THESIS

IMAGES IN MOTION?: A FIRST LOOK INTO VIDEO LEAKAGE IN FEDERATED LEARNING

Submitted by

Md Fazle Rasul

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2025

Master's Committee:

Advisor: Indrakshi Ray

Anura P. Jayasumana
Bruhadeshwar Bezawada
Steve Simske

ABSTRACT

IMAGES IN MOTION?: A FIRST LOOK INTO VIDEO LEAKAGE IN FEDERATED LEARNING

Federated learning (FL) allows multiple entities to train a shared model collaboratively. Its core, privacy-preserving principle is that participants only exchange model updates, such as gradients, and never their raw, sensitive data. This approach is fundamental for applications in domains where privacy and confidentiality are important. However, the security of this very mechanism is threatened by gradient inversion attacks, which can reverse-engineer private training data directly from the shared gradients, defeating the purpose of FL. While the impact of these attacks is known for image, text, and tabular data, their effect on video data remains an unexamined area of research. This paper presents the first analysis of video data leakage in FL via gradient inversion attacks. We evaluate two common video classification approaches: one employing pre-trained feature extractors and another that processes raw video frames with simple transformations. Our results indicate that the use of feature extractors offers greater resilience against gradient inversion attacks.

We also demonstrate that image super-resolution techniques can enhance the frames, extracted through gradient inversion attacks, enabling attackers to reconstruct higher-quality videos. Our experiments validate this across scenarios where the attacker has access to zero, one, or more reference frames from the target environment. We find that although feature extractors make attacks more challenging, leakage is still possible if the classifier lacks sufficient complexity. We, therefore, conclude that video data leakage in FL is a viable threat and the conditions under which it occurs warrant further investigation.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Indrakshi Ray, for her invaluable advice, support, and encouragement throughout my Master's program. Her expertise and insights were instrumental to the success of this work. I am also deeply grateful to Dr. Bruhadeshwar Bezawada for his continuous guidance.

I extend my thanks to Dr. Indrajit Ray for his valuable contributions during our weekly Cybersecurity Center Group (DBSec Group) meetings. I am also grateful to Dr. Anura P. Jayasumana and Dr. Steve Simske for their time and dedication as members of my master's advisory committee.

I wish to thank all my teachers for their motivation, as well as my friends and colleagues in the DBSec Group for their unwavering support, which was essential in achieving this goal.

Finally, I am grateful to the funding agencies for their generous support, which made this work possible. This work was supported in part by NIST funding under Award Number 60NANB23D152 and NSF under Award Numbers CNS 2335687, DMS 2123761, CNS 1822118, NIST, ARL, Statnett, AMI, NewPush and Cyber Risk Research, and by the Cybersecurity Center funded by the State of Colorado.

DEDICATION

I dedicate this work to my mother, for her love and belief in me, and to my father, for his unwavering support. To my wonderful wife, Mim, your patience and partnership were my anchor on this journey. To my dear sister, Nimu, thank you for being my constant cheerleader. This achievement is a reflection of the love and strength you have all given me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The advancement of machine learning is driven by the availability of large, diverse datasets. However, the use of sensitive data is fundamentally at odds with the critical need to protect individual privacy and comply with stringent regulations. This tension necessitates innovative approaches that allow for collaborative model training without centralizing sensitive information.

1.1 Motivation

The application of machine learning to video analysis offers transformative potential across numerous sensitive fields. In healthcare, for instance, analyzing surgical recordings can lead to improved training, automated skill assessment, and new diagnostic insights. Similarly, law enforcement agencies rely on surveillance footage for forensic analysis and public safety, while the insurance industry uses video from accident documentation to streamline claims processing and assess risk. The effectiveness of the machine learning models in these domains is highly dependent on the quality and diversity of the training data.

A significant challenge, however, is that the data held by any single institution is often limited in scope and diversity. A model trained exclusively on one hospital's surgical videos may be biased towards that hospital's specific patient demographics or surgical techniques, and thus fail to generalize well to other settings. To build a truly robust and unbiased model, it is necessary to learn from the varied data of multiple organizations. This need for collaborative training, however, runs directly into a critical barrier: privacy.

Directly sharing raw, sensitive data for collaborative analysis is often legally impossible. Stringent data protection regulations, such as the General Data Protection Regulation (GDPR) in Europe, the California Consumer Privacy Act (CCPA), and the Australian Privacy Principles (APP), are designed to protect individuals' private information and impose severe penalties for violations. This legal and ethical landscape makes Federated Learning (FL) an attractive paradigm.

Federated Learning is a decentralized machine learning approach where multiple parties collaborate to build a shared global model without exchanging their local, private data. In the FL process, each organization trains a local model on its own dataset. Instead of sharing the data itself, they share the resulting model updates, known as gradients. These gradients mathematically represent the training errors and learnings from the local data and are used by a central server to adjust and improve the weights of the global model [21, 27, 24]. Because the raw data never leaves its owner's secure environment, FL was initially regarded as a fundamentally privacy-preserving solution.

However, this foundational assumption of privacy has been challenged by several recent and significant attack methodologies. A growing body of research has demonstrated that models trained in a federated manner are susceptible to training data leakage. These studies show that the shared gradients, once thought to be secure and abstract representations, can be exploited to reveal sensitive information about the private data used to compute them. This vulnerability points to a fundamental privacy risk, particularly in high-stakes applications where confidentiality is a primary concern.

This work focuses on a particularly severe form of this vulnerability known as "deep leakage" from video data. Deep leakage is a process that enables an adversary to achieve a full, high-fidelity reconstruction of the original raw video samples directly from the model gradients. This stands in stark contrast to "shallow leakage" [40, 13], which is less severe as it typically only reveals general data properties, such as confirming whether a specific individual's data was part of the training set (membership information) or exposing statistical details about the dataset as a whole. The ability to perfectly reconstruct a frame from a sensitive surgical or surveillance video represents a catastrophic privacy breach, thereby motivating the investigation presented in this thesis.

1.2 Research Gap

While significant research has explored privacy vulnerabilities in machine learning, the existing literature has largely concentrated on leakage attacks against image classification systems [40, 35,

38, 12, 23, 9, 10, 39] and other data formats like text and tabular data [28, 19]. Consequently, the unique security challenges posed by video data have remained comparatively under-explored. Our work addresses this gap by focusing on practical video analysis systems, such as video captioning [31] and modern collaborative intruder detection systems that rely on security camera feeds [2]. The findings from image-based attacks cannot be directly extrapolated to video, as video data introduces distinct complexities rooted in its temporal dimension.

This time-dependent structure presents two fundamental challenges. First, it leads to a dramatic increase in data volume; a short video clip is composed of hundreds or thousands of individual frames, each a high-dimensional image. Second, it creates significant inter-frame redundancy, where consecutive frames contain a large amount of overlapping visual information. While this redundancy is useful for understanding motion and context, it also represents a potential security vulnerability. An adversary might exploit this overlap to reconstruct a high-fidelity video sequence even if they only successfully leak information corresponding to a small fraction of the total frames. To manage these challenges, training on video data typically follows one of two main strategies. A simplistic approach is to treat a video as an unordered collection of frames, effectively reducing the problem to image classification and discarding the vital temporal information. However, a more common and sophisticated approach acknowledges the sequential nature of video. In this pipeline, a video is first segmented into short clips. Then, pre-trained convolutional neural networks (e.g., ResNet, AlexNet) are used as powerful feature extractors to generate feature vectors for each frame within a clip. These individual frame-level features are subsequently aggregated—often by averaging—to create a single, compact video-level representation. It is this aggregated feature vector that is used for the final classification task [31, 2].

This aggregation step marks a key distinction from standard image processing. In image classification, even when using feature extractors, each image typically retains its own distinct feature matrix. In contrast, the video analysis pipeline condenses the features from many frames into a single, abstract representation. This distinction leads to the central question of our research. Most existing studies on gradient inversion attacks that successfully reconstruct images assume that the

gradients are calculated directly from the raw input pixels. Our work investigates a more complex and practical scenario where the classifier is trained not on raw data, but on these aggregated, intermediate features. We seek to determine whether this type of training on features rather than on the data itself—inherently mitigates the risk of gradient inversion attacks. By doing so, this work addresses a critical gap in security analysis for real-world video processing systems.

1.3 Our Framework

Our research began by establishing a baseline for gradient inversion attacks against video data within a federated learning context. We first investigated whether existing, well-documented attacks could be extended to reconstruct meaningful information from complex video streams. To accomplish this, we adapted the Deep Leakage from Gradients (DLG) algorithm [40], a prominent method originally designed for image reconstruction. This investigation was conducted under two distinct and realistic scenarios: one where the model's gradients are computed directly from raw video frames, and another where gradients are computed from abstract features generated by a pre-trained feature extractor. Initial experiments revealed a significant limitation: the DLG attack, being optimized for static images of a specific resolution, yielded low-resolution and visually degraded video outputs when applied to video frames.

