Adversarial Face Recognition: Detection Framework for Video-Based Attack Sequences

James Wilson III
jiw2@fordham.edu
CIS Department, Fordham University
Dr. Thaier Hayajneh, Ph.D

***Abstract*—This project investigates the vulnerability of facial recognition (FR) systems to adversarial attacks in video-based surveillance scenarios and proposes a detection framework for identifying and detecting these attacks. The primary attack that the detection framework hopes to defend against are Adversarial Transformation Network (ATN) based attacks, which are capable of generating imperceptible perturbations that can fool modern FR models while maintaining the same visual realness in real time. To counter this type of threat, a detection framework that encompassed the combination of temporal drift analysis and ensemble disagreement was developed, achieving 89.3% accuracy on ATN-generated adversarial examples and demonstrating notable cross-attack generalization to gradient-based attacks (PGD: 90.0%, FGSM: 83.3%). Our findings reveal that temporal consistency analysis significantly outperforms ensemble-based detection methods when compared individually, but a fusion of both methods in conjunction with a simple binary classifier yields the best results.**

***Index Terms*—adversarial machine learning, face recognition, attack detection, temporal analysis, ensemble models**

I. Introduction

A. Background and Motivation

Facial recognition (FR) systems are increasingly embedded in critical infrastructure to enhance security, such as access control systems, airports, public transport systems, and law enforcement applications, to name a few. Although these systems achieve high accuracy under benign conditions, they are vulnerable to adversarial evasion attacks. Adversarial attacks on AI models are becoming an increasingly common attack vector that pose serious threats to critical systems due to the adoption of these systems that utilize convolutional neural networks (CNNs) and metric learning methods such as ArcFace, FaceNet, CosFace, and SphereFace. Despite the advancements of these deep learning (DL) models, FR systems remain vulnerable to adversarial attacks. Threat actors can add human-imperceptible perturbations to facial images or video frames to manipulate embedding vectors, which causes the system to misclassify or misverify identities. Recent advances in real-time attacks as well such as the ReFace framework proposed by S. Hussain et. al in 2022 is a real-time adversarial attack designed specifically to target Deep Neural Network (DNN)-based FR systems. Traditional adversarial attacks like Projected Gradient Descent (PGD) or Fast Gradient Sign Method (FGSM) are not effective or practical for real-time applications since both attacks require solving a data-dependent optimization problem for each new input using multiple forward-backward passes through the victim model. This means that these attacks are limited by the time and computational power necessary to generate adversarial examples. The ReFace framework was proposed to address this limitation by demonstrating that real-time attacks are possible through the use of Adversarial Transformation Networks (ATNs).

Using ATNs, attackers can generate perturbations fast enough to bypass commercial facial recognition APIs like AWS Rekognition or Microsoft Azure Face in real time. ATNs model the adversarial example generation process as a feed-forward neural network, which allows the generation of an adversarial perturbation in a single forward pass, without requiring multiple gradient computations during inference. As surveillance systems increasingly rely on video-based FR pipelines, adversarial ML research must take into account these temporal dynamics. Video streams pose distinct challenges that include things like perturbations remaining consistent across frames and enduring natural variations in pose and illumination. Traditional image-based attacks applied independently to individual frames typically introduce unnatural frame by frame embedding alterations, which produce easily detectable behavior. This motivates the need for temporally coherent attack generation methods as well as sequence-level detection mechanisms. There

is a gap between he defensive landscape for this attack vector and the research of this attack itself, in which most existing defenses rely on adversarial training or focus on low-level image artifacts in the case of deepfake, and do not account for transferable, high-speed perturbations generated in real-time like that of ATNs. Therefore, this project's objective is to propose a real-time detection framework that leverages temporal embedding drift to ensemble embedding disagreement to detect both ATN and gradient-based attacks on FR systems. By focusing on how perturbations affect the model's internal representation over time, rather than only the input image, this approach aims to provide a lightweight yet robust attack-agnostic solution that could eventually be suitable for real-world deployment.

B. Contribution

This research builds upon this rising field by designing a U-Net based ATN able to generate subtle, temporally consistent perturbations that can fool multiple FR models at once. Our core contribution lies in the design of a dual-detector adversarial detection framework that makes the most of two distinct signals:

1) Temporal Drift Detection: Benign video sequences produce stable embeddings over time, while adversarial perturbations, produce inconsistent temporal trajectories. By analyzing certain temporal drift metrics, we were able to identify adversarial sequences without relying on model gradients or assumptions about the knowledge of the attack.

2) Ensemble Disagreement Detection: Multiple FR models with different architectures are able to process the same input and attempt to achieve robust classification or detection results based on context. However, adversarial perturbations can shift their embeddings in divergent directions, and while this technique has been explored in previous works, we demonstrate that the most common modern FR architectures share similar vulnerabilities, reducing the effectiveness of ensemble-based disagreement as an individual detection method.

