as follows:

$$\mathcal{L}_{focal} = \frac{1}{N} \sum_{i=1}^{N} \left(-\alpha \cdot P_i \cdot (1 - \hat{P}_i)^{\gamma} \cdot \log(\hat{P}_i) \right), \tag{4.1}$$

where i is i-th batch, P are the ground truth logits, and $\hat{P}$ are the model predicted logits.

4.1.3 Training Details

OMEGA was trained on our dataset for 194 epochs with a batch size of 16, using an initial learning rate of $1e^{-4}$. The learning rate was halved at the 20th, 40th, 80th, and 120th epochs. Optimization was performed using AdamW [54], and the SiLU activation function [55] was employed in constructing the Video Encoder and MLP components. Prior to training, the Video Encoder underwent Kaiming initialization [56], setting a robust foundation for model learning.

4.2 SmokeMamba

MAMBA architectures [57], characterized by their Selective State Mechanisms (SSM), have recently gained prominence in the field of machine learning. These architectures offer tailored processing of temporal data, making them highly suited for dynamic applications. SmokeMamba represents a novel application of this technology, specifically designed to handle the complex demands of smoke detection in video data. The system begins with a sophisticated video encoder that extracts critical features from raw footage, which are crucial for identifying and analyzing smoke patterns over time.

The core of SmokeMamba is its custom MAMBA block, which utilizes the linear complexity of SSMs to efficiently process the extracted features. This block is specifically tuned to enhance and capitalize on temporal dynamics, thereby ensuring that subsequent analysis is both accurate and relevant to the temporal context of the video.

Following the MAMBA block, a Long Short Term Memory (LSTM) layer takes over. The choice of LSTM is strategic, leveraging its ability to remember significant details over extended time frames, which is essential for tracking smoke development and movement. This layer's final output is an opacity prediction, which quantifies the presence and density of smoke in each frame, offering valuable insights for environmental monitoring and public safety.

SmokeMamba utilizes the same input present in 4.1.1, loss function 4.1.2, and initial training parameters of OMEGA 4.1.3.

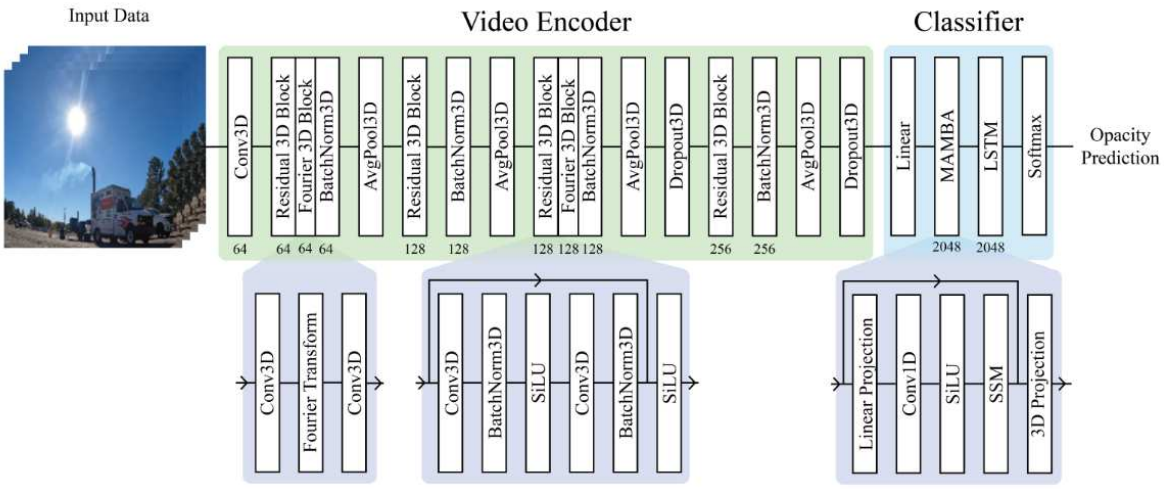


Figure 4.3: Overall architecture of SmokeMamba. Taking in a small number of frames from video data, it employs a video encoder utilizing both Residual3D blocks, and Fourier3D blocks, before being passed into a custom MAMBA block, which feeds into an LSTM which outputs an opacity prediction.

Chapter 5

Experiments

Our experiments employ a precise classification schema that categorizes ground truth opacity readings into 21 distinct classes. Specifically, a reading of 0 is assigned to class 0, while the remaining readings are segmented into 20 classes via the mapping function:

$$class = \frac{score}{\alpha} + 1, \tag{5.1}$$

where 'score' denotes the actual opacity label and α, chosen as 5, defines the classification granularity. This parameterization not only yields 21 classes but also offers flexibility to refine the schema for precise opacity value predictions in future work.

5.1 Metrics

To evaluate our baseline models, we focus on accuracy (top 1, top 3, and top 5), as well as the number of parameters, and the number of videos per second the model can process (VPS). Top 1 accuracy is when we take only the top prediction from the model, and compare it against the ground truth, while top 3 and top 5, see if the correct label is in either the highest 3 or 5 predictions. We choose to utilize this, because when humans make predictions themselves, they are allowed a $\pm$ 15% range on eitherside of the true opacity, which correlates to a top 5 accuracy. Additionally, a low number of parameters and a high amount of VPS is ideal, due to the nature of the task. When implemented in a real-world setting performing live predictions, a machine learning model will need to make predictions not only accurately, but efficiently.

5.2 Baseline Models

To establish a comprehensive baseline for evaluating OMEGA and SmokeMamba, we selected a diverse array of video-based models, including 3D ResNet 18 (R3D18) [48, 58, 59], and Video

Table 5.1: Comparative performance of the SmokeMamba, OMEGA, and baseline models across the full dataset, highlighting Top 1, Top 3, and Top 5 accuracies within a 15% tolerance window for human opacity measurements. The table also compares inference speeds, measured in videos per second (VPS), between 3D models (VST, R3D18, SmokeMamba, OMEGA) that process temporal features and 2D models (ResNet18, sm-ViT) that handle single-frame inputs.

Model	Top 1 ↑	Top 3 ↑	Top 5 ↑	3D	Parameters	VPS
ResNet18	12.70%	34.08%	49.91%	✗	11.2M	–
Sm-ViT	24.20%	54.36%	68.21%	✗	21.7M	–
VST	30.64%	66.47%	82.36%	✓	49.5M	57.6
R3D18	33.17%	69.18%	83.44%	✓	33.2M	104.4
SmokeMamba	**30.64%**	**66.77%**	**82.72%**	✓	**55.6M**	**133.14**
OMEGA	**35.28%**	**69.60%**	**83.62%**	✓	**4.6M**	**170.6**

Table 5.2: Performance evaluation of the custom model over individual days, with a focus on Top-1 accuracy among 21 classes. Performance peaks were observed at Wyoming, whereas METEC posed significant challenges, particularly on days 2 and 3, due to adverse conditions such as occlusion from snow, high cloud coverage, and complete overcast skies. These conditions notably impacted the model's performance, with OMEGA experiencing a discernible decrease in accuracy on these days.

Model	Test Set	Top 1	Top 3	Top 5
OMEGA	METEC Day 1	50.7%	86.3%	**96.3%**
	METEC Day 2	22.9%	59.1%	75.6%
	METEC Day 3	20.2%	48.3%	71.5%
	Wyoming Day 1	47.3%	85.0%	91.8%
	Wyoming Day 2	**52.3%**	**88.3%**	95.1%

Swin Transformer (VST) [60], as well as image-based models like ResNet18 [61] and a small version of the Vision Transformer (SmViT) [62–64], for application to our dataset.