To address this, we proposed a novel enhancement to the attack pipeline by incorporating image and video super-resolution algorithms [29, 34]. This post-processing step was designed to substantially improve the visual fidelity of the reconstructed frames, making the leaked information more intelligible. Furthermore, to simulate more sophisticated threat models, we explored scenarios where an attacker might possess some prior knowledge. We conducted attacks assuming the adversary has access to zero, one, or multiple reference frames from the target environment, using this information to aid in the reconstruction of adjacent video frames. Finally, to isolate the specific impact of feature extractors, we performed a controlled analysis on an image dataset where the effectiveness of these attacks is already well-demonstrated, allowing us to systematically measure how feature extraction influences attack outcomes.

1.4 Problem Statement

We consider that n parties are participating in the collaborative learning algorithm; $\{P_1, P_2, \ldots, P_n\}$ and, are respectively, in possession of n distinct sets of data: $\{D_1, D_2, \ldots, D_n\}$. Each D_i, $\forall i \leq n$, consists of M video data frames of certain dimensions and each video frame is denoted by V_j^i where $j \leq M$. We denote the local machine learning model for P_i as C_i and the number of training epochs per data sample by t. Let $\delta^k(V_j^i)$, where $k \leq t$, denote the gradients generated by training C_i on V_j^i in the k^{th} epoch. At the end of each training epoch, these gradients are shared by each participant with the other participants.

The problem is thus: Given $\delta^k(V_j^i)$ where $i \neq x$ and $\forall k \leq t$, a participant P_x needs an algorithm to reconstruct V_j^i without any additional information from P_i. Additionally, we assume that P_x is a semi-honest participant, , P_x follows the rules of the collaborative protocol but will attempt to learn the private training data of the other participants.

1.5 Core Research Questions

Our work was guided by a set of core research questions aimed at closing a significant gap in the understanding of privacy vulnerabilities in video-based federated learning.

- Feasibility of Video Reconstruction: To what extent can existing gradient inversion attacks, which were originally designed for static images, be successfully adapted to reconstruct meaningful visual information from complex video data?
- Enhancing Attack Potency: Can the visual quality and practical utility of video frames reconstructed from gradients be significantly improved by applying super-resolution techniques as a post-processing step?
- Impact of Prior Knowledge: How does an attacker's access to auxiliary information, such as one or more authentic reference frames from the target's environment, affect the fidelity and success rate of the video reconstruction attack?

- The Role of Feature Extractors as a Defense: Does the common practice of training a classifier on abstract features, rather than directly on raw video frames, function as an inherent defense against gradient inversion? Furthermore, how does the choice of the feature extractor architecture itself influence the degree of data leakage?

The rest of the thesis unfolds as follows. We first establish the context in Section 2 by reviewing the landscape of existing gradient inversion attacks and defenses. Section 3 details our proposed approaches. Section 4 rigorously evaluates our methods, presenting comprehensive experimental results. Section 5 explores pertinent open challenges and potential future directions. Finally, Section 6 concludes the paper, summarizing our findings.

Chapter 2

Background

This chapter provides an overview of the primary methods used in gradient-based privacy attacks and the corresponding defensive strategies developed to counter them. We will explore the progression of attack techniques and then examine the strengths and limitations of various protective measures.

2.1 Gradient Leakage Attacks

The vulnerability of gradients was starkly exposed in seminal works that formulated data reconstruction as an optimization problem. The DLG paper was a pioneering effort [40], demonstrating that by initializing a dummy input and label, an attacker could iteratively optimize them to produce a gradient that closely matches the target gradient, thus recovering the original data and label pixel-for-pixel. However, this foundational attack was often slow to converge. A subsequent work, iDLG [38] enhanced this process by showing that ground-truth labels could be extracted analytically from the gradients of a model's final fully-connected layer, which significantly accelerated the image reconstruction process by removing the need to optimize the label.

Concurrently, researchers in [12] framed the attack as minimizing the cosine distance between gradients rather than the L2 distance, which proved more robust to variations in gradient magnitude. This work also introduced total variation as a regularization term to improve the visual quality of reconstructed images, demonstrating successful recovery even from batches of images. Further solidifying these findings, the the GradInversion attack demonstrated robust image batch recovery and introduced an effective method for recovering labels from a batch, even when multiple classes are present [35]. These early works collectively established that gradients, far from being secure, contain substantial, extractable information about the training data. The core principle of these attacks was further analyzed in [6] which provided a layer-wise theoretical understanding of how architectural properties contribute to data leakage.

Building on these foundations, subsequent research aimed to improve the potency, efficiency, and generality of gradient inversion attacks. One line of work introduced powerful priors to guide the reconstruction process. Jeon et al. showed that employing a pre-trained generative model could produce significantly more realistic and faithful reconstructions, as the optimization is performed in the generator's latent space rather than the high-dimensional pixel space [18]. This marked a major leap in the quality of recovered data, moving from noisy or blurry images to highly plausible ones.

Other research focused on different facets of the attack. For instance, Zhu et al. proposed a novel analytical approach that recursively solves for layer inputs in a closed-form manner, offering a much faster alternative to iterative optimization and providing insights into which network architectures are inherently more vulnerable due to rank deficiency [39]. The issue of applicability was tackled by demonstrating that these attacks are not limited to standard computer vision tasks but also pose a severe threat to sensitive domains, as shown in [23], which successfully reconstructed multi-label medical images.

As the FL field matured, so did the attacks designed to break it. Recognizing that defenses like dropout were being considered, the work [25] demonstrated that while standard inversion attacks fail against dropout's stochasticity, a modified attack that jointly optimizes for both the private data and the dropout mask can successfully bypass this defense. Similarly, to address the common use of gradient compression to save bandwidth, researchers developed an improved attack in [9] showing that even heavily compressed or pruned gradients could be vulnerable.

The scope of attacks also broadened beyond the standard horizontal FL setting. Jin et al. revealed that vertical federated learning (VFL), where clients hold different features of the same data samples, is particularly vulnerable [19]. In VFL, since the server often knows the indices of data samples used in a batch, this side-channel information can be exploited to drastically improve reconstruction, even for very large batches. Furthermore, the problem of attacking aggregated gradients from multiple users was explored in [22] which showed that by observing aggregated updates over several rounds and leveraging side-channel information about user participation, an

attacker could potentially disentangle individual user updates, thereby enabling traditional gradient inversion on the disaggregated gradients.

Newer advancements in federated learning also have to contend with the dual-use nature of technologies like steganography. Traditionally, steganography is the art of concealing data within non-secret files, such as embedding a text file within an image. In the context of federated learning, this technique can be used defensively to embed watermarks or other metadata within model updates to trace the source of a data leak or to verify the integrity of the models [7]. However, this same principle can be weaponized. Malicious actors could employ steganography to create a covert channel for data exfiltration, hiding sensitive information within the gradients of seemingly benign model updates [8]. For instance, a compromised client could encode fragments of private data into the least significant bits of the video frames used for training, which would then be subtly embedded in the resulting gradient update. To a central server, this would appear as normal model training, but to a colluding party, it represents a secret communication channel for leaking private data that is difficult to detect.

Finally, frameworks have been proposed to formalize and evaluate these privacy risks. Balunović et al. established a theoretical basis for understanding the problem by defining a Bayes optimal adversary, allowing existing attacks to be interpreted as approximations of this theoretical ideal [4]. Other works, such as [10]," and evaluation frameworks like those in [15] provide a comprehensive overview and structured analysis of the attack landscape, categorizing existing methods and critically assessing their effectiveness under practical conditions. Together, this body of literature confirms that gradient inversion is a persistent and evolving threat that challenges the core privacy promises of federated learning.

2.2 Defensive Mechanisms

The defense strategies can be broadly categorized into three main categories: perturbing or modifying gradients, obfuscating data before training, and employing cryptographic techniques to protect gradient information.

A foundational defense strategy involves perturbing the gradients before they are shared. The most formally recognized approach in this category is Differential Privacy, which adds precisely calibrated noise to the gradients. As detailed in [1], this method provides a provable guarantee that the inclusion of any single data point in the training set has a limited and quantifiable impact on the output, thereby making accurate data reconstruction from the noisy gradients computationally infeasible. Other perturbation techniques focus on information reduction. Revisiting Gradient Pruning explores sparsifying gradients by setting a significant fraction of the values with the smallest magnitudes to zero. This intentional removal of information degrades the quality of the gradient signal available to an attacker [32]. A more advanced gradient manipulation technique is proposed in [37], which defends against attacks by projecting the true gradient onto a randomly sampled orthogonal subspace. This censoring process is designed to remove sufficient information to thwart reconstruction while preserving enough to maintain model utility.

Another class of defenses operates on the data itself, before any gradients are computed. The core idea is to make the relationship between the shared gradient and any single, original data sample ambiguous. The mixup framework, introduced in [36], achieves this by training the model on convex combinations of pairs of data samples and their labels. The resulting gradients correspond to these virtual, mixed samples, not the original ones, confounding reconstruction attempts. Taking this concept further, InstaHide proposes a more complex data-hiding scheme where each private training image is mixed with several images from a public dataset using a unique, invertible mixing matrix [16]. This defense places a significant computational burden on the attacker, who would need to solve a difficult disentanglement problem to recover the original private sample.