To fuse these methods, we introduce a classifier component, implemented utilizing logistic regression, that integrates both detectors' outputs into a final adversarial prediction. This classifier yields optimal performance by learning the best threshold within the feature space of the temporal and ensemble metrics.

Ultimately, the contributions of this work are as follows: 1) Enhanced ATN Architecture, 2) Dual-Detector Adversarial Detection Framework, 3) Classifier Fusion Module, 4) Extensive Cross-Attack Generalization Analysis, and 5) Enhanced Experimental Evaluation.

II. Related Works

The security of face recognition (FR) systems has become a focal point in the adversarial machine learning research community due to FR systems being widely utilized in authentication and surveillance. The research focuses on both attack strategies/vectors that threat actors may use to target deep neural networks, which are the underlying models of FR systems, and the development of defensive mechanisms designed to mitigate these attacks. This section reviews prior work across two primary areas: 1) adversarial attacks on FR systems and 2) defense strategies including temporal and ensemble-based methods.

A. Adversarial Attacks on Face Recognition Systems

Adversarial attacks involve deliberately introducing small perturbations to the input data to deceive a ML/DL model into producing false predictions. As FR systems have become a cornerstone of modern biometric authentication that are widely used within critical infrastructure, they currently benefit the most from the representational power of deep neural networks, and subsequently are vulnerable to adversarial attacks. Y. Xu et al. show traditional adversarial attacks such as FGSM and PGD have been able to effectively exploit these models, which could allow adversaries to impersonate authorized individuals or bypass identification systems altogether [1]. The authors illustrate the impacts of these attacks by showing that, when a fixed threshold is used, the attacks lead to a higher False Match Rate (FMR) [1], which can allow attackers to bypass the verification process. Research of these attacks are primarily focused on black-box attacks since the black-box setting is the most common threat model in real-world scenarios since attackers have limited knowledge on the models they target. Other literature that mention these adversarial attacks, that will also be talked about in this paper, focus on different domains like network security or Natural Language Processing or LLMs.

Beyond standard image-based attacks, a growing body of work examines transferable adversarial examples capable of fooling multiple FR models simultaneously.

Since the attack methods mentioned require iterative optimization that limit their reliability in real-world scenarios, research on real-time adversarial evasion attacks has begun to evolve as a new threat vector. The ReFace framework introduced by S. Hussain et al. realize the

limitations of previous techniques and utilize Adversarial Transformation Networks (ATNs) to fool face recognition classifiers instead of gradient-based attacks. They used ATNs since they can produce adversarial images at a much faster rate than PGD or FGSM and perform well if trained properly. The real-time component is due to the fact that ATNs generate adversarial images from an input without multiple passes back and forth. Their developed real-time attacks on face recognition systems reduced accuracy from 82% to 16.4%. This is why defensive mechanisms against this attack vector are vital to be researched and tested [2]

B. Defense Strategies

Recently, there has been focused research concerning defense strategies for ML and DL models using temporal and spatio-temporal analysis. This essentially constitutes a focus on detecting anomalies in time-series data, primarily in video-based security like Deepfake and network security. Temporal analysis has proven to be quite useful in these domains for identifying manipulated content. For example, U. Muhammad et al. showed that in presentation attack detection (PAD), methods like Temporal Sequence Sampling (TSS) show strong results in identifying spoofing by compensating for inter-frame motion and encoding the temporal data into a compact representation vector [3]. This shows that temporal patterns are effective in distinguishing between benign frames and manipulated ones, but more specifically, natural facial movements and unnatural ones. In the world of deepfake detection, since this has also been an increased attack vector as a result of stronger performing AI models, spatio-temporal consistency was leveraged by Chen et al. that addresses the limitations of novel detection techniques. They used a convolutional neural network (CNN) to extract spatial domain features of deepfake images and used an optical flow technique to describe these images. They use both temporal and spatial features and train their model to treat the detection as a binary classification task [4]. In another domain, unsimilar to Deepfake or multimedia, we see that temporal and time-series analysis has been effective in network intrusion detection. Since many advanced attacks like Advanced Persistent Threats (APTs) and stealthy denial-of-service (DoS) attacks occur over time, it is intuitive to see why this type of analysis is useful. However, Z. Liu et al. conducted a study of adversarial attacks on RNN models with time steps in network intrusion detection systems. They designed an attack model called the Temporal Adversarial Examples Attack Model (TEAM) to generate adversarial examples with temporal dynamics as a way to exploit a model as well. They dubbed this phenomenon as the "next moment" attack since it is able to perturb early inputs and subsequently degrade model performance over time [5]. This research highlights that applying temporal anomaly detection in FR systems offers both new opportunities and new challenges. Ensemble learning has been extensively studied as a strategy to enhance robustness and generalization in ML models, not only for security purposes but for performance. In the context of DLL or FR systems specifically, combining multiple models trained with different architectures or conditions, ensemble learning can provide diverse perspectives on input data, making them less sensitive to single perturbations [6]. In APT detection, the TSE-APT framework,introduced by Cheng et al., employed multiple base learners such as Random Forest, Multi-Layer Perceptron, and Bidirectional Long Short-Term Memory Network models along with an adaptive self-attention mechanism, which demonstrated how ensembles improve detection accuracy in complex attack environments [7].