5.2.1 Data Split

For effective model evaluation, we structured our dataset into training, testing, and validation splits following an 80-15-5 percentage distribution. Specifically, the testing set comprised three randomly selected runs from each day, while the validation set included one randomly selected run from each day. The remaining runs formed the training set. This resulted in a division where the training set contained 11,304 clips, the testing set 1,667 clips, and the validation set 571 clips.

5.2.2 Data Augmentation and Training

To prepare the data for training, we applied several augmentation techniques to enhance model robustness and prevent overfitting. These included: random resized crop, random horizontal flip, and random augment for comprehensive and varied transformations.

For image-based models like ResNet 18 and Vision Transformers (ViTs), we utilized additional mixup augmentation, which was not applied to video-based models. All inputs were standardized by normalizing the RGB channels using the ImageNet means (0.485, 0.456, 0.406) and standard deviations (0.229, 0.224, 0.225) for the red, green, and blue channels respectively.

In terms of processing, the 2D models (e.g., ResNet18 and SmViT) processed individual frames from each clip, labeling them with the ground truth opacity values of their corresponding clips. Frames were resized to 224 x 224 pixels from the original high resolution of 1920 x 1080.

The training duration was set to 150 epochs for baseline models, while the custom OMEGA model underwent an extended training period of 300 epochs to fully leverage its complex architecture.

5.3 Model Results

Table 5.1 compares the performance of baseline models to OMEGA, highlighting the R3D18 model as the most effective baseline with a top 1 accuracy of 33%. This underscores the importance of temporal feature analysis in accurate smoke opacity prediction.

Further analysis, detailed in Table 5.2, assessed OMEGA's performance across different days, illustrating how variable conditions affect accuracy. OMEGA outperformed the R3D18 baseline by approximately 2%, achieving a top 1 accuracy of 35.3%. All models performed the best during the initial days at Wyoming but faced challenges on the second and third days at METEC, primarily due to adverse weather conditions. Notably, OMEGA's efficiency is highlighted by its parameter count of only 4.6 million, in contrast to the best baseline, R3D18's 33.2 million, underscoring the model's suitability for real-time opacity prediction tasks.

5.4 Ablation Study

5.4.1 OMEGA

An ablation study was conducted to ascertain the effectiveness of our proposed architecture, exploring the impact of different temporal data configurations on model performance. This investigation encompassed a series of experiments with varied combinations of N_f (number of frames) and N_t (interval between frames). The results revealed that configurations with a broader time gap between four frames achieved the highest accuracy, highlighting the critical role of temporal information in enabling the model to detect subtle changes in smoke dynamics across frames. These findings, detailed in Table 5.3, underscore the balance between temporal resolution and model efficiency. Furthermore, our study demonstrates that reducing the frame sample to approximately 10% of the video significantly accelerates inference times (170.6 videos per second (Table 5.1)), a crucial factor for real-time applications. This suggests that effective opacity predictions can be achieved with a more compact model, optimizing both performance and computational efficiency.

Table 5.3: Results of the ablation study on temporal data configurations. This table presents the comparative analysis of model performance across varying combinations of N_f (number of frames) and N_t (frame intervals). Highlighting the optimal configuration of a wider temporal gap between four selected frames, the table illustrates the significant role of temporal information in enhancing opacity prediction accuracy.

N_f	N_t	Top 1	Top 3	Top 5
2	5	29.8%	65.4%	81.1%
2	10	34.0%	70.3%	85.2%
2	15	31.8%	**70.5**%	**86.2**%
4	5	31.7%	66.7%	82.8%
4	10	**35.3**%	69.6%	83.6%
8	5	34.3%	68.8%	83.9%

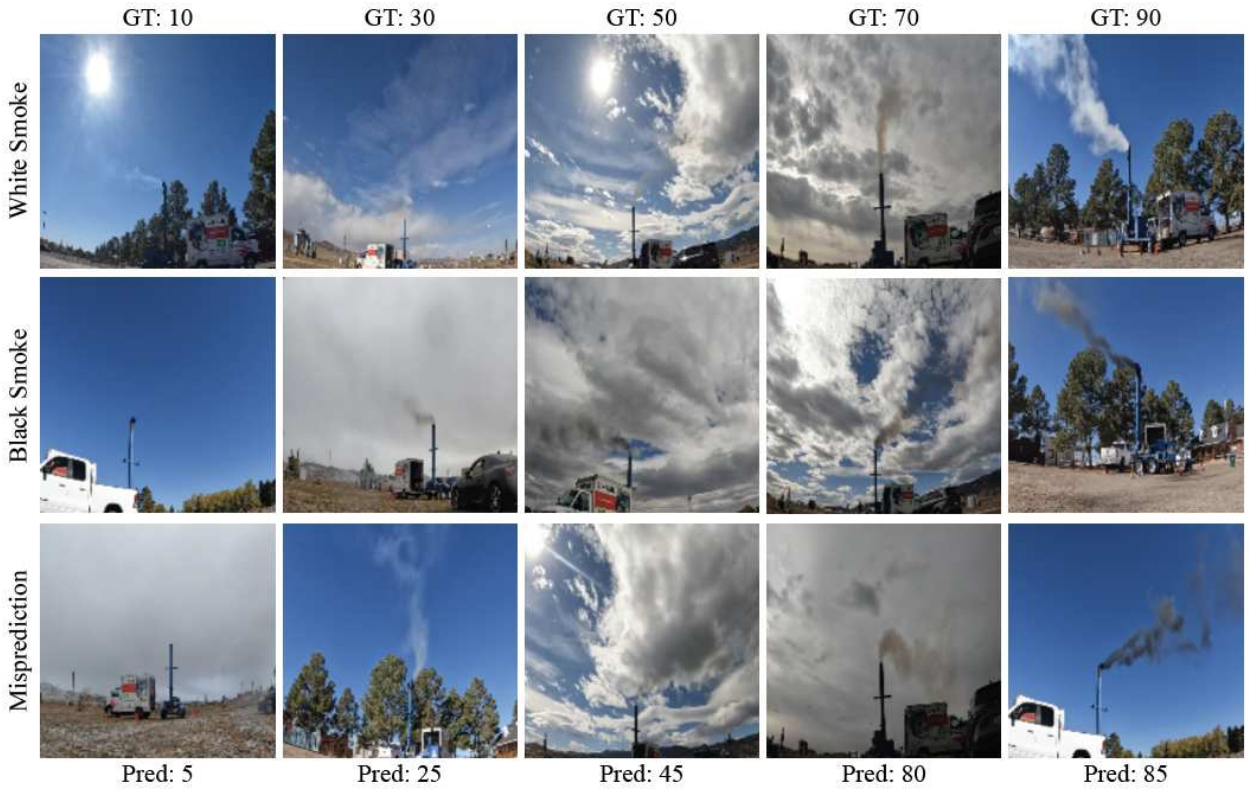


Figure 5.1: Visualization of the OMEGA model's performance, showcasing correct classifications of black and white smoke across a range of opacities from 10% to 90%, and instances of missed predictions. The bottom row presents mispredictions and highlights the challenges posed by adverse weather conditions. In 85% of cases (see Table 5.1), the model's predictions were within an acceptable range of ±10% of the ground truth.

5.4.2 SmokeMamba

To validate the effectiveness of each component in the SmokeMamba architecture, I conducted an ablation study. This study involved systematically removing key components—one at a time—and observing the impact on the model's performance. The focus was on three critical elements: the Fourier3D blocks, the MAMBA block, and the LSTM layer. The results demonstrate that each component contributes significantly to the architecture's ability to detect and analyze smoke in video data. I chose to evaluate the model primarily on Top 1 accuracy, but also include Top 3 and Top 5 accuracy's in the ablation study. These metrics are visualized in Table 5.4, which presents a detailed comparison of the architecture's performance with and without each component.