Finally, cryptographic methods offer powerful protections by preventing the server from ever observing individual client gradients. Secure Aggregation protocols, outlined in works like [5], ensure that the central server only receives the sum of gradients from a cohort of users, not the individual contributions, making it impossible to isolate and attack a single user's update. Recognizing the potential communication and computation overhead of such methods, subsequent research such as FastSecAgg [20] and LightSecAgg [33] has focused on developing more efficient

and scalable secure aggregation protocols. An alternative cryptographic approach, discussed in Privacy-Preserving Deep Learning via additively homomorphic encryption [3], uses encryption schemes that allow the server to compute the sum of encrypted gradients without decrypting them, achieving a similar privacy guarantee. Beyond these categories, novel architectural changes are also being explored. For instance, in [14] the authors propose a two-phase training process that aims to break the invertible relationship between gradients and private data by creating a more complex and indirect mapping. Together, these diverse strategies highlight critical ongoing efforts to balance model performance with the robust privacy preservation required for secure collaborative machine learning.

Chapter 3

Methodology

This study undertakes a systematic examination of potential data leakage pathways inherent in video-based machine learning models. Our research was structured around two primary objectives: first, to explore how different data handling and preprocessing methodologies affect privacy, and second, to conduct a detailed analysis of any leaked information to determine its practical value to a potential adversary. This included assessing the quality of reconstructed data and investigating techniques, such as super-resolution, that an attacker might use to enhance it.

3.1 Experimental Design and Pipeline

The core of our investigation centered on the video processing pipeline illustrated in Figure 3.1, which was designed to analyze the impact of neural network preprocessing on the retrievability of video frames from model gradients. The experimental pipeline features two distinct branches to simulate different real-world scenarios:

- **Pipeline with Neural Network Preprocessing:** In this configuration, video data is first processed by a neural network to extract abstract features before being fed to the main classifier. Our findings indicate that this preprocessing step can add more difficulty to gradient inversion attacks. Specifically, when a moderately complex classifier was used, the intermediate feature representations became sufficiently abstract, making it computationally infeasible to reverse the process and reconstruct the original video frames. The transformation effectively obscured the necessary information, breaking the link between the gradients and the source data.

- **Pipeline without Preprocessing:** In this scenario, the model is trained on gradients derived more directly from the raw video data. As a result, this pathway proved vulnerable to leakage, allowing for the extraction of low-resolution versions of the original frames. These

recovered frames, while degraded, still contain sensitive information. We then explored how an adversary could employ super-resolution techniques to significantly enhance the clarity and detail of these leaked frames, thereby increasing their utility.

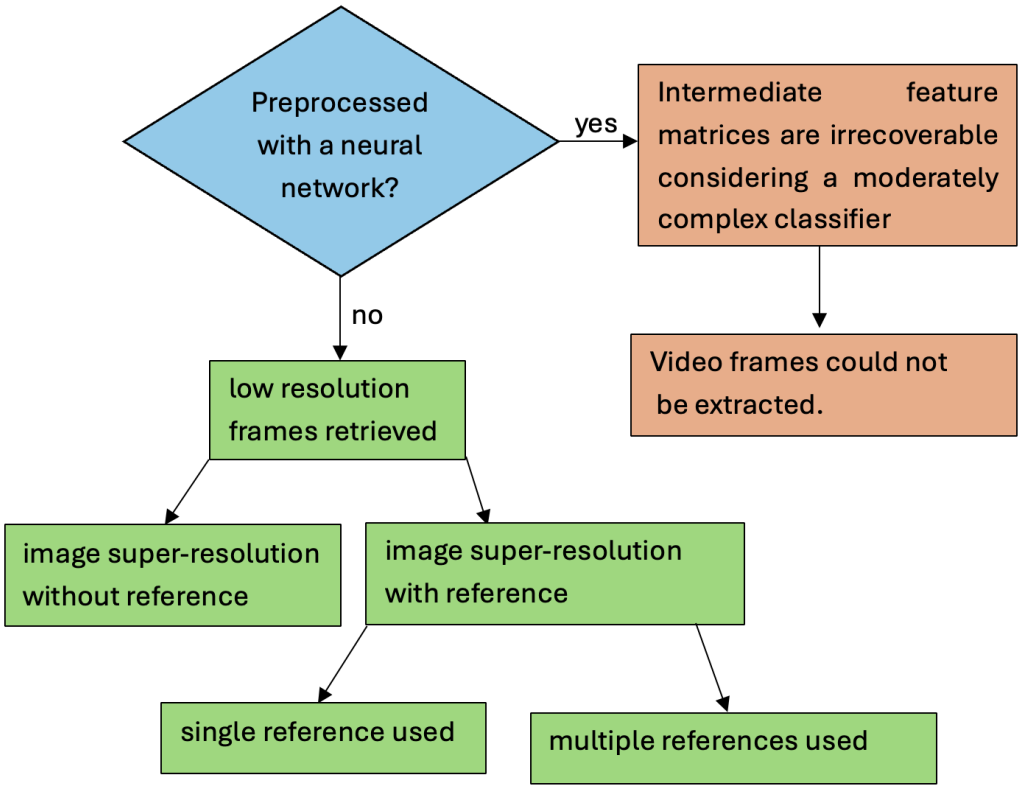


Figure 3.1: Video Processing Pipeline

To provide a comprehensive security analysis, our simulations evaluated the system's resilience against adversaries with varying levels of capability. We defined and tested three distinct threat scenarios:

- **No-Frame Access:** In this scenario the attacker does not have access to any video frame.
- **Single-Frame Access:** A scenario testing the impact of a partial leak, where an attacker gains access to only one frame from the video used for training.

- **Multiple-Frame Access:** The most severe scenario, where an attacker has access to multiple frames.

By analyzing the system's performance across these tiered levels of compromise, this research provides nuanced insights into how the threat landscape changes based on an attacker's resources and success, offering a more complete picture of the vulnerabilities in video-based federated learning.

3.2 Gradient Inversion on Raw Frames with Simple Transformation

To establish a foundational understanding of data leakage in video-centric federated learning, our initial experiments focused on a simplified scenario. In this phase, we investigated the feasibility of video reconstruction when the data undergoes only basic preprocessing transformations—such as resizing, cropping, and normalization—without the added complexity of neural network feature extraction. This approach allowed us to create a controlled baseline to isolate and measure the effectiveness of the gradient inversion attack itself. While prior research has already confirmed that gradient attacks can successfully reconstruct images across various batch sizes and dimensions, our research specifically investigated methods for improving the fidelity of extracted video frames.

Our reconstruction pipeline, illustrated in Figure 3.2, consists of three primary stages: data preprocessing, gradient-based reconstruction, and quality enhancement through super-resolution.

1. **Preprocessing and Federated Learning Simulation:** The first step involves significant downsampling of the input video frames. Specifically, each frame's resolution is reduced from its original 240x320 pixels to a much smaller 32x32 pixels, and processed individually with a batch size of one. This dimensional reduction is a critical prerequisite, as it aligns with the operational constraints of the Deep Leakage from Gradients (DLG) attack, which is computationally intensive and performs more effectively on smaller data inputs. The use of

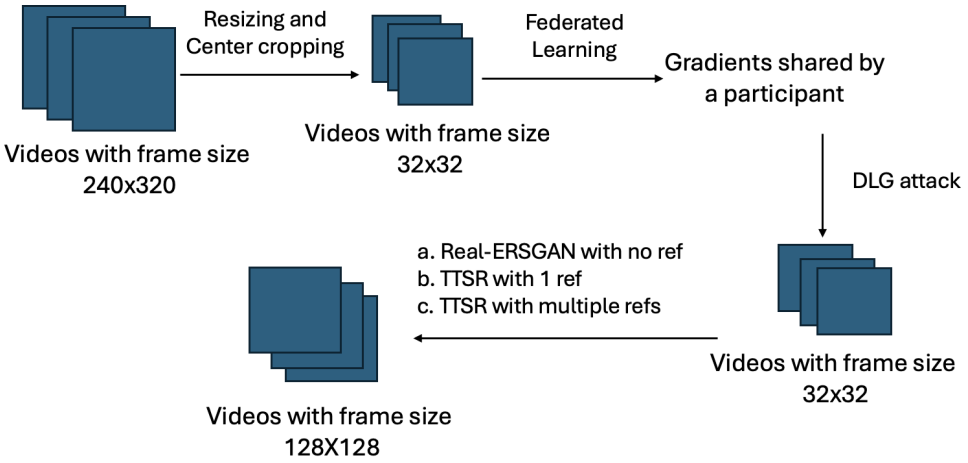


Figure 3.2: Video Reconstruction Pipeline Without Neural Network Pre-processing.

a single-item batch also represents a worst-case privacy scenario, where the shared gradient is directly correlated with a single data sample, maximizing potential information leakage. These low-resolution frames are then processed within a simulated federated learning environment, from which we extract the resulting gradients for the attack.