III. Problem Formulation

Modern FR systems rely on deep metric learning models such as ArcFace, FaceNet, and other various models to generate identity-discriminative embeddings by enforcing angular or cosine margin constraints during training. ArcFace in particular demonstrated that additive angular margins significantly improve distinctness in the embedding space by increasing inter-class separation while reducing intra-class variance [8]. Since these systems map input frames to high-dimensional embeddings, even small perturbations to the input may alter the embedding vector enough to cause misclassification in an FR system, an outcome well-documented in adversarial machine learning research. This section's purpose is to formalize the threat model, the video-based FR pipeline, and the adversarial detection objective. Let a video sequence consist of $\boldsymbol{T}$ frames:

$$X = \{x_1, x_2, \ldots, x_T\}, \quad x_t \in \mathbb{R}^{H \times W \times 3} \tag{1}$$

where each frame represents a normalized and aligned RGB facial image. In benign conditions, these embeddings evolve smoothly over time due to the physical continuity of natural facial motion. This assumption of temporal smoothness aligns with observations in adversarial video research that benign sequences exhibit consistent temporal dynamics, whereas perturbed sequences often

break this pattern [9]. In the adversarial scenario, a threat actor is able to generate perturbations (δ_t) such that

$$x'_t = x_t + \delta_t \tag{2}$$

The perturbations are constrained within the max norm (ℓ_∞), which ensures that perturbations are small at every pixel, spread out and not localized, and most importantly, imperceptible to human eyes. Black box threat models commonly assume the attacker uses surrogate architectures to craft perturbations that transfer to the target model, another inference supported by extensive research that adversarial subspaces overlap across different deep networks, even when trained independently [10]. The attacker's goal is to:

- Cause misclassification or misverification
- Maintain temporal consistency
- Avoid detection

Therefore, we consider three attack types:

- ATN-based attacks (learned perturbations
- PGD attacks (iterative gradient optimization)
- FGSM attacks (single-step gradient perturbations)

The classifier detection objective is to assign a binary label $y \in \{0, 1\}$ to the entire sequence based on deviations from expected temporal and cross-model behavior. The proposed detection system relies on two measurable signals:

- temporal drift: abrupt changes in embedding trajectories inconsistent with natural facial motion, and
- ensemble disagreement: divergence in pairwise embedding distances among multiple different face recognition models

The ideal detection system would have no dependence on the target model architecture (agnostic), support real-time performance, generalize across multiple types of attacks, and be robust/tolerable to natural facial motion and variation.

IV. Methodology and Framework Design

This section presents the proposed attack generation and the detection mechanisms. The methodology consists of three major components: (1) an Adversarial Transformation Network (ATN) for generating temporally coherent perturbations, (2) a dual-detector defense mechanism based on temporal drift and ensemble disagreement, and (3) a classifier fusion module for final decision-making.

A. System Architecture

A video sequence $X = \{x_1, x_2, \ldots, x_T\}$ containing T frames is processed by two systems:

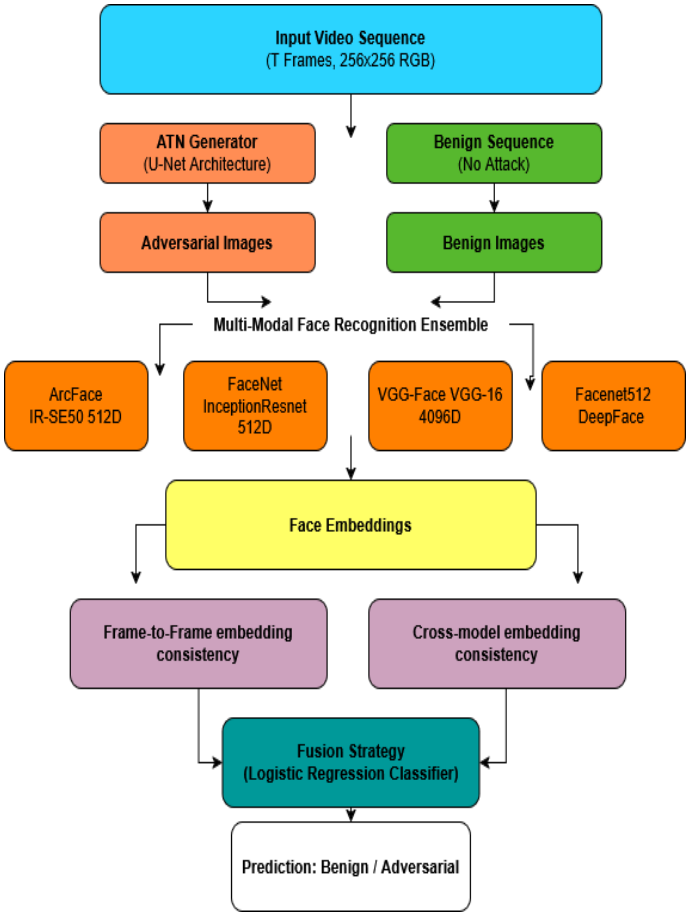

Fig. 1. System Architecture

1) Adversarial Generator: Produces an adversarial sequence $X' = \{x'_1, \ldots, x'_T\}$ by applying perturbations δ_t to each frame such that:

$$x'_t = x_t + \delta_t, ||\delta||_\infty \leq \epsilon \tag{3}$$

2) Ensemble Model Embedding Extraction: Four face recognition models extract embedding features of input frames in a given sequence (10 frames per sequence on average).

The relationship between these two components is illustrated in an architectural diagram seen in Figure 1.

B. Adversarial Transformation Network (ATN)

Adversarial Transformation Networks were first introduced by Baluja and Fischer [11] as a feed-forward, generator-style alternative to iterative gradient-based attacks. Unlike PGD or FGSM, which require backward passes, ATNs learn the mapping that is able to directly transform an input image into its adversarial equivalent in a single forward pass. This framework has recently been extended to the face recognition domain with,

for example, the ReFace method which trains an ATN to generate subtle, real-time perturbations that transfer across FR models [2]. Building on this research, we constructed our own U-Net-based ATN that simultaneously enforces adversarial embedding alteration and maintains perceptual similarity. Early attempts using Learned Perceptual Image Patch Similarity (LPIPS) based perceptual losses proved to be unstable for optimizing a series of frames because of its significantly higher computational complexity. Our formula replaces LPIPS with Structural Similarity Index (SSIM) + cosine-margin losses, inspired by the effectiveness of angular margin constraints in FR embeddings spaces [8] and the robustness of SSIM for modeling structural similarity in perceptual tasks [12]. SSIM measures how similar two images are in terms of luminance, contrast, and structure.

1) ATN Architecture: We adopted an enhanced U-Net based encoder-decoder structure. U-Net is a convolutional neural network (CNN) architecture with a U-shaped design that was initially developed for image segmentation. However, the U-Net structure is ideal for this research since it excels at pixel-level classification tasks as a result of its encoder and decoder functions. The encoder progressively reduces the spatial dimensions of an input image while increasing the number of feature channels, while the decoder mirrors the encoder by progressively upsampling the feature maps [13]. The U-net also skips connections which, for this use case, enable the preservation of fine-grained spatial details essential for maintaining visual imperceptibility.

The encoder that we built consists of:

- Four convolutional blocks with channel progression: $3 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512$
- Each block consists of: Conv(3x3) $\rightarrow$ BatchNorm $\rightarrow$ LeakyReLU(0.2) $\rightarrow$ MaxPool(2x2)

The decoder that we build consists of:

- Four upsampling blocks with reverse channel flow: $512 \rightarrow 256 \rightarrow 128 \rightarrow 64$
- Each block uses: Upsample(2x) $\rightarrow$ Conv(3x3) $\rightarrow$ BatchNorm $\rightarrow$ ReLU

The final layer of the ATN architecture is a:

- A 3 x 3 convolution to produce a perturbation map of equal spatial size
- Tanh activation followed by scaling ensures the perturbation remains within the max norm constraint

2) Loss Functions: The ATN is trained using a loss formula that attempts to balance two objectives: (1) generating adversarial perturbations that can effectively shift embeddings away from their benign equivalent, and (2) remaining visually imperceptible across video frames. After extensive experimentation, we adopted this loss function equation that uses only two terms:

$$L_{total} = L_{adv} + \lambda_{img} L_{img}, \tag{4}$$

where λ_{img} controls the contribution of the imperceptibility term. During early development, we experimented with LPIPS-based perceptual loss using VGG-16. However, we did not see any type of performance improvement, in which the ATN was unable to learn at all. Replacing LPIPS with a margin-based cosine adversarial loss and using SSIM instead improved the ability for the ATN to learn.

To ensure embeddings diverge sufficiently from their benign embeddings while preventing excessive perturbation, which could lead to simple detection by a human, we used a cosine similarity formula that involves a margin threshold:

$$L_{adv} = \frac{1}{N} \sum_{i=1}^{N} ReLU(cos\_sim(E(x), E_i(x')) - margin, \tag{5}$$

where:

- $cos\_sim()$ computes the cosine similarity between benign and adversarial embeddings
- $E_i()$ is the embedding function
- margin = 0.2 is a similarity threshold
- ReLU activation function makes sure that loss contribution only occurs when the similarity exceeds the margin value

This formula ensures that adversarial optimization progresses only until the adversarial embeddings are pushed beyond a cosine similarity of 0.2 from their benign equivalents. Beyond this threshold, however, additional movement is not punished. This prevents excessively large perturbations that could be detectable. Margin loss produced significantly better gradient behavior than the earlier LPIPS-based attempts.