Table 5.4: An ablation study exploring the capabilities of individual components of SmokeMamba. These experiments are some of the first of their kind, and I found the MAMBA block highly successful in increasing a models accuracy.

Component	Top 1	Top 3	Top 5
No-Fourier/LSTM	25.8%	61.4%	79.4%
No-Fourier/MAMBA	27.3%	64.6%	81.1%
No-MAMBA	28.4%	63.1%	78.6%
SmokeMamba	**30.6%**	**66.7%**	**82.7%**

Chapter 6

Synthetic Data Experiments

In this chapter, we explore the balance and quantity needed when incorporating synthetic data into a real-world smoke dataset. These experiments incorporate the Unreal Engine generated data and the NVIDIA Omniverse data into Smoke+, for the task of smoke detection.

6.1 Data Splits

We utilize a small portion of SMOKE+, comprising of 1,774 video clips, with 1,554 featuring smoke and 220 without. It is divided into training, validation, and testing sets to facilitate a thorough evaluation: 370 smoke and 190 non-smoke clips for training, 319 smoke and 6 non-smoke clips for validation, and 865 smoke and 24 non-smoke clips for testing. The lack of data size and distribution is already focused on in other studies [65–68], but we propose to address this problem by incorporating synthetic data into the training set in our study. In addition, our dataset distribution, particularly the expansive unseen testing set, is proposed for enhanced generalization on evaluation, despite the constraints on the size of training set, which may help in future works, such as opacity predictions.

To address the complexity of smoke detection, we focus on the opacity of smoke, quantified by the equation:

$$Opacity = \left(1 - \frac{I}{L}\right) \times 100, \tag{6.1}$$

where I/L represents the transmittance of light through the smoke plume, which after being subtracted from one may be converted into a percentage [69]. Given that detecting smoke can be straightforward, our study only includes clips with opacity values between 5-30%, thus elevating the detection challenge by focusing on subtler smoke patterns.

6.2 Experiments

In our experiments, we address the challenge of limited real-world data availability by starting with a modest set of 560 real-world samples. To explore the effectiveness of synthetic data in enhancing model performance, we incrementally introduced additional synthetic samples, each increment amounting to 5% of the original real-world dataset size, aiming to identify the optimal synthetic-to-real data ratio for improved model accuracy.

Moreover, to intensify the challenge and more closely mimic real-world complexities, the smoke featured in the real-world data was deliberately chosen to have an opacity of 30% or less. This choice was made to simulate the difficulty models face in detecting low-opacity smoke, which is often more subtle and harder to distinguish. Conversely, the synthetic data was generated with higher opacity levels, with the intention of facilitating the model's learning process by providing clearer examples of smoke features. This experimental setup was designed not only to test the model's ability to learn from limited data but also to evaluate the impact of synthetic data quality, in terms of opacity, on the learning outcomes.

Table 6.1 presents a comparison between the SemiS model and established baseline models, specifically a 3D ResNet model (R3D) [48] and a tiny Video Swin Transformer (VST) [70], with training conducted solely on real-world data. R3D extends the traditional 2D ResNet framework into three dimensions, adapting it for action recognition tasks in video sequences. Similarly, VST adapts the Swin Transformer [71] architecture for video analysis, leveraging its strengths in capturing complex spatial-temporal relationships. The SemiS model outperformed these baselines, achieving the highest accuracy of 89.99%, demonstrating its capability in accurately detecting smoke features and indicating its potential for advancing computer vision tasks involving amorphous objects.

To investigate the impact of synthetic data on model performance, we augmented the initial set of 560 real-world training samples with synthetic data generated from both Unreal Engine and Omniverse. As depicted in Figure 6.1, the classification accuracy of SemiS consistently showed higher results with the Omniverse-generated synthetic data, indicating its superior alignment with

Method	Parameters	Accuracy (↓)	VPS (↑)
R3D	33.4 M	72.55 % %	37.1
VST (t)	28.2 M	84.70 % %	35.5
SemiS	**9.2 M**	**89.99 %**	**41.5**

Table 6.1: Results for the baseline tests for smoke detection over the real world dataset only. Two baseline models, one CNN (R3D) and one video transformer (VST Tiny), were compared to an efficient (9.2M parameters) model, named SemiS. SemiS achieved the highest accuracy by 8.5% over the best baseline VST.

the nuances of real-world smoke detection. The optimal integration of synthetic data, enhancing accuracy maximally, was found to be an additional 30% of the original dataset size for both sources.

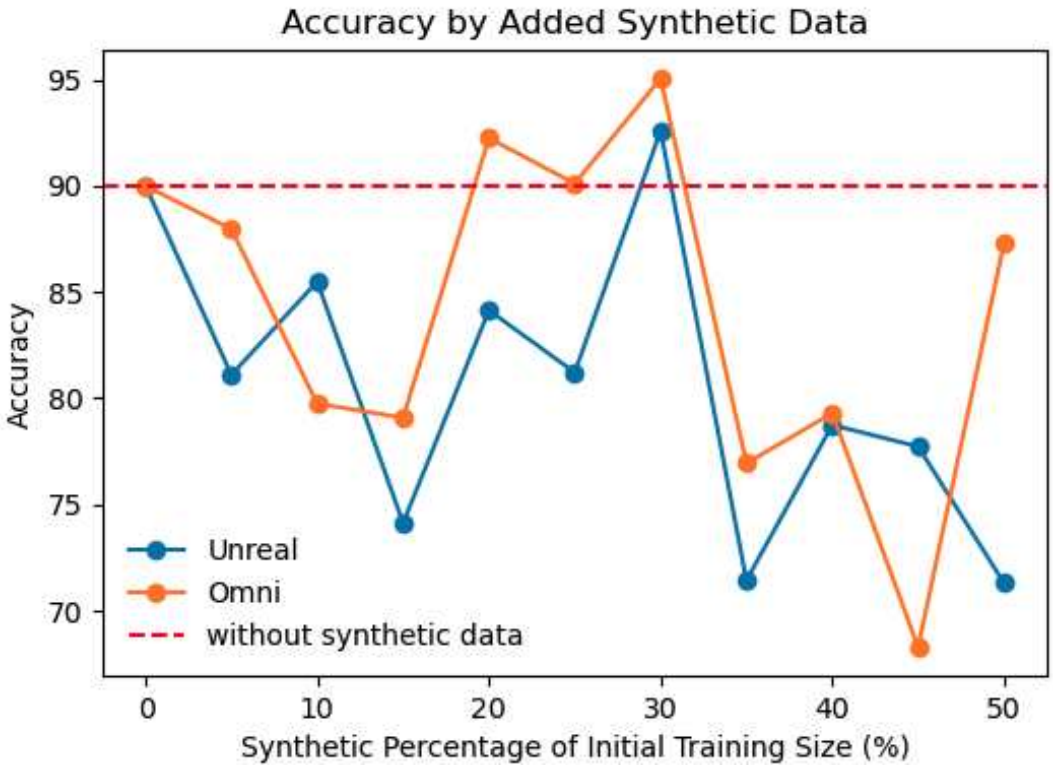


Figure 6.1: Graph of synthetic data integration vs. model accuracy: This graph plots the model's accuracy as a function of the synthetic data proportion added to the training set, where 50 indicates that 50% of the initial real-world data size was incorporated as synthetic data. The red line highlights the baseline accuracy achieved with no synthetic data. The trend illustrates how incorporating synthetic data impacts the model's performance, providing a visual comparison to the baseline scenario.