2. **Gradient Inversion via DLG:** Using only the shared gradient information from the previous step, we employ the DLG attack to reverse the process and reconstruct the video frames. This attack iteratively optimizes a "dummy" input until its resulting gradient matches the target gradient, thereby revealing the original private data. The output of this stage is a sequence of low-resolution (32x32) video frames that are a direct reconstruction from the gradient data.

3. **Fidelity Enhancement with Super-Resolution:** The final and most crucial stage focuses on enhancing the quality of the recovered low-resolution frames. To accomplish this, we utilize advanced super-resolution techniques to upscale the frames to a higher quality. Recognizing that an attacker's capabilities can vary, we implemented two distinct approaches tailored to different threat models:

    - **Zero-Knowledge Scenario (Real-ESRGAN):** For a scenario where an attacker has no prior access to any of the original video data, we employ Real-ESRGAN [29]. This is

a "blind" super-resolution model capable of enhancing image quality without needing a reference image, making it suitable for a general-purpose attack. This method was used to enhance the frame resolution to 128x128 pixels without utilizing any external reference frames. It improves blind super-resolution for real-world images through advanced degradation modeling and a U-Net discriminator, achieving superior restoration. For our experiments, we employed the open-source implementation available at the GitHub repository https://github.com/xinntao/Real-ESRGAN-ncnn-vulkan.

- **Partial-Knowledge Scenario (TTSR):** For a more sophisticated attacker who might possess one or more original frames (a plausible scenario in video analysis), we utilize the Texture Transformer Network for Image Super-Resolution (TTSR) [34]. This model can leverage the texture information from a known reference frame to significantly improve the reconstruction quality of the leaked frames, making it particularly effective for video data where frames often share similar textures. This approach utilizes a single reference frame to guide the super-resolution process, upscaling the frames to 128x128 pixels. It leverages attention mechanisms for effective texture transfer from reference images. We used the open-source TTSR implementation available on GitHub (https://github.com/researchmm/TTSR). The TTSR approach is also extended by leveraging multiple reference frames to refine the super-resolution reconstruction, also resulting in 128x128 pixel frames.

3.3 Gradient Inversion on Extracted Video Features

To assess the practical impact of feature extraction on privacy, our research investigated the potential for video reconstruction within a state-of-the-art video analysis framework. We selected the Collaborative Learning of Anomalies with Privacy (CLAP) framework [2] as a representative use case for this analysis. CLAP is a particularly relevant choice as it embodies a modern, privacy-preserving paradigm for unsupervised video anomaly detection, designed specifically for sensitive contexts like collaborative surveillance where raw data cannot be shared. Using the official open-

source implementation of CLAP (https://github.com/AnasEmad11/CLAP), we aimed to determine if its complex pre-processing pipeline provided inherent protection against gradient-based attacks. Our methodology involved creating a systematic procedure that adapts and extends the principles of the DLG attack to this more complex scenario. The objective was to intercept the gradients during the collaborative training process and reverse-engineer them to reconstruct the original video frames. However, our attempts to reconstruct the video data were unsuccessful.

The failure of the attack can be attributed to the restrictive effects of the extensive video pre-processing inherent to the CLAP architecture. Unlike simpler models where gradients are computed directly from input data, CLAP first processes videos through a feature extractor and then aggregates these features. This aggregation step, which condenses information from multiple frames into a single, abstract representation, proved to be a critical defense mechanism. Consequently, the gradients are calculated based on this highly processed, abstract data, not the original pixels. This abstraction effectively severs the direct informational link back to the source frames, making it impossible to retrieve the distinct intermediate feature matrices required for the inversion attack to succeed and ultimately blocking any reconstruction of the video.

3.4 Effect of Feature Extractors on Gradient Inversion Attacks

To empirically evaluate how feature extractors affect the vulnerability of a system to gradient inversion attacks, our study implemented three prominent attack methodologies: Deep Leakage from Gradients (DLG), improved Deep Leakage from Gradients (iDLG), and Recursion-based Gradient Attack on Privacy (R-GAP). These methods were selected to represent two distinct philosophical approaches to data reconstruction: iterative optimization and analytical recursion.

The first category includes the iterative recovery attacks, DLG and iDLG. The DLG attack conceptualizes data reconstruction as an optimization challenge. The **DLG** attack frames data reconstruction as an iterative optimization problem that begins with a random guess, $(\mathbf{x}', \mathbf{y}')$. The random data is then repeatedly updated using gradient descent to minimize the distance between

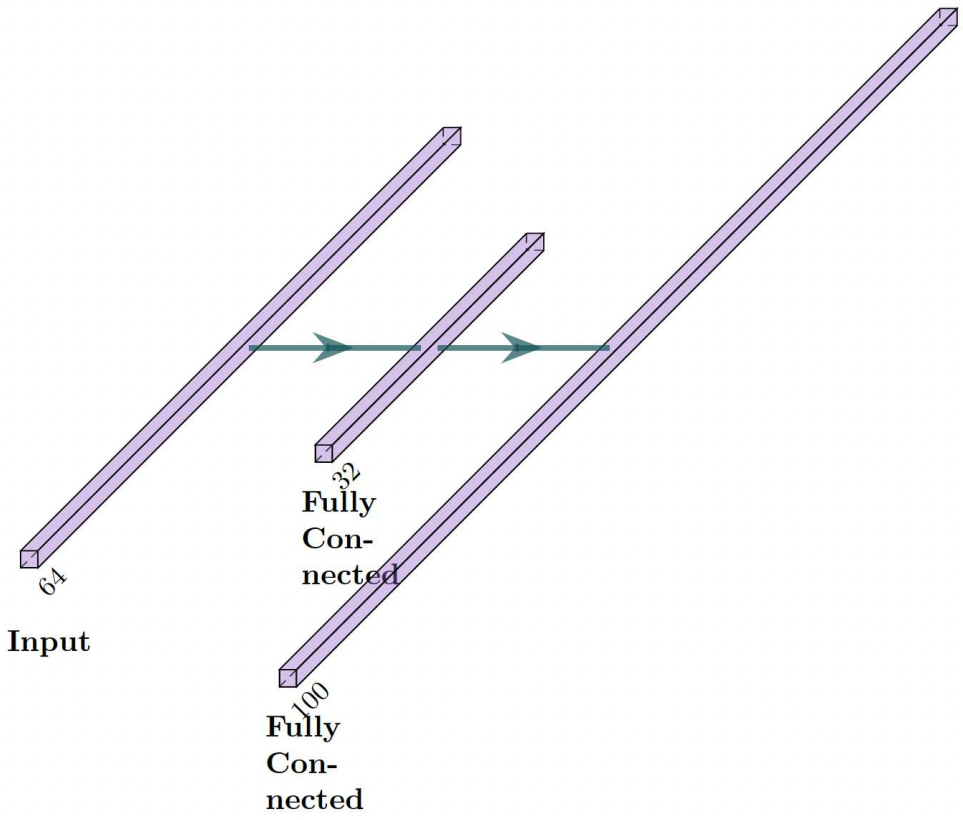


Figure 3.3: Architectures of the simple classifier .

its gradients and the victim's true gradients, eventually converging on the private data.

$$\begin{aligned} \mathbf{x}'^*, \mathbf{y}'^* &= \arg\min_{\mathbf{x}',\mathbf{y}'} \left\| \nabla W' - \nabla W \right\|^2 \\ &= \arg\min_{\mathbf{x}',\mathbf{y}'} \left\| \frac{\partial \mathcal{L}(\mathcal{F}(\mathbf{x}', W), \mathbf{y}')}{\partial W} - \nabla W \right\|^2 \end{aligned} \tag{3.1}$$

The core of the attack is formulated as the optimization problem shown in Eq. 3.1. The objective is to find the optimal reconstructed data $\mathbf{x}'^*$ and corresponding labels $\mathbf{y}'^*$ that best reproduce the leaked information. This is achieved by searching for the arguments $(\mathbf{x}', \mathbf{y}')$ that minimize the squared euclidean distance between two vectors: the ground-truth gradient leaked from the victim, ∇W, and a synthetically generated gradient, $\nabla W'$. The minimization is performed over the set of dummy inputs $\mathbf{x}'$ and $\mathbf{y}'$, which are initialized randomly and iteratively updated. The synthetic

gradient $\nabla W'$ is calculated with respect to the model's weights W by using the model's forward pass function, denoted $\mathcal{F}(\cdot)$, and its loss function, $\mathcal{L}(\cdot)$, evaluated on these dummy inputs.