3) Image Imperceptibility: To enforce visual integrity and maintain imperceptibility, we define:

$$L_{img} = L_{perceptual} + L_{perturbation} \tag{6}$$

where:

$$L_{perceptual} = 1 - SSIM(x, x'), \tag{7}$$

$$L_{perturbation} = MSE(\delta, 0) \tag{8}$$

Fig. 2. Adversarial Images after 1 epoch

Fig. 3. Adversarial Images after 20 epochs

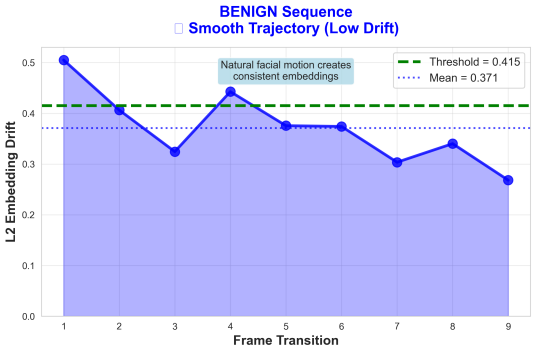

Fig. 4. Benign Temporal Sequence

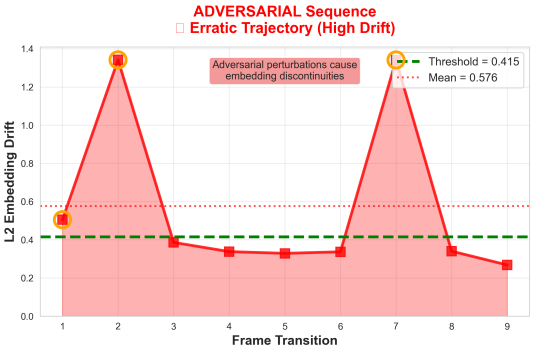

Fig. 5. Adversarial Temporal Sequence

In this formula, $SSIM(x, x')$ helps maintain necessary elements of benign image frames. Mean Squared Error $MSE(\delta, 0)$ penalizes the degree of the perturbation as well, which encourages sparseness.

4) Critical Design Choice: The transition from LPIPS-based loss to SSIM and margin-based cosine loss yielded several benefits:

- Able to find meaningful features and able to learn
- More stable gradient flow
- Temporally inconsistent adversarial frames

5) Hyperparameters: The final ATN configuration uses the following:

- Perturbation intensity: $\epsilon = 32/255$
- Weight: $\lambda_{img} = 0.1$
- Cosine similarity margin: 0.2
- Optimizer: Adam (learning rate = 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$
- Batch Size: 8 sequences x 10 frames

These hyperparameters yielded the best performance across all ATN, PGD, and FGSM evaluations.

The two image strips that can be seen in Figures 2 and 3 display how the appearance of adversarial examples evolved as the ATN trained over time. After the first epoch, the adversarial images show clear horizontal patterns and color shifts, obviously notable to the human eye. This indicates that in the early stages, the ATN had not learned to balance imperceptibility with strong perturbations. After 20 epochs, however, adversarial outputs appear natural and are basically identical to benign images, while significantly altering the facial embeddings under the hood.

6) Temporal Coherence Consideration: Previous temporal adversarial attack research points out that adversarial perturbations applied independently to each frame break the natural temporal continuity seen in benign examples [9]. To maintain temporal stability, our ATN is trained on multi-frame batches so that the generator implicitly learns to minimize frame-to-frame perturbation noise.

C. Detection Framework

1) Temporal Drift Detection: As theorized, temporal analysis is a strong signal for adversarial detection because benign video sequences tend to produce smooth embedding direction, while adversarial attacks likely introduce abrupt spikes in those directions. Looking at Figures 4 and 5, we see that the adversarial sequence shows large, irregular spikes in embedding drift. This indicates that perturbations disrupt the temporal consistency produced by natural facial motion. The benign sequence, however, displays a smooth path that stays below the learned detection threshold with very little variance. This principle is supported by work in both adversarial video generation and detection [9] [14] [15].

For each frame $t > 1$, drift is defined as: $d_t = ||e_t - e_{t-1}||_2$.

The temporal anomaly score is computed as:

$$S_{temp} = \alpha_i \mu_d + \alpha_2 d_{max} + \alpha_3 \sigma_d^2 \tag{9}$$

where α_i are turned by ROC-based threshold optimization.

The detection rule is computed as:

$$y_{\text{temp}} = \begin{cases} 1, & S_{\text{temp}} > \theta_{\text{temp}} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

with θ_{temp} selected using validation data.