Chapter 7

Discussions

7.1 Overview on the Importance of Smoke Monitoring

Smoke, whether from industrial emissions or wildfires, presents significant challenges for environmental and public safety. Effective monitoring and regulation of smoke are crucial for mitigating its impact on health and safety. This thesis explores the application of computer vision techniques in industrial settings to enhance smoke detection and monitoring, building on my previous contributions to the field.

In the study [72], we focused on predicting the spread of wildfires using a top-down view of a 1km by 1km region previously affected by forest fires. This work demonstrated the potential of computer vision to provide valuable insights into smoke behavior and spread in natural environments.

Similarly, the SMOKE+ dataset, utilized in this thesis, comprises multiple views of a single smoke release in an industrial context. The methods developed here are broadly applicable, underscoring the versatility of computer vision. These tools are not only vital for researchers but also for on-the-ground experts like firefighters and opacity inspectors who rely on accurate, real-time data to make critical decisions.

The development and refinement of datasets such as SMOKE+ are pivotal for training accurate machine learning models. These models can significantly enhance the capabilities of computer vision systems, making them indispensable in both emergency response scenarios and routine environmental monitoring. The continued evolution of such technologies promises greater efficacy in our ongoing efforts to manage and mitigate the effects of smoke across diverse settings.

7.2 SMOKE+ Discussions

7.2.1 Smoke Colors

Our analysis, highlighted in Chapter 5.1, shows the model's precision in predicting smoke opacity, particularly under clear conditions. Performance drops in adverse weather or obscured views, conditions which are similarly challenging for human evaluators. A significant discrepancy was observed in accuracy for black versus white smoke: OMEGA recorded top 1, top 3, and top 5 accuracies of 50.8%, 88.1%, and 96.2% for black smoke, against 29.3%, 63.2%, and 79.5% for white smoke. This indicates that while our model adeptly identifies smoke patterns, the contrast between smoke types, especially in challenging conditions, requires further refinement."

7.2.2 Weather Conditions Across Days

During the initial two days of data collection in Wyoming, we were fortunate to experience nearly perfect weather conditions, characterized by bright skies and minimal cloud cover. These conditions are ideal for computer vision tasks, as they ensure consistent lighting and visibility, making these segments of the dataset particularly suitable for efficient model training.

At the METEC facility, the weather presented more of a challenge. The first day there was marked by heavy cloudiness, posing difficulties as the smoke from tests could easily blend into the gray backdrop, complicating the task of smoke detection and tracking. The following day continued with overcast conditions, though without precipitation, which, while slightly better than the first, still posed issues with lower light levels and reduced contrast in the visual data.

The third day at METEC was the most challenging in terms of weather conditions. Data recording commenced early in the morning amidst a light snowfall, introducing additional visual noise and obstruction as snowflakes accumulated on the camera lenses. Though the snow ceased later, the persistent overcast skies and residual droplets on the camera lenses continued to significantly impact the visual quality of the recordings. This day's data, therefore, provides a crucial test case for assessing and improving the robustness of our vision algorithms under adverse weather conditions.

7.3 Future Works

Future efforts will aim to utilize stereo-vision data for 3D smoke modeling, leveraging NERF techniques [73, 74], despite potential challenges posed by diverse weather conditions and the range of smoke opacities. Additionally, we intend to develop a Bayesian framework for opacity prediction, inspired by existing frameworks [75], and to employ supervised contrastive learning [76, 77] for model refinement. This approach will focus on enhancing model sensitivity to subtle class distinctions, such as differentiating between closely ranged opacity levels. Finally, adding in heatmaps or EgoPose labels such as seen in [78], could be interesting.

7.3.1 Future Bayesian Network

Utilizing the human score sheets, we intend to train a Bayesian Neural Network (BNN), where we will utilize OMEGA as a black box component that provides input into our BNN, which has been trained on the human labels. the goal of this is to introduce a new model, that is capable of predicting exactly as a human expert does. With this in mind, we will enforce the strict rules on our model, where any miss on a prediction by 15% or more is an automatic failure, and train it to comply with EPA regulations.

7.3.2 Synthetic Data Expansion

One limitation with our synthetic data is the omission of opacity labels. In game engines, it's possible to emit a light source, and a light reader on the other side of an object such as smoke, and get accurate light measurements, similar to how our real dataset was constructed. utilizing this technique, we can record an infinite amount of data, specifically for higher end opacities, where real-world implications of smoke releases make it difficult to obtain. Additionally, as we found weather to be a large factor in poor model performance on parts of the dataset, training a model with synthetic opacity labels in poor conditions, will help with more fine grained estimations.

7.3.3 Virtual Reality Adaptation

As virtual reality (VR) becomes more and more common, fields such as education will begin adopting it on a rapid basis. Similarly to what I presented in [79], I can make an educational VR environment, utilizing 3D smoke for opacity reading training. Using VR for a system such as this could eliminate the need not only for experts to travel and take time off work for re-certification, but also reduce the amount of emissions caused by re-certification every year. A fully VR environment allows for experts to be trained in any weather condition, and lighting scenario, providing a far more robust certification event, than what is currently provided.

7.4 Limitations

7.4.1 3D Modeling

The current approach to 3D modeling of smoke utilizing NeRF (Neural Radiance Fields) models, which are commonly used for 3D scene reconstruction, typically require imagery from a single camera that has recorded the entire scene. This dataset, however, consists of footage captured from six static cameras, each providing a different viewpoint. This setup has made it difficult to successfully apply NeRF techniques, as they have mostly produced extremely blurry scenes when attempting to render from these multiple, disjointed perspectives.

7.4.2 Model Deployments

Furthermore, the dataset was collected during a private event and does not include footage of real-world industrial-scale smoke stacks. This specificity limits the dataset's applicability and realism for broader smoke analysis applications. While this dataset is suitable for extracting basic smoke features and for pre-training the model, additional, more varied data will likely be necessary to fine-tune and effectively deploy the model in real-world environments where conditions and smoke behaviors can be vastly different.

7.5 Dataset Interest

In the course of this research, the methodologies and techniques developed from the SMOKE+ were not only validated but also refined for enhanced application. These refined techniques played a crucial role in the subsequent curation of another dataset, titled CSU101. Specifically designed for educational purposes within the field of computer vision, the CSU101 dataset incorporates lessons learned from SMOKE+ to ensure more robust and applicable data collection and labeling processes. Chapter 8 provides a detailed account of the CSU101 dataset, discussing its design rationale, the enhanced data collection methodologies employed, and the specific improvements made over previous techniques.

Chapter 8

CSU101 Introduction

8.1 Segway From SMOKE+

Building on the refined techniques for data collection developed during the SMOKE+ project, I embarked on creating a new dataset focused on educational applications in computer vision. This dataset, titled CSU101, adopts the same successful methodologies used during the SMOKE+ data collection phase, including specific equipment choices (such as the GoPro Hero Black 10), a streamlined labeling process, and rigorous label verifications. This chapter introduces the CSU101 dataset and delves into its statistics and curation process.

8.2 CSU101 Overview

In this work, I introduce CSU101, the first dataset, to my knowledge, that is specifically tailored for educational use in computer vision and designed to engage university students globally. This building-centric dataset features images extracted from video footage of university buildings, annotated with object detection labels for objects commonly found on campuses. CSU101 facilitates learning both image classification and object detection, enabling students to master core computer vision techniques through a unified dataset. Initially presented as processed images, the dataset supports a structured learning progression from simple image-based tasks to more complex video data processing. To aid in educational delivery, I have included starter Jupyter Notebooks structured as lesson plans, which have proven beneficial in classroom environments [80].