A critical insight of the **iDLG** attack is the ground-truth label, c can be accurately recovered from the shared gradients by applying the heuristic shown in Equation 3.2. This rule identifies the correct class index, i by exploiting the property that its final-layer weight gradient, $\nabla \mathbf{W}_L^i$, is geometrically distinct, having a non-positive dot product with the gradients of all other incorrect classes

$$c = i, \quad \text{s.t.} \quad \nabla \mathbf{W}_L^{i\,T} \cdot \nabla \mathbf{W}_L^j \leq 0, \quad \forall j \neq i \tag{3.2}$$

So, iDLG attack optimizes only the data and not the label makeing the attack lot faster than DLG.

On the other hand, recursion-based attacks analytically recover data by recursively working backward through the network from the final layers to the input. **R-GAP** exploits the direct mathematical relationships between a layer's inputs, weights, and gradients, they solve for the input of each preceding layer until the original data is revealed [39]. This system directly relates a layer's input x_i to its weights W_i and feature map Z_i, allowing for its reconstruction by solving equation 3.3. The gradients of the loss with respect to the weights and the feature map are given by ∇W_i and ∇Z_i, respectively.

$$\begin{aligned} W_i \cdot \mathbf{x}_i &= Z_i \\ \nabla Z_i \cdot \mathbf{x}_i &= \nabla W_i \end{aligned} \tag{3.3}$$

For our analysis, we selected an image dataset. This choice was deliberate, as the selected attacks have been demonstrably successful in prior literature at reconstructing high-fidelity images, providing a reliable baseline for our investigation. Our experiments were structured in a two-stage process to isolate the impact of the feature extractor.

- **Baseline Analysis:** In the first stage, we applied the attacks directly to gradients generated from raw input images. This was conducted using both a simple and a moderately complex classifier architecture, with the primary goal of reconstructing the original image. This stage

served to validate our implementation of the attacks and establish a performance benchmark in a standard setting.

- **Feature Vector Analysis:** In the second stage, we investigated the core research question. Here, the attacks were directed at gradients derived from feature vectors, which had been generated by pre-trained feature extractors. The objective was to determine the feasibility of reconstructing these intermediate feature vectors themselves, thereby assessing whether the feature extraction process provides a layer of security against gradient inversion.

The specific architectures for the simple and moderately complex classifiers employed throughout these experiments are detailed in Figure 3.3 and 3.4

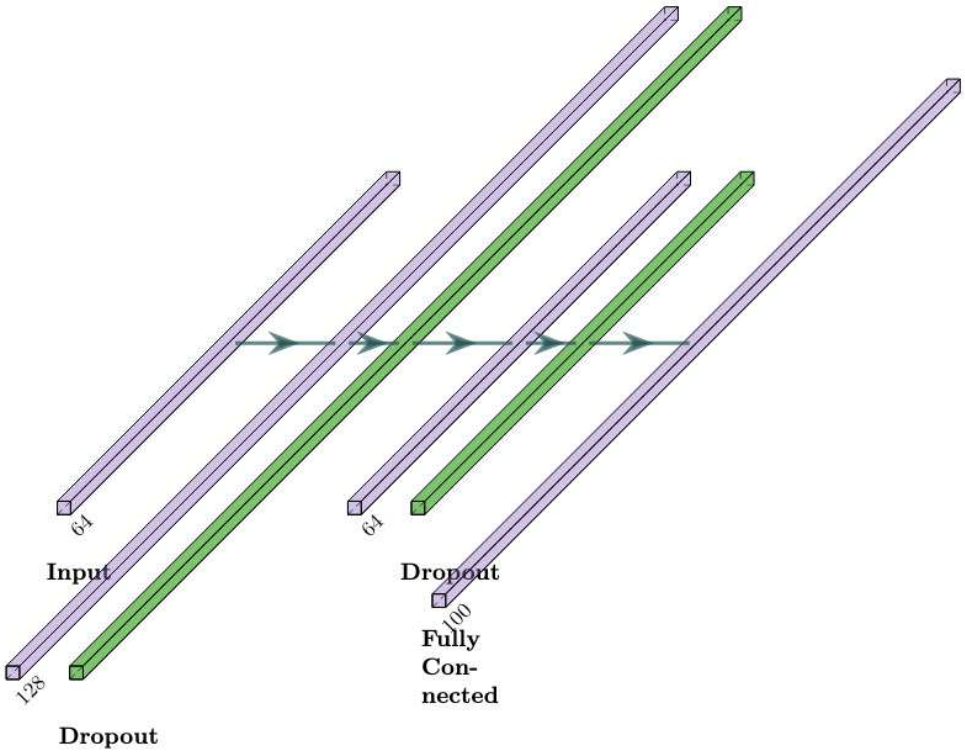


Figure 3.4: Architectures of the moderately complex classifier .

Chapter 4

Experiments

Our experiments were conducted using two standard benchmarks: the UCF-Crime dataset [26] for video analysis and the CIFAR-100 dataset for image-based tasks.

UCF-Crime: This dataset is designed for video anomaly detection and contains 1,610 training videos and 290 testing videos, which are classified into 13 distinct types of anomalous events. Instead of using raw video files, which are computationally intensive, our experiments leveraged pre-extracted features. Specifically, we used the I3D features, which were generated using an inflated 3D ResNet50 model. This approach is consistent with established methodologies, such as those in the CLAP paper, and allows for the efficient representation of a video's spatial and temporal characteristics.

CIFAR-100: This is a widely-used image dataset containing 60,000 color images distributed across 100 classes. To simulate a realistic feature-extraction pipeline, we employed a ResNet20 model that was pre-trained on 50% of the CIFAR-100 data. The model's final classification layer was removed, allowing it to function as a feature extractor that converts raw images into intermediate feature representations.

The experiments were implemented using the PyTorch framework, a leading deep learning library chosen for its flexibility and robust support for GPU acceleration via NVIDIA CUDA. The optimization process, central to the gradient-based reconstruction attacks, was performed using the L-BFGS optimizer. This is a powerful quasi-Newton method that is highly effective for solving complex optimization problems by approximating the inverse Hessian matrix to guide its search. We adopted the standard honest-but-curious threat model. In this scenario, the adversary is assumed to follow the prescribed protocol correctly (i.e., it is "honest"). However, it is also "curious" and will attempt to analyze any legitimately received information, such as shared model gradients, to infer sensitive data from the other participants. To ensure the scientific validity and reproducibility of our findings, a fixed manual seed was set for the PyTorch random number generator

at the beginning of each experiment. This practice eliminates variability from stochastic processes (such as model weight initialization) and guarantees that our results are deterministic and can be precisely replicated by other researchers.

4.1 Results on Video Frames with Simple Transformations

The primary objective of this research was to determine if gradient inversion attacks, previously demonstrated on static images, could be effectively extended to video data. To establish a baseline, we initially treated videos as a sequence of independent frames and conducted a series of reconstruction experiments.

We selected three distinct videos from the UCF crime dataset to represent different levels of scene complexity based on camera motion: a static scene with no camera movement (Video 2), a scene with minimal camera movement (Video 1), and a scene with frequent and significant camera movement (Video 3). This variation allowed us to assess how motion impacts the reconstruction quality. To manage computational demands and align with common practices in foundational privacy attack literature, the native 240x320 resolution of the video frames was downsampled to 32x32. All experiments were conducted using a batch size of one, ensuring that the gradients generated were specific to a single frame at a time.

To replicate a collaborative learning scenario, we simulated a federated learning environment where gradients, rather than raw data, are shared for global model training. We employed a LeNet classifier for our experiments, a choice made to directly mirror the foundational methodology of the Deep Leakage from Gradients (DLG) attack [29], which serves as our primary attack vector. The specific architecture of the LeNet model is consistent with that detailed in the original DLG study, as illustrated in Figure 4.1. The block diagrams were prepared using the PlotNeuralNet software package [17].

By applying the DLG attack algorithm to the gradients produced for each frame, we aimed to reconstruct the original training data. The reconstruction process for each frame was an optimization task run for 300 iterations. A reconstruction was deemed successful if the loss—calculated as

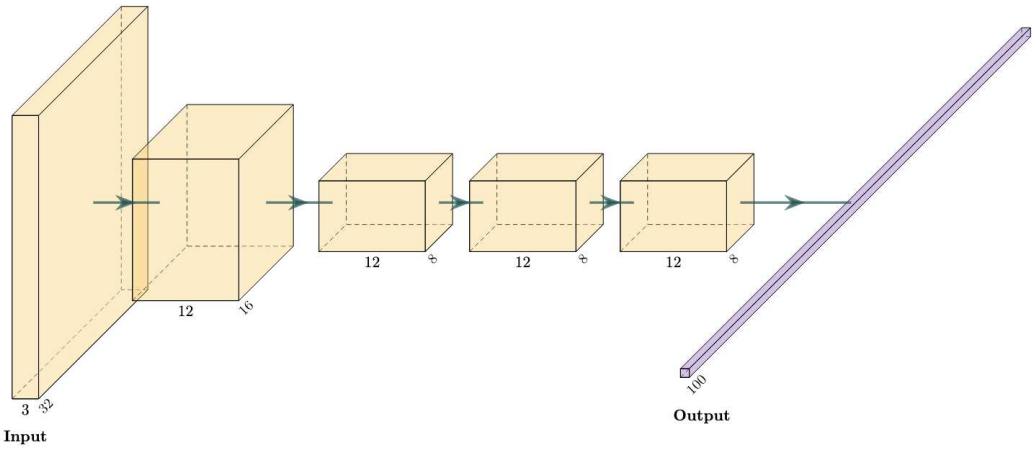


Figure 4.1: LeNet used in the DLG attack.

the sum of squared differences between the original gradients and the gradients of the reconstructed frame—dropped below a stringent threshold of 0.00001. If this threshold was not achieved, the optimization was re-initialized with a different random seed and repeated up to a maximum of ten attempts to ensure robustness. Recognizing that the reconstructed 32x32 frames have limited visual utility, we applied advanced image super-resolution techniques to upscale them to a more interpretable 128x128 resolution. This step was crucial for evaluating the practical severity of the information leakage. We investigated three distinct scenarios reflecting varying levels of prior knowledge an attacker might possess as mentioned before zero number of frame access, single frame access and multiple frame access. The final, upscaled frames were compiled into video files at a frame rate of 30 frames per second for qualitative assessment. A visual summary of this entire experimental pipeline is presented in Figure 3.2.