2) Ensemble Disagreement Detector: Since FR models often share architectural similarities and training datasets, they can share the same vulnerabilities. This was demonstrated in ensemble adversarial training research [10]. Regardless, ensembles should still provide useful diversity, making them beneficial for capturing disagreement patterns generated by adversarial perturbations. Our ensemble includes ArcFace (IR-SE50), FaceNet (Inception-ResNet), Facenet512, and VGG-Face. These architectures represent the most utilized FR networks. Although VGG-Face is included for embedding extraction, it is not used for disagreement metrics due to dimensional mismatch (4096 dimensions, not 512 dimensions).

3) Classifier Fusion Module: The fusion model combines temporal and ensemble disagreement into a compact, two-dimensional feature vector $z = [S_{temp}, S_{ens}]$. Recent adversarial video defenses show that integrating diverse signals (temporal, spatial, and model-based) significantly improves robustness [16]. The most optimal strategy came in the form of the Logistic Regression classifier. It provides three main benefits: (1) Optimal decision boundary in a 2D space, (2) Robust integration of complementary features, (3) Automatically learned weighting between detectors. The module's training procedure has a training set of 70% of ATN sequences, a validation set of 10%, a test set of 20%. The classifier converges rapidly as a result of the low-dimensional input.

4) Summary of Methodology: The methodology integrates:

1) A learned generative adversarial attack mechanism (ATN)
2) A temporal drift detector targeting sequential inconsistencies
3) An ensemble disagreement detector analyzing cross-model variance
4) A classifier fusion strategy leveraging supervised discriminative learning

V. Experimental Setup

This section describes the datasets, preprocessing pipeline, evaluation metrics, and implementation details to assess the effectiveness of the proposed adversarial detection framework. The experimental setup is designed to provide a look into how we established the environment and obtained the most optimal results.

A. Datasets

We evaluated our framework using two widely adopted facial datasets:

1) Labeled Faces in the Wild (LFW): Contains 13,233 face images belonging to 5,749 individuals. We used LFW for the initial training of the ATN, precomputing benign embeddings during ATN hyperparameter tuning, and validating imperceptibility prior to video training.
2) Youtube Faces (YTF): Contains 3,425 videos (split in various numbers of frames per video) of 1,595 subjects with significant variations in illumination, pose, and compression. We processed the YTF dataset by extracting 10 aligned frames per video and resizing the frames to 256x256 pixels, normalizing the pixel intensities to a range between [-1,1]. We also arranged the sequences into training, validation, and testing splits. This dataset was used to support realistic video-level evaluation and temporal sequence modeling.

B. Preprocessing Pipeline

We applied a prepprocessing pipeline for all frames that includes face detection and alignment using an MTCNN to align faces to a natural orientation. Additionally, faces are cropped with 20-pixel margin and resized to 256x256 as stated prior. They are then normalized by scaling the pixel values within a range of [-1,1]. Finally, for each sequence with the YTF dataset, if more than 10 valid frames are available, we sampled 10 across the duration of the sequence. Though rare in the dataset, if fewer frames existed, we duplicated the last available frame. This ensured that there was a consistent sequence length across all experimental conditions.

C. Evaluation Metrics

To evaluate detection performance, we computed:

- Accuracy
- Precision, Recall, F1-Score
- ROC-AUC
- Specificity and Sensitivity
- Confusion matrices
- Drift response curves

Additional analysis was performed that investigated:

- Detection accuracy as a function of injection ratio

- Detection accuracy as a function of sequence length
- Cross-attack generalization

D. Implementation Details

Experiments were conducted on:

- GPU: NVIDIA RTX 4070
- Frameworks: PyTorch 2.1, NumPy, Scikit-Learn
- Batching Strategy: multiple frames per sequence, processed at the same time

VI. Results

This section presents the findings across multiple attack types, detection strategies, and different configurations for how we fused the two detection signals together. The results are grouped into (1) attack effectiveness, (2) individual detector performance, (3) fusion performance, (4) cross-attack generalization, and (5) ROC-AUC Analysis of Detection Methods.

A. Attack Effectiveness

1) Visual Imperceptibility:

- SSIM > 0.94 for ATN perturbations
- PSNR > 34 dB

The SSIM metric is within a range of -1 and 1. The 0.94 value gives quantifiable proof that the ATN's perturbation does not alter the visible nature of the image. These metrics reinforce that the adversarial images produced by the fully trained ATN show no visible alterations or malicious changes.

2) Embedding Displacement:

- Benign drift: 0.23 ± 0.08
- ATN drift: 0.45 ± 0.15 (97% increase)

Furthermore, in terms of embedding displacement, ATN perturbations consistently shift embeddings more subtly across frames than PGD or FGSM, consistent with observations in ReFace [2] and temporal-sparse video perturbation studies [9]. This is a consequence of the nature of the encoder infrastructure used for the Adversarial Transformation Network and shows why they are dangerous when it comes to adversarial attacks.