8.3 Contributions

In summary, CSU101 provides four main contributions. First, CSU101 to the best of my knowledge, is the first ever Computer Vision dataset specifically tailored to university students for educational purposes. Second, CSU101 provides comprehensive labels for both image clas-

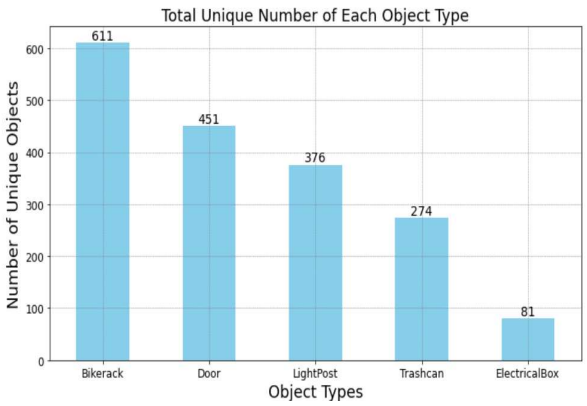
sification and object detection, allowing learners to utilize a single dataset for mastering multiple core Computer Vision tasks. Third, CSU101 includes pre-processed images from video footage, enabling beginners to focus initially on simpler image-based tasks before advancing to video data processing, thus offering a graduated learning experience. Fourth, CSU101 is supplemented with documented Jupyter Notebooks that serve as structured lesson plans, facilitating immediate, practical engagement and reinforcing theoretical concepts within an educational framework. Table 8.1 provides an in depth comparison of CSU101 and other building datasets.

Table 8.1: Comparison of *CSU101* with other Building Datasets. This table highlights differences in the number of labeled frames, unique buildings, and objects, and indicates the presence of bounding box labels for object detection. Notably, *CSU101* ranks second highest in the number of labeled frames and leads in the number of unique labeled objects.

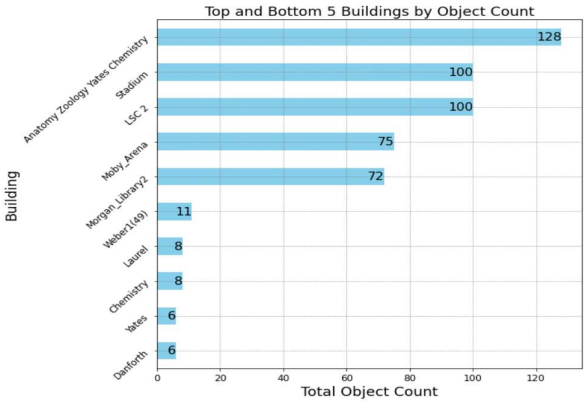
Dataset	Video Data	Frames	Buildings	Objects	Bounding Boxes
Obeso et al. [81]	✗	284	–	–	✗
Shalunts et al. [82]	✗	400	–	–	✗
Shao et al. [83]	✗	1,005	201	–	✗
Taoufiq et al. [84]	✗	1,033	–	–	✗
Xu et al. [85]	✗	$\sim$ 5,000	–	–	✗
Taoufiq et al. [84]	✗	6,297	–	–	✗
Cordts et al. [86]	✓	25,000	–	–	✗
Zheng et al. [87]	✗	$\sim$ 21.4 million	5,312	–	✗
Philben et al. [88]	✗	5,062	11	11	✓
Philben et al. [89]	✗	6,300	11	11	✓
Barz and Denzler [90]	✗	9,485	9,346	566	✓
CSU101(Ours)	✓	**277,056**	**48**	**750**	✓

Chapter 9

CSU101 Dataset



(a) Unique objects across the dataset.



(b) The top and bottom 5 building object counts.

Figure 9.1: Statistics of the *CSU101* dataset. Figure (a) showcases the number of *unique* objects present in the dataset, highlighting a common campus trend: a high frequency of bike racks and a relatively low occurrence of electrical boxes. Figure (b) details the top and bottom five counts of total objects per building. For instance, The Stadium, a larger area, contains 100 unique objects, while Danforth, a smaller building, includes only 6 unique objects.

9.1 CSU101 Construction

9.1.1 Dataset Collection

Similarly to SMOKE+, data was collected using a GoPro Hero 10 camera, set to record at 24 FPS with a resolution of 1080p. The data collection procedure involved systematically walking around each building's perimeter and capturing all of its faces using the GoPro cameras. In total, 48 university buildings were recorded, with researchers fully encircling each building when possible, resulting in a total of 252,837 labeled frames. During these recordings, care was taken to ensure that the entire building was within the frame, along with surrounding elements such as the ground, to facilitate the labeling of objects like bike racks and trashcans in addition to the buildings themselves. This comprehensive approach to filming was designed to capture a wide array

of angles and perspectives, enhancing the dataset's utility for both object detection and building classification tasks.

9.1.2 Labels

For building classification, we provide JSON files containing the labels along with corresponding data splits. To ensure robust model training and evaluation, each building video was labeled at every 20th frame, with some final frames intentionally excluded to avoid overlap between training and testing datasets, thereby preventing wraparound issues. For object detection, labels are formatted according to the YOLO [91] standard. This involves labeling every frame of each video in a separate text file, where each line represents an object with five attributes: the class identifier, the center x and center y coordinates of the bounding box, and the dimensions (height and width) of the bounding box. This detailed labeling facilitates precise object localization and is crucial for training effective detection models

9.1.3 Intrarater Reliability

To validate the accuracy of our object detection annotations in the CSU101 dataset, we conducted an intrarater reliability assessment using the Kappa coefficient [92, 93]. The five annotators who contributed to CSU101 were tasked with independently annotating the same video (Forestry), each without specific prior instructions for this exercise. After annotation, the YOLO-formatted labels from each annotator were collected and compared. The primary expectation was that each annotator would identify a consistent number of objects (15 objects were present in the video), with a focus on maintaining high accuracy in their annotations. From this comparison, we calculated a Kappa score of .80, demonstrating the reliability and consistency of our annotations across different annotators. The Kappa coefficient (κ) is calculated using the formula: $\kappa = (P_0 - P_e) * (1 - P_e)^{-1}$, where P_o is the observed agreement among raters, and P_e is the hypothetical probability of chance agreement. This measure adjusts for the agreement that would naturally occur by chance, providing a more accurate reflection of the annotators' reliability.