To rigorously assess the effectiveness of our video quality enhancement methods, we employed a quantitative evaluation framework using two widely recognized metrics: the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM). For a consistent and standardized comparison across all test cases, each video frame was resized to a uniform resolution of 128×128 pixels before analysis. The quality metrics were then computed on a frame-by-frame basis, providing a detailed assessment of reconstruction fidelity throughout the video sequence.

Peak Signal-to-Noise Ratio (PSNR): PSNR is a well-established metric used to objectively measure the quality of a reconstructed image or video by comparing it to its original version. It calculates the ratio between the maximum possible pixel value (the "signal") and the amount of error, or "noise," introduced during processing, such as in lossy compression. This metric is expressed on a logarithmic decibel (dB) scale. For standard 8-bit images, PSNR values typically fall between 30 and 50 dB, where higher values signify a smaller amount of distortion and a closer match to the original [11]. While its computational simplicity has made it a staple in the field, PSNR has a notable limitation: it operates on a pixel-by-pixel error calculation and does not always align with human visual perception. An image with a high PSNR may still appear unnatural or distorted to a human observer because the metric does not account for the structural context of pixel errors [30].

Structural Similarity Index (SSIM): To address the perceptual shortcomings of PSNR, we also utilize the SSIM. Unlike PSNR, SSIM is designed to evaluate reconstruction quality in a way that more closely mirrors how the human visual system works. It assesses the similarity between two images by comparing three key components:

- Luminance: The similarity in the brightness between the images.
- Contrast: The similarity in the range between the brightest and darkest areas.
- Structure: The similarity in the structural information, such as edges and textures.

The resulting SSIM value is a decimal between -1 and 1, where a value of 1 indicates a perfect match between the original and reconstructed images. By incorporating these perceptual principles, SSIM provides a more nuanced and meaningful measure of visual quality. Using both PSNR and SSIM together offers a more comprehensive and balanced evaluation, capturing both the raw pixel-level fidelity and the preservation of structural integrity.

To provide a comprehensive quantitative assessment of the reconstructed frames, a detailed analysis was conducted, with the results summarized in Table 4.1. This evaluation compares video

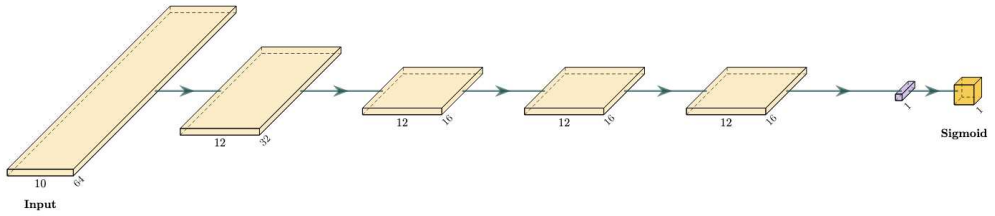


Figure 4.2: Modified LeNet-style classifier.

frames across three distinct categories to measure the effectiveness of the attack and subsequent enhancement techniques. The terms used in Table 4.1 are defined as follows:

The **High** quality category consists of the original, unaltered video frames from the UCF-Crime dataset. The **Low** quality category includes the video frames as they were initially reconstructed by the DLG attack, representing the baseline success of the privacy breach before any post-processing. Finally, the **Enhanced** category comprises frames that were refined by applying various image super-resolution (SR) techniques to the "Low" quality reconstructions.

A prerequisite for both PSNR and SSIM metrics is that the compared frames must have identical dimensions. Therefore, all video frames across all three categories were standardized to a uniform resolution of 128x128 pixels to ensure a valid and consistent comparison. Our initial experiments, comparing the "High" and "Low" quality videos, confirm that the DLG attack is independently capable of reconstructing recognizable video frames solely from shared model gra-

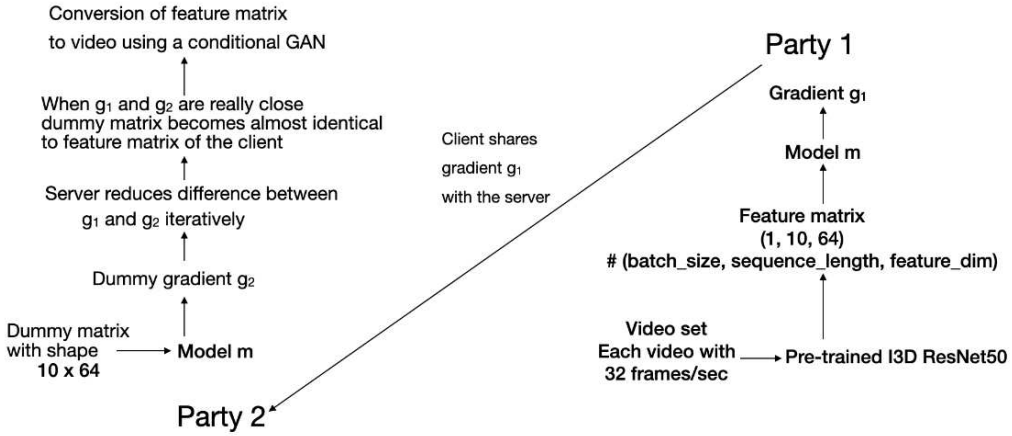


Figure 4.3: Gradient Leakage Attack: Proposed Video Reconstruction Steps

dients. This baseline reconstruction, however, can be significantly improved. By applying super-resolution techniques, we were able to further refine the leaked frames, thereby increasing their clarity and potential utility for an adversary. Figure 4.4 provides a visual comparison, showcasing sample frames in their "Low" quality state post-DLG attack and their "Enhanced" state after super-resolution has been applied.

The effectiveness of different super-resolution strategies was found to be dependent on the video content. For videos with static backgrounds or minimal camera motion, super-resolution methods that use no reference or a single reference frame provided the most significant quality boost. In contrast, for dynamic scenes characterized by frequent camera movement, employing multiple reference frames from the video sequence yielded superior performance improvements. The experimental results conclusively show that applying these super-resolution techniques leads to a measurable increase in the quality of the reconstructed videos, as quantified by a notable improvement in their SSIM scores, indicating that the enhanced frames are structurally more similar to the original videos.

Table 4.1: Quantitative Evaluation of Reconstructed Frame Fidelity for Different image super-Resolution Strategies.

		Video SR w/o Ref.		Video SR w/ 1 Ref.		Video SR w/ Mult. Refs.	
	Baseline (High vs. Low)	Enhanced vs. High	Enhanced vs. Low	Enhanced vs. High	Enhanced vs. Low	Enhanced vs. High	Enhanced vs. Low
Video 1	28.40/0.2026	**28.75**/0.2145	29.52/0.6103	28.68/0.2254	30.12/**0.6693**	28.68/**0.2263**	**30.14**/0.6744
Video 2	28.47/0.1968	**28.86/0.2311**	30.01/0.6842	28.76/0.2206	**30.57/0.7006**	28.76/0.2205	**30.57**/0.6998
Video 3	28.63/0.2881	28.74/0.2956	31.37/0.7978	**28.95**/0.3177	32.76/0.8771	**28.95/0.3178**	**32.77/0.8780**

For the experimental setup involving pre-extracted features, our methodology was designed to test the feasibility of reconstructing intermediate data representations within a FL framework. Here we detail the data source, preprocessing steps, model architecture, and the specific configuration of the gradient inversion attack.

4.2 Results of Gradient Inversion Attacks on Extracted Video Features

To simulate a realistic video processing pipeline, we utilized pre-extracted features from a public GitHub repository associated with the RTFM project (https://github.com/tianyu0207/RTFM). This repository provides features extracted using the Inflated 3D ConvNet (I3D) model, which is highly effective at capturing the complex spatio-temporal dynamics inherent in video data. Each video clip in the dataset is represented by a feature matrix with the shape (1, 10, 2048).