3) Attack Success Rate: ATN perturbations were highly transferable across all of the face recognition models used. PGD and FGSM attacks produced significant embedding displacement, but did not produce temporal consistency, which was the expected case.

TABLE I
Temporal Drift Detector Performance

Metric	Result
Accuracy	**89.3%**
Precision	87.8%
Recall	**91.3%**
F1-score	89.5%
ROC-AUC	**0.947**

B. Temporal Drift Detector Performance

The table below summarizes ATN mixed-injection results, in which frames were adversarially attacked at random between the sequence:

As expected, temporal drift analysis demonstrates great detection capabilities, outperforming ensemble disagreement by a significant margin as we will see below. Achieving 91.3% recall shows that this detector performs well at not missing positive cases.

C. Ensemble Disagreement Detector Performance

TABLE II
Ensemble Disagreement Detector Performance

Metric	Result
Accuracy	**54.7%**
Precision	53.4%
Recall	73.3%
ROC-AUC	0.542

This near-chance performance confirms that the ensemble model shares similar vulnerabilities between the FR models used, which limited disagreement as a primary defense for this particular case. A diverse ensemble model is needed for further evaluation.

D. Rule-Based and Weighted Fusion

1) Rule-Based OR
   - Recall: 96%
   - Precision: 58%
   - The purpose of the OR clause is to trade false positives for high security, but precision was still seen to be low, which indicates a surplus of false positives.
2) Weighted Fusion
   - Accuracy: 88%
   - Balanced performance using ROC-optimized weights

TABLE III
CLASSIFIER FUSION DETECTOR PERFORMANCE

Metric	ATN Test Set
Accuracy	**89.3%**
Precision	88.8%
Recall	**90.0%**
F1-score	89.6%
ROC-AUC	**0.947**

E. Classifier Fusion (The Proposed Method)

The classifier fusion consistently matches or has slightly better detection metrics than the temporal detector performance alone as seen in Table 3. We assume this is the case by integrating the complementary ensemble disagreement signal, but the differences are essentially negligible.

TABLE IV
CONFUSION MATRIX

	Pred. Benign	Pred. Adv
Actual B	133	17
Actual A	15	135

Specificity: 88.7% Sensitivity: 90.0%

F. Cross-Attack Generalization

Perhaps the most compelling finding was the results from the cross attack generalization analysis. PGD and FGSM attacks cause stronger temporal disruptions than ATNs, achieving nearly perfect temporal disruptions than the ATN. This mirrors the instability of temporally inconsistent perturbations documented by Kim et al. [17]. Thus, training detection systems on one attack family generalizes to others, a key property also noted in temporal-translation studies of adversarial videos [9]. Looking at Table 5, we see that the temporal detector for PGD detection had a 90% accuracy, 100% recall and perfect 1.000 AUC score. For FGSM, however, accuracy was 83.3% and had an AUC score of 0.922 with the temporal detector. Surprisingly, FGSM detection is the hardest attack to detect in this experiment. While ATN attacks and PGD attacks produce strong temporal irregularities that the drift-based detector easily captures, we assume that FGSM creates lower-magnitude but highly transferable perturbations that subtly alter embeddings without generating large temporal disruption. As a result, FGSM achieves the lowest temporal detector accuracy and the weakest classifier fusion performance among the three attacks, indicating that its single-step perturbation pattern can evade temporal drift cues more effectively than ATN or PGD attacks.