9.1.4 Object Classes

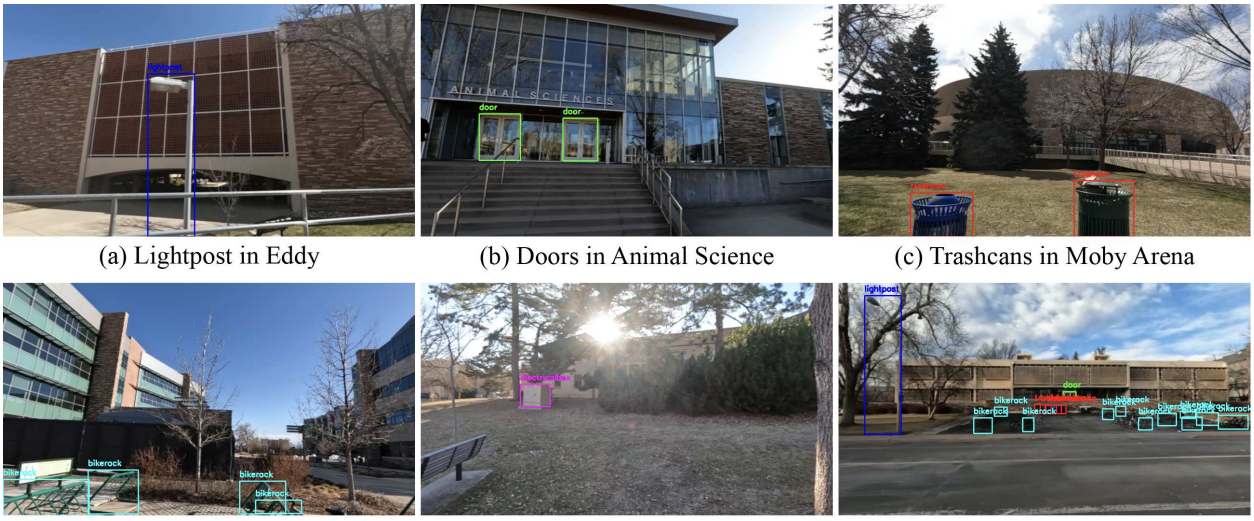
One common issue with many educational datasets is the overwhelming number of object classes, which can complicate the learning process. To address this, we carefully selected only five classes that are often present on campus: bike racks, light posts, doors, trash cans, and electrical boxes. These classes were chosen for their familiarity to university students, ensuring that the objects are easily identifiable and relatable. This focused approach not only simplifies the learning experience but also makes it easier for students to understand how models function and to interpret their outputs. By reducing complexity, we aim to simplify the principles of object detection and enhance the practicality of experiments conducted with this dataset. The overall distribution of unique objects (counting only each occurrence of an object once per video), can be seen as right tailed in Figure 9.1a. As expected on a college campus, there are a large number of bike racks present (509 in total), while electrical boxes were the least present (68 in total). Additionally, per building, we saw averages of: (Bike racks: 9.98, Doors: 7.71, Light Posts: 6.27, Trashcans: 5.30, Electrical Boxes: 1.42), depicted in Figure 9.1b.

9.1.5 Data Annotation

The annotation process was conducted using the Computer Vision Annotation Tool (CVAT). Five object categories relevant to campus infrastructure were annotated: light posts, bike racks, trash cans, electrical boxes, and doors (specifically of buildings). Annotations were performed by annotators who manually drew bounding boxes around each instance of the target objects. To ensure accuracy and consistency, the bounding boxes were adjusted every 20 frames. Each bounding box is represented using two points that define a square enclosing the object of interest.

9.1.6 Dataset Split (Building Classification)

As each video contains a building that has been fully encircled, we needed to be careful about how we chose to split the data. To minimize scene overlap at 24 FPS, we selected every 80th frame from each video, resulting in a total of 7,440 frames. After pruning the frames, the average number



(a) Lightpost in Eddy (b) Doors in Animal Science (c) Trashcans in Moby Arena

(d) Bikeracks in Biology (e) Electrical Box in Ammons (f) Multiclass Objects in Education

Figure 9.2: This figure presents an up-close view of each object type included in the dataset, along with an image featuring multiple objects. A common theme observed throughout the dataset is the grouping of bike racks, as exemplified in image (f).

of frames per video is 155, with a minimum of 31 frames and a maximum of 395 frames per video. From the selected frames, we distributed 60%, 20%, and 20% of the shuffled frames per video to the training, validation, and testing sets, respectively.

9.1.7 Dataset Split (Object Detection)

Unlike image classification, not all buildings needed to be present in each split for object detection. As we had 56 videos (48 buildings, some were unable to be completed in a single video due to fencing, overlapping the same scenes, and etc.), we decided to withhold 5 videos for validation, and 5 videos for testing, while the remaining 38 videos went into the training set. To ensure consistency and quality, we decided to choose 10 videos for validation and testing that were a sizeable amount of frames (4500), and contained a good distribution of objects. For example, for the Forestry building in the testing set, it contained 4728 frames, 10 light posts, 4 doors, 2 trashcans, 4 bike racks, and 4 electrical boxes. As similarly as Section 9.1.6, we selected every 20th frame from each video, and then assigned the pruned frames to corresponding sets by the building name. The training, validation, and testing sets contain 7,334, 808, and 643 frames respectively.

(a) View 1 (b) View 2 (c) View 3

Figure 9.3: Three separate views illustrate the diversity of objects found in a single video, in this case, the Computer Science Building. View 1 displays four distinct objects, View 2 features two, and View 3 highlights a grouping of multiple objects, specifically bike racks. This demonstrates the wide variety of objects that a single building can present within the dataset.

9.1.8 CSU101 as an Educational Resource.

CSU101 is designed to be an accessible and valuable resource for introductory Machine Learning courses. The dataset's manageable size allows for efficient training and experimentation on standard hardware, making it ideal for educational settings. The inclusion of diverse scenes and lighting conditions introduces students to the challenges of real-world object detection, preparing them for more complex tasks.

To further enhance its educational value, we provide a suite of Jupyter notebooks tailored for beginners. These notebooks offer step-by-step guidance, from dataset loading and visualization to model training and evaluation. Extensive comments and explanations within the notebooks aim to aid students in understanding the underlying concepts and techniques. Additionally, pre-trained models will be made available, enabling students to explore object detection without requiring extensive computational resources.

Chapter 10

Conclusions

10.1 SMOKE+

In this thesis, I have introduced the SMOKE+ dataset (Systematic Measurement of Opacity for smoKe Evaluation with Synthetic Data), which serves as a pivotal resource for real-time smoke opacity evaluation. This dataset is distinguished by its comprehensive labeling of opacity levels across a wide range of weather conditions, which enhances its utility for environmental monitoring. A significant innovation in SMOKE+ is the integration of synthetic data, which addresses gaps in real-world data and improves the robustness of machine learning models trained on this dataset.

Through rigorous experiments and baseline comparisons, it has been demonstrated that SMOKE+ can effectively support the application of machine learning techniques to assess and interpret smoke opacity. This work not only advances the field of smoke opacity measurement but also has practical implications for real-time environmental monitoring in both industrial and public sectors.

Future work will focus on expanding the dataset to include more diverse environmental conditions and further refining the balance of synthetic and real data to optimize model training and performance.

10.2 CSU101

Building on the methodologies developed for SMOKE+, the CSU101 dataset has been curated to serve as an educational tool in computer vision. This dataset comprises video footage of 48 university buildings, segmented into 277,056 labeled images and featuring 750 unique objects across five distinct classes. Designed specifically for introductory computer vision courses, CSU101 offers a practical, hands-on learning experience through its focus on fundamental tasks such as image classification and object detection.

To enhance the educational value of CSU101, I have provided detailed tutorials through Jupyter Notebooks. These include six baseline models for image classification and five for object detection, enabling students to engage with both the theoretical aspects and practical challenges of computer vision.

The ultimate aim of CSU101 is to empower learners globally, fostering innovations in educational methods within the field of computer vision. By making learning both accessible and engaging, this dataset encourages students to explore advanced studies and pursue careers in this vibrant field.

Plans for the future development of CSU101 include expanding the range of environments and objects covered in the dataset, thus increasing its robustness and relevance to real-world applications. Such enhancements will better prepare learners to address the challenges they will encounter in professional settings.

Bibliography

[1] Shubhangi Chaturvedi, Pritee Khanna, and Aparajita Ojha. A survey on vision-based outdoor smoke detection techniques for environmental safety. *ISPRS Journal of Photogrammetry and Remote Sensing*, 185:158–187, 2022.

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[3] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points, 2019.

[4] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification, 2015.