However, the high dimensionality of these raw feature matrices poses computational challenges for gradient inversion algorithms like the DLG attack. Therefore, a preprocessing step was necessary to downsample the data. We applied a max pooling operation to reduce the feature dimension, transforming each matrix to a more manageable shape of (1, 10, 64). This downsampled feature matrix served as the input for the model within our simulated FL environment. The primary objective was then to apply the DLG algorithm to the shared gradients to ascertain if these original, albeit downsampled, feature matrices could be successfully reconstructed. Figure 4.3 illustrates the complete workflow for this feature-based reconstruction approach.

The choice of the client-side classifier is critical in evaluating gradient leakage. For this experiment, we designed a modified LeNet-style architecture, as depicted in Figure 4.2, which accepts the I3D features as input. This architecture was intentionally selected because its simple, well-understood structure is known to be vulnerable to DLG attacks. Using a susceptible model provides a clear baseline to determine if the feature extraction process itself offers any inherent privacy protection.

The DLG attack was executed for 20,000 iterations to provide ample opportunity for the algorithm to converge on a solution. A significant challenge in gradient-based reconstruction is the risk of the optimization process stalling in a local minimum. To counteract this, we implemented an adaptive optimization strategy: The reconstruction loss was monitored at intervals of 1,000 iterations to detect stagnation. If the loss failed to improve, a small amount of random noise, sampled

from a normal distribution ($\mu = 0$, $\sigma = 0.001$), was added to the dummy data being optimized. This helps perturb the process out of a local minimum. Simultaneously, the learning rate of the optimizer was halved to encourage finer-grained convergence once a more promising search direction was found. To ensure a thorough evaluation, we experimented with two different optimizers, Adam and L-BFGS, to assess their relative effectiveness in performing the reconstruction task. But DLG attack failed to reconstruct the feature matrices accurately this time.

4.3 Influence of Feature Extractors

To comprehensively evaluate the impact of using a feature extractor on the efficacy of gradient inversion attacks, we designed a series of controlled experiments. For this analysis, we selected three well-established gradient inversion algorithms from the literature: Deep Leakage from Gradients (DLG), improved Deep Leakage from Gradients (iDLG), and R-GAP (Recursive gradient attack on privacy).

The CIFAR-100 dataset was chosen for this study due to its known vulnerability to image reconstruction attacks, providing a reliable benchmark for our investigation. We employed a ResNet20 model as the feature extractor. This model was pre-trained on 50% of the CIFAR-100 dataset to ensure it could generate meaningful representations, and its final classification layer was removed to isolate its feature-extraction capabilities.

Within a simulated federated learning environment, we conducted two primary experiments. The core objective was to determine if an attacker could successfully reconstruct the intermediate feature matrices (the output of the ResNet20 extractor) from shared gradients. In the first experiment, the feature extractor was paired with a simple, low-complexity classifier. In the second, it was paired with a moderately complex classifier, as illustrated in Figure 3.3 and 3.4. This dual-classifier approach allowed us to assess how the complexity of the model component trained on the features influences vulnerability.

Our findings, summarized in Table 4.2, reveal a nuanced relationship between feature extraction and privacy. The results demonstrate that the mere use of feature matrices is not sufficient to

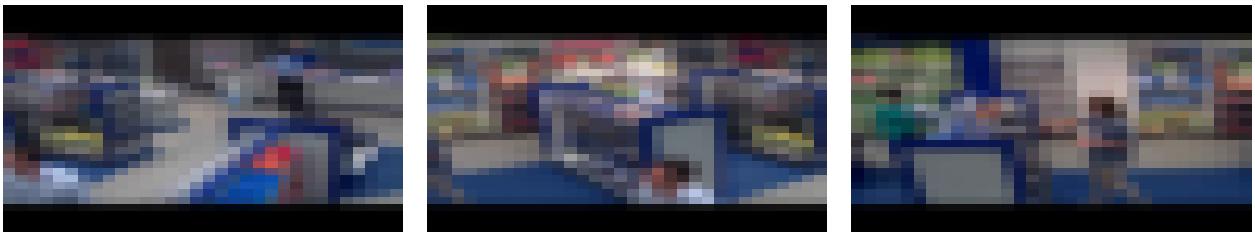
Table 4.2: Reconstruction Performance of DLG, iDLG, and R-GAP on Feature Matrices with Varying Classifier Complexity (N - feature matrices could not retrived from gradients, Y - feature matrices were retrived from gradients.)

	Moderately complex classifier with feature extractor	Moderately complex classifier with raw images	Simple classifier with feature extractor	Simple classifier with raw images
DLG	N	Y	Y	Y
iDLG	N	Y	Y	Y
R-GAP	N	N	N	N

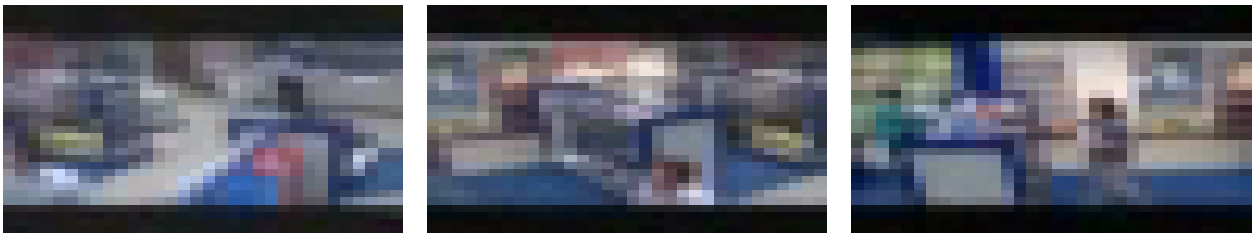
guarantee security. When a simple classifier was attached to the feature extractor, all three attack algorithms were able to successfully reconstruct the feature matrices from the gradients.

However, a significant shift was observed in the second experiment. While moderately complex classifiers are known to be vulnerable to data leakage when trained directly on raw images, we found that these same classifiers effectively prevented leakage when operating on the features provided by the extractor. In this configuration, the attacks failed to reconstruct the intermediate features. From this, we conclude that the integration of a feature extractor substantially increases the difficulty of performing successful gradient inversion attacks, effectively protecting the data when the subsequent classifier possesses even a moderate level of complexity.

(a) Representative video frames illustrating scenes from the UCF-crime dataset.



(b) Representative video frames used as input for training the model.



(c) Sample video frames obtained via the DLG attack.

(d) Sample enhanced video frames produced by the no-reference super-resolution method.

(e) Sample enhanced video frames produced by super-resolution with one reference.

(f) Sample enhanced video frames produced by super-resolution with multiple references.

Figure 4.4: Frames obtained through DLG and enhanced using super-resolution.

Chapter 5

Discussion

In our experimental evaluation, we first investigated the efficacy of standard gradient inversion techniques against video processing pipelines that utilize feature extractors. Our initial results were definitive: the canonical DLG algorithm proved ineffective, consistently failing to reconstruct the pre-processed feature matrices from their gradients. Further experiments with various attack models on the CIFAR-100 dataset confirmed that while using a feature extractor is not an absolute defense against inversion, it introduces a substantial barrier that significantly complicates the attack process.

The challenges of attacking feature matrices derived from video data are twofold. First, there is a fundamental difficulty in evaluating the quality of a successful reconstruction. Unlike attacks on images or text, where success can be judged by visual fidelity or semantic coherence, reconstructed feature matrices are abstract numerical vectors. They lack any intuitive, human-interpretable form, compelling reliance on quantitative metrics like Mean Squared Error, which may not fully capture the functional impact of the leaked information.

Second, the inherent nature of these features presents a formidable challenge to the inversion process itself. Gradient-based attacks on images implicitly leverage strong spatial priors—the knowledge that adjacent pixels are highly correlated. Similarly, attacks on text exploit sequential dependencies between words. In contrast, the feature matrices in our scenario are abstract representations that have been aggregated from multiple frames. They lack this inherent spatial or sequential structure, which an attacker could otherwise exploit. This absence of discernible patterns results in a less constrained optimization problem for the attacker, making it significantly more difficult to converge to the original private values. Situating our work within the existing literature, it is clear that while attacks on image and text data have seen tremendous success, gradient inversion on numerical feature matrices remains an active and challenging area of research. The most pertinent work in this domain is TabLeak [28], which established the state-of-the-art for

attacking tabular data composed of both numerical and categorical features. However, a critical aspect of TabLeak's success is its reliance on the presence of categorical columns, as the discrete nature of this data type significantly narrows the search space for an attacker. The authors of TabLeak themselves report a significant performance disparity, noting that reconstruction accuracy for continuous features is up to 30% lower than for categorical features in the same batch.

This distinction is crucial, as our experimental scenario involves feature matrices composed exclusively of continuous numerical values. This specific context, which lacks the structural and categorical aids that methodologies like TabLeak exploit, represents a more challenging and previously unaddressed case in the privacy literature. Our experiments, therefore, aim to assess the practical security of these common data representations where no such structural shortcuts exist for an attacker.