[5] Shengxin Zha, Florian Luisier, Walter Andrews, Nitish Srivastava, and Ruslan Salakhutdinov. Exploiting image-trained cnn architectures for unconstrained video classification, 2015.

[6] Gao Xu, Yongming Zhang, Qixing Zhang, Gaohua Lin, Zhong Wang, Yang Jia, and Jinjun Wang. Video smoke detection based on deep saliency network. *Fire Safety Journal*, 2019.

[7] Gao Xu, Qixing Zhang, Dongcai Liu, Gaohua Lin, Jinjun Wang, and Yongming Zhang. Adversarial adaptation from synthesis to reality in fast detector for smoke detection. *IEEE Access*, 2019.

[8] Gao Xu, Yongming Zhang, Qixing Zhang, Gaohua Lin, and Jinjun Wang. Deep domain adaptation based video smoke detection using synthetic smoke images. *Fire safety journal*, 2017.

[9] Rui Ba, Chen Chen, Jing Yuan, Weiguo Song, and Siuming Lo. Smokenet: Satellite smoke scene detection using convolutional neural network with spatial and channel-wise attention. *Remote Sensing*, 2019.

[10] Gaohua Lin, Yongming Zhang, Qixing Zhang, Yang Jia, Gao Xu, and Jinjun Wang. Smoke detection in video sequences based on dynamic texture using volume local binary patterns. *KSII Transactions on Internet and Information Systems*, 2017.

[11] Feiniu Yuan, Lin Zhang, Boyang Wan, Xue Xia, and Jinting Shi. Convolutional neural networks based on multi-scale additive merging layers for visual smoke recognition. *Machine Vision and Applications*, 2019.

[12] Marin Bugarić, Toni Jakovčević, and Darko Stipaničev. Adaptive estimation of visual smoke detection parameters based on spatial data and fire risk index. *Computer vision and image understanding*, 2014.

[13] ByoungChul Ko, JunOh Park, and Jae-Yeal Nam. Spatiotemporal bag-of-features for early wildfire smoke detection. *Image and Vision Computing*, 2013.

[14] Kosmas Dimitropoulos, Panagiotis Barmpoutis, and Nikos Grammalidis. Spatio-temporal flame modeling and dynamic texture analysis for automatic video-based fire detection. *IEEE transactions on circuits and systems for video technology*, 2014.

[15] B Uğur Töreyin, Yiğithan Dedeoğlu, and A Enis Cetin. Wavelet based real-time smoke detection in video. In *2005 13th European Signal Processing Conference*. IEEE, 2005.

[16] Alexander Filonenko, Laksono Kurnianggoro, and Kang-Hyun Jo. Smoke detection on video sequences using convolutional and recurrent neural networks. In *International Conference on Computational Collective Intelligence*. Springer, 2017.

[17] Yen-Chia Hsu, Ting-Hao Kenneth Huang, Ting-Yao Hu, Paul Dille, Sean Prendi, Ryan Hoffman, Anastasia Tsuhlares, Jessica Pachuta, Randy Sargent, and Illah Nourbakhsh. Project rise: Recognizing industrial smoke emissions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14813–14821, 2021.

[18] Ralph C Stahman, Kenneth D Mills, and Merrill W Korth. Federal air pollution efforts, the early years. *SAE transactions*, pages 446–455, 1989.

[19] M L Brubacher. Reduction of diesel smoke in california. *SAE Prog. Technol.; (United States)*, 12, 1966.

[20] Roger C. Bascom, Wen S. Chiu, and Ronald J. Padd. Measurement and evaluation of diesel smoke. *SAE Transactions*, 82:814–830, 1973.

[21] Frank Uekoetter. The strange career of the ringelmann smoke chart. *Environmental Monitoring and assessment*, 106:11–26, 2005.

[22] Shawn D Dolan. After considerable struggle, epa method 082 modernized visible emissions monitoring. *Natural Gas & Electricity*, 34(5):21–25, 2017.

[23] Michael J McFarland, Spencer H Terry, Daniel A Stone, Steven L Rasmussen, and Michael J Calidonna. Evaluation of the digital opacity compliance system in high mountain desert environments. *Journal of the Air & Waste Management Association*, 53(6):724–730, 2003.

[24] Michael J McFarland, Spencer H Terry, Michael J Calidonna, Daniel A Stone, Paul E Kerch, and Steven L Rasmussen. Measuring visual opacity using digital imaging technology. *Journal of the Air & Waste Management Association*, 54(3):296–306, 2004.

[25] Michael J Calidonna, Paul E Kerch, Daniel A Stone, Michael J McFarland, Tomas J Logan, John C Bosch Jr, and ENVIRONMENTAL SECURITY TECHNOLOGY CERTIFICATION PROGRAM ALEXANDRIA VA. An alternative to epa method 9: Field validation of a digital opacity compliance system (docs). 2004.

[26] Zhijian Yin, Boyang Wan, Feiniu Yuan, Xue Xia, and Jinting Shi. A deep normalization and convolutional neural network for image smoke detection. *Ieee Access*, 5:18429–18438, 2017.

[27] Yaocong Hu and Xiaobo Lu. Real-time video fire smoke detection by utilizing spatial-temporal convnet features. *Multimedia Tools and Applications*, 77(22):29283–29301, 2018.

[28] Gaohua Lin, Yongming Zhang, Gao Xu, and Qixing Zhang. Smoke detection on video sequences using 3d convolutional neural networks. *Fire Technology*, 55(5):1827–1847, 2019.

[29] Oscar D Pedrayes, Rubén Usamentiaga, and Daniel F García. Fully automated method to estimate opacity in stack and fugitive emissions: A case study in industrial environments. *Process Safety and Environmental Protection*, 170:479–490, 2023.

[30] Jun Mao, Change Zheng, Jiyan Yin, Ye Tian, and Wenbin Cui. Wildfire Smoke Classification Based on Synthetic Images and Pixel- and Feature-Level Domain Adaptation. *Sensors*, 21(23):7785, January 2021. Number: 23 Publisher: Multidisciplinary Digital Publishing Institute.

[31] Gao Xu, Yongming Zhang, Qixing Zhang, Gaohua Lin, and Jinjun Wang. Deep domain adaptation based video smoke detection using synthetic smoke images. *Fire Safety Journal*, 93:53–59, October 2017.

[32] Hang Yin, Yurong Wei, Hedan Liu, Shuangyin Liu, Chuanyun Liu, and Yacui Gao. Deep Convolutional Generative Adversarial Network and Convolutional Neural Network for Smoke Detection. *Complexity*, 2020:e6843869, November 2020. Publisher: Hindawi.

[33] Qi-xing Zhang, Gao-hua Lin, Yong-ming Zhang, Gao Xu, and Jin-jun Wang. Wildland Forest Fire Smoke Detection Based on Faster R-CNN using Synthetic Smoke Images. *Procedia Engineering*, 211:441–446, January 2018.

[34] Chao Xie and Huanjie Tao. Generating Realistic Smoke Images With Controllable Smoke Components. *IEEE Access*, 8:201418–201427, 2020. Conference Name: IEEE Access.

[35] Fida K Dankar and Mahmoud Ibrahim. Fake it till you make it: Guidelines for effective synthetic data generation. *Applied Sciences*, 11(5):2158, 2021. Publisher: MDPI.

[36] Ahmed Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Ma-*

chine Learning, volume 162 of *Proceedings of Machine Learning Research*, pages 290–306. PMLR, July 2022.

[37] Joshua Snoke, Gillian M Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(3):663–688, 2018. Publisher: Oxford University Press.

[38] Fida K Dankar, Mahmoud K Ibrahim, and Leila Ismail. A multi-dimensional evaluation of synthetic data generators. *IEEE Access*, 10:11147–11158, 2022. Publisher: IEEE.

[39] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[40] JoÅ¾e M RoÅ¾anec, Patrik Zajec, Spyros Theodoropoulos, Erik Koehorst, BlaÅ¾ Fortuna, and Dunja MladeniÄ‡. Synthetic data augmentation using GAN for improved automated visual inspection. *Ifac-Papersonline*, 56(2):11094–11099, 2023. Publisher: Elsevier.

[41] Oleksii Sidorov, Congcong Wang, and Faouzi Alaya Cheikh. Generative smoke removal. In *Machine Learning for Health Workshop*, pages 81–92. PMLR, 2020.

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[43] Keith Man and Javaan Chahl. A review of synthetic image data and its use in computer vision. *Journal of Imaging*, 8(11):310, 2022.

[44] Pablo Martinez-Gonzalez, Sergiu Oprea, Alberto Garcia-Garcia, Alvaro Jover-Alvarez, Sergio Orts-Escolano, and Jose Garcia-Rodriguez. Unrealrox: an extremely photorealistic virtual reality environment for robotics simulations and synthetic data generation. *Virtual Reality*, 24:271–288, 2020. Publisher: Springer.

[45] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 969–977, 2018.

[46] Ethan Seefried, Changsoo Jung, Jack Fitzgerald, Mariah Bradford, Trevor Chartier, and Nathaniel Blanchard. Balancing quality and quantity: The impact of synthetic data on smoke detection accuracy in computer vision. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2024.

[47] Kailai Zhou, Yibo Wang, Tao Lv, Yunqian Li, Linsen Chen, Qiu Shen, and Xun Cao. Explore spatio-temporal aggregation for insubstantial object detection: Benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3104–3115, June 2022.

[48] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[49] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[50] Guancheng Chen and Huabiao Qin. Class-discriminative focal loss for extreme imbalanced multiclass object detection towards autonomous driving. *The Visual Computer*, 38(3):1051–1063, 2022.

[51] Jie Chang, Xiaoci Zhang, Minquan Ye, Daobin Huang, Peipei Wang, and Chuanwen Yao. Brain tumor segmentation based on 3d unet with multi-class focal loss. In *2018 11th Interna-*

tional Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pages 1–5. IEEE, 2018.

[52] Ty Nguyen, Tolga Ozaslan, Ian D Miller, James Keller, Giuseppe Loianno, Camillo J Taylor, Daniel D Lee, Vijay Kumar, Joseph H Harwood, and Jennifer Wozencraft. U-net for mav-based penstock inspection: an investigation of focal loss in multi-class segmentation for corrosion identification. *arXiv preprint arXiv:1809.06576*, 2018.

[53] Xuezheng Jiang, Junyi Wang, Qinggang Meng, Mohamad Saada, and Haibin Cai. An adaptive multi-class imbalanced classification framework based on ensemble methods and deep network. *Neural Computing and Applications*, 35(15):11141–11159, 2023.

[54] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[55] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

[56] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[57] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024.

[58] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017.

[59] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[60] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[61] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[62] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[63] Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain vit baselines for imagenet-1k. *arXiv preprint arXiv:2205.01580*, 2022.

[64] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[65] Bjorn Barz and Joachim Denzler. Deep learning on small datasets without pre-training using cosine loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[66] Muzammil Khan, Muhammad Taqi Mehran, Zeeshan Ul Haq, Zahid Ullah, Salman Raza Naqvi, Mehreen Ihsan, and Haider Abbass. Applications of artificial intelligence in covid-19 pandemic: A comprehensive review. *Expert systems with applications*, 185:115695, 2021.

[67] Guangzhou An, Masahiro Akiba, Kazuko Omodaka, Toru Nakazawa, and Hideo Yokota. Hierarchical deep learning models using transfer learning for disease detection and classification based on small number of medical images. *Scientific reports*, 11(1):4250, 2021.

[68] Hidetoshi Matsuo, Mizuho Nishio, Tomonori Kanda, Yasuyuki Kojita, Atsushi K Kono, Masatoshi Hori, Masanori Teshima, Naoki Otsuki, Ken-ichi Nibu, and Takamichi Murakami. Diagnostic accuracy of deep-learning with anomaly detection for a small amount of imbalanced data: discriminating malignant parotid tumors in mri. *Scientific Reports*, 10(1):19388, 2020.

[69] Kirk Foster Karen Randolph. Visible emissions field manual epa methods 9 and 22. December 1997.

[70] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[71] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[72] Jack Fitzgerald, Ethan Seefried, James E Yost, Sangmi Pallickara, and Nathaniel Blanchard. Paying attention to wildfire: Using u-net with attention blocks on multimodal data for next day prediction. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 470–480, 2023.

[73] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *CoRR*, abs/2201.05989, 2022.

[74] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020.

[75] Derek S. Prijatelj, Mel McCurrie, Samuel E. Anthony, and Walter J. Scheirer. A bayesian evaluation framework for subjectively annotated visual recognition tasks. *Pattern Recognition*, 123:108395, 2022.

[76] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[77] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10042–10051, 2021.

[78] Changsoo Jung, Nathaniel Blanchard, Ross Beveridge, Benjamin Clegg, et al. Egoroom: egocentric 3d pose estimation through multi-coordinates heatmaps. 2021.

[79] Ethan Seefried, Mariah Bradford, Swagatalaxmi Aich, Caspian Siebert, Nikhil Krishnaswamy, and Nathaniel Blanchard. Learning foreign language vocabulary through task-based virtual reality immersion. In *International Conference on Human-Computer Interaction*, pages 203–213. Springer, 2024.

[80] Jeremiah W Johnson. Benefits and pitfalls of jupyter notebooks in the classroom. In *Proceedings of the 21st annual conference on information technology education*, pages 32–37, 2020.

[81] Abraham Montoya Obeso, Jenny Benois-Pineau, Alejandro Álvaro Ramirez Acosta, and Mireya Saraí García Vázquez. Architectural style classification of mexican historical buildings using deep convolutional neural networks and sparse features. *Journal of Electronic Imaging*, 26(1):011016–011016, 2017.

[82] Gayane Shalunts, Yll Haxhimusa, and Robert Sablatnig. Architectural style classification of building facade windows. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Song Wang, Kim Kyungnam, Bedrich Benes, Kenneth Moreland, Christoph Borst, Stephen

DiVerdi, Chiang Yi-Jen, and Jiang Ming, editors, *Advances in Visual Computing*, pages 280–289, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[83] H. Shao, T. Svoboda, and L. Van Gool. ZuBuD — Zürich buildings database for image based recognition. Technical Report 260, Computer Vision Laboratory, Swiss Federal Institute of Technology, March 2003.

[84] Salma Taoufiq, Balázs Nagy, and Csaba Benedek. Hierarchynet: Hierarchical cnn-based urban building classification. *Remote Sensing*, 12(22):3794, 2020.

[85] Zhe Xu, Dacheng Tao, Ya Zhang, Junjie Wu, and Ah Chung Tsoi. Architectural style classification using multinomial latent logistic regression. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 600–615, Cham, 2014. Springer International Publishing.

[86] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[87] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven. Tour the world: Building a web-scale landmark recognition engine. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1085–1092, 2009.

[88] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[89] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[90] Björn Barz and Joachim Denzler. Wikichurches: A fine-grained dataset of architectural styles with real-world challenges. *CoRR*, abs/2108.06959, 2021.

[91] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[92] Kilem L Gwet. Intrarater reliability. *Wiley encyclopedia of clinical trials*, 4, 2008.

[93] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.