Chapter 6

Conclusion and Future Work

We designed the experimental framework to evaluate the vulnerability of video data to deep leakage attacks, which reconstructs private data from shared model gradients. Our investigation systematically explores how different data preprocessing choices impact the efficacy of such attacks within a federated learning context. The experiments were structured to address three primary objectives: First, we sought to quantify the fidelity of video frames that can be reconstructed through gradient inversion. Second, we aimed to assess the practical utility of the leaked information, determining if the recovered frames are sufficiently clear for malicious use, such as subject identification. Finally, we investigated whether the quality of these reconstructed frames could be adversarially enhanced using publicly available image super-resolution technologies. Our empirical results confirm that super-resolution techniques can markedly improve the visual quality of the reconstructed frames, transforming initially indistinct images into more intelligible ones. This finding is particularly significant in the context of the evolving threat landscape. Recent literature has demonstrated that modern gradient inversion attacks are increasingly effective, successfully compromising even high-resolution imagery and larger batch sizes—two factors that were previously thought to provide a degree of protection. Our work reinforces the conclusion that relying on these properties as implicit safeguards is an insufficient defense strategy for privacy-sensitive collaborative systems.

In contrast, our experiments identify a highly promising mitigation technique. We demonstrate that employing a pre-trained feature extractor—and training the collaborative model on these intermediate features rather than raw video data—significantly impedes the attack's effectiveness. This layer of abstraction appears to disrupt the gradient leakage process, making it substantially more difficult for an adversary to reconstruct the original video content. Given these findings, this work provides empirical evidence that the security of collaborative video analysis systems depends heavily on the data preprocessing pipeline. While some configurations are highly vulnerable, the

use of feature extraction offers a robust path toward mitigation. Therefore, our future work will be dedicated to a more comprehensive investigation of this defensive strategy, analyzing the trade-offs between model accuracy and privacy protection across various system parameters. We will also investigate the significant privacy threat that generative AI introduces to federated learning. A malicious actor may not even start from random noise to reconstruct the private training data. Instead, they can leverage a pre-trained generative model to produce a much more informed starting point that is already close to the distribution of the private data. Our research will examine how this technique improves the reconstruction of private video data from shared gradients, thereby posing a greater risk to data privacy.

References

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. pp. 308–318 (2016)

[2] Al-Lahham, A., Zaheer, M.Z., Tastan, N., Nandakumar, K.: Collaborative learning of anomalies with privacy (clap) for unsupervised video anomaly detection: A new baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12416–12425 (2024)

[3] Aono, Y., Hayashi, T., Wang, L., Moriai, S., et al.: Privacy-preserving deep learning via additively homomorphic encryption. IEEE transactions on information forensics and security **13**(5), 1333–1345 (2017)

[4] Balunović, M., Dimitrov, D.I., Staab, R., Vechev, M.: Bayesian framework for gradient leakage. arXiv preprint arXiv:2111.04706 (2021)

[5] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for federated learning on user-held data. arXiv preprint arXiv:1611.04482 (2016)

[6] Chen, C., Campbell, N.D.: Understanding training-data leakage from gradients in neural networks for image classification. arXiv preprint arXiv:2111.10178 (2021)

[7] Chen, W., Zhang, C., Zhang, W., Cai, J.: Class-hidden client-side watermarking in federated learning. Entropy **27**(2), 134 (2025)

[8] Dalal, M., Juneja, M.: Steganography and steganalysis (in digital forensics): a cybersecurity guide. Multimedia Tools and Applications **80**(4), 5723–5771 (2021)

[9] Ding, X., Liu, Z., You, X., Li, X., Vasilakos, A.V.: Improved gradient leakage attack against compressed gradients in federated learning. Neurocomputing **608**, 128349 (2024)

[10] Du, J., Hu, J., Wang, Z., Sun, P., Gong, N.Z., Ren, K., Chen, C.: Sok: On gradient leakage in federated learning. arXiv preprint arXiv:2404.05403 (2024)

[11] Fowler, C.: What is a good psnr value for image? (2022), https://poletoparis.com/what-is-a-good-psnr-value-for-image/, accessed: Apr. 11, 2025

[12] Geiping, J., Bauermeister, H., Dröge, H., Moeller, M.: Inverting gradients-how easy is it to break privacy in federated learning? Advances in neural information processing systems **33**, 16937–16947 (2020)

[13] Gong, H., Jiang, L., Liu, X., Wang, Y., Gastro, O., Wang, L., Zhang, K., Guo, Z.: Gradient leakage attacks in federated learning. Artificial Intelligence Review **56**(Suppl 1), 1337–1374 (2023)

[14] Guo, P., Zeng, S., Chen, W., Zhang, X., Ren, W., Zhou, Y., Qu, L.: A new federated learning framework against gradient inversion attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 16969–16977 (2025)

[15] Huang, Y., Gupta, S., Song, Z., Li, K., Arora, S.: Evaluating gradient inversion attacks and defenses in federated learning. Advances in neural information processing systems **34**, 7232–7241 (2021)

[16] Huang, Y., Song, Z., Li, K., Arora, S.: Instahide: Instance-hiding schemes for private distributed learning. In: International conference on machine learning. pp. 4507–4518. PMLR (2020)

[17] Iqbal, H.: Plotneuralnet. https://github.com/HarisIqbal88/PlotNeuralNet (2018)

[18] Jeon, J., Lee, K., Oh, S., Ok, J., et al.: Gradient inversion with generative image prior. Advances in neural information processing systems **34**, 29898–29908 (2021)

[19] Jin, X., Chen, P.Y., Hsu, C.Y., Yu, C.M., Chen, T.: Cafe: Catastrophic data leakage in vertical federated learning. Advances in neural information processing systems **34**, 994–1006 (2021)

[20] Kadhe, S., Rajaraman, N., Koyluoglu, O.O., Ramchandran, K.: Fastsecagg: Scalable secure aggregation for privacy-preserving federated learning. arXiv preprint arXiv:2009.11248 (2020)

[21] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[22] Lam, M., Wei, G.Y., Brooks, D., Reddi, V.J., Mitzenmacher, M.: Gradient disaggregation: Breaking privacy in federated learning by reconstructing the user participant matrix. In: International Conference on Machine Learning. pp. 5959–5968. PMLR (2021)

[23] Li, Z., Hubchak, M., Zhu, Y.: Deep leakage from gradients in multiple-label medical image classification. In: 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI). pp. 447–448. IEEE (2021)

[24] Reddi, S.J., Kale, S., Kumar, S.: On the convergence of adam and beyond. arXiv preprint arXiv:1904.09237 (2019)

[25] Scheliga, D., Mäder, P., Seeland, M.: Dropout is not all you need to prevent gradient leakage. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 9733–9741 (2023)

[26] Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6479–6488 (2018)

[27] Tran, P.T., Phong, L.T.: On the convergence proof of amsgrad and a new version. IEEE Access **7**, 61706–61716 (2019). https://doi.org/10.1109/ACCESS.2019.2916341

[28] Vero, M., Balunović, M., Dimitrov, D.I., Vechev, M.: Tableak: Tabular data leakage in federated learning. arXiv preprint arXiv:2210.01785 (2022)

[29] Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: International Conference on Computer Vision Workshops (ICCVW)

[30] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)

[31] Wu, Z., Yao, T., Fu, Y., Jiang, Y.G.: Deep learning for video classification and captioning. In: Frontiers of multimedia research, pp. 3–29 (2017)

[32] Xue, L., Hu, S., Zhao, R., Zhang, L.Y., Hu, S., Sun, L., Yao, D.: Revisiting gradient pruning: A dual realization for defending against gradient attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6404–6412 (2024)

[33] Yang, C.S., So, J., He, C., Li, S., Yu, Q., Avestimehr, S.: Lightsecagg: Rethinking secure aggregation in federated learning. arXiv e-prints pp. arXiv–2109 (2021)

[34] Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: CVPR (June 2020)

[35] Yin, H., Mallya, A., Vahdat, A., Alvarez, J.M., Kautz, J., Molchanov, P.: See through gradients: Image batch recovery via gradinversion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16337–16346 (2021)

[36] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)

[37] Zhang, K., Cheng, S., Shen, G., Ribeiro, B., An, S., Chen, P.Y., Zhang, X., Li, N.: Censor: Defense against gradient inversion via orthogonal subspace bayesian sampling. arXiv preprint arXiv:2501.15718 (2025)

[38] Zhao, B., Mopuri, K.R., Bilen, H.: idlg: Improved deep leakage from gradients. arXiv preprint arXiv:2001.02610 (2020)

[39] Zhu, J., Blaschko, M.: R-gap: Recursive gradient attack on privacy. arXiv preprint arXiv:2010.07733 (2020)

[40] Zhu, L., Liu, Z., Han, S.: Deep leakage from gradients. Advances in neural information processing systems **32** (2019)