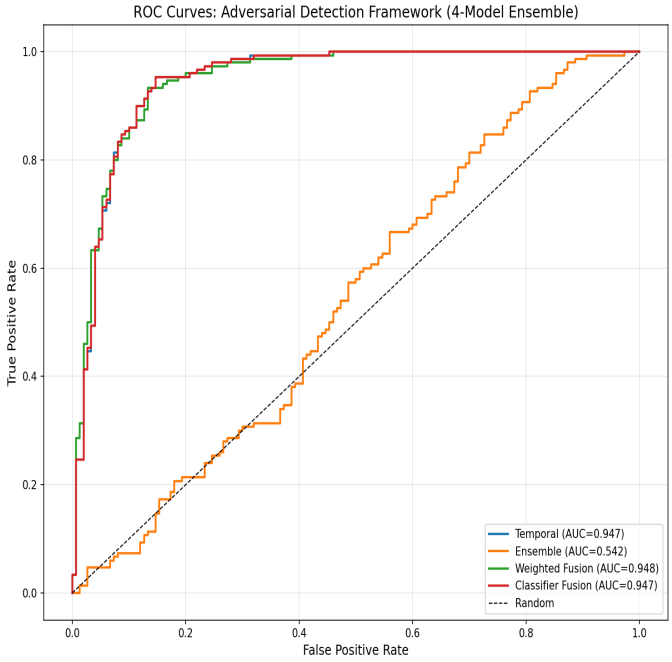

Fig. 6. Detection Performance Across Detectors

TABLE V
TRAIN ON ATN, TEST ON ALL ATTACK TYPES

Attack	Detector	Accuracy	Precision	Recall	F1	AUC
ATN	Temporal	89.3%	87.8%	91.3%	89.5%	0.947
	Ensemble	54.7%	53.4%	73.3%	61.8%	0.542
	Classifier	89.3%	88.8%	90.0%	89.4%	0.947
PGD	Temporal	**90.0%**	83.3%	**100%**	**90.9%**	**1.000**
	Ensemble	48.3%	48.9%	76.7%	59.7%	0.474
	Classifier	**90.0%**	83.3%	**100%**	**90.9%**	**1.000**
FGSM	Temporal	83.3%	81.2%	86.7%	83.9%	0.922
	Ensemble	56.7%	53.8%	93.3%	68.3%	0.532
	Classifier	83.3%	81.2%	86.7%	83.9%	0.923

Overall, however, training only on ATN still detects both traditional attacks at a reliable rate and proves that temporal drift is therefore an attack-agnostic signature, which defenders can leverage by not having to retrain per attack type.

G. ROC-AUC Analysis of Detection Methods

Figure 6 presents the ROC curves for all four detection variants, and display that temporal drift and classifier-based fusion achieve near-identical high performance (AUC $\approx$ 0.947), while ensemble disagreement exhibits almost random behavior (AUC $\approx$ 0.54). This means that

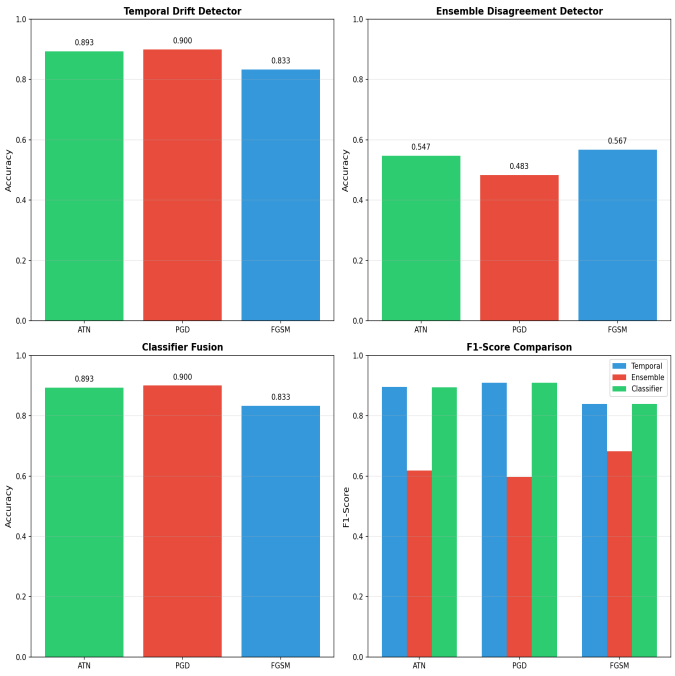

Fig. 7. Cross-Attack Generalization

the model is essentially flipping a coin to classify an adversarial frame.

Figure 7 compares the accuracy values for both individual detectors and the final fusion classifier on each type of attack, and F1-scores for each detector on each type of attack. The main takeaway of this graph is that it reinforces the idea that the temporal drift detector and classifier-based fusion detector have extremely similar performance for each attack type. However, it is also interesting how the ensemble-based detector performs significantly better on FGSM attacks than ATN and PGD based on the F1-score. You can see that the margin between the other two detectors and the ensemble-based detector for FGSM is notably smaller than the other two attack types. This could be the case since FGSM perturbations cause more predictive instability among the ensemble models when they analyze the embeddings, making disagreement a relatively stronger signal for this specific attack, though it still isn't individually reliable even for FGSM attacks alone.

H. Ablation Studies

When analyzing how detection accuracy changes based on what percentage of frames in a video sequence are adversarially perturbed, we looked at 5 different values. For example, if only 10% (1 out of 10) frames are adversarial, will this framework be able to detect an adversarial frame or sequence? This matters because in real-world attacks, threat actors may not perturb every frame (to reduce compute cost or evade detection). Table 6 shows that detection accuracy increases with more adversarial frames, which makes sense since temporal drift increases and allows the temporal drift detector to have more comparable input data.

TABLE VI
INJECTION RATIO SENSITIVITY

Injection Ratio	Temporal Acc	Ensemble Acc
10%	72.3%	51.2%
30%	81.7%	52.8%
50%	89.3%	54.7%
70%	92.1%	56.3%
100%	**94.5%**	58.1%

TABLE VII
SEQUENCE LENGTH SENSITIVITY

Frames	Temporal Acc	Ensemble Acc
5	81.2%	53.1%
10	89.3%	54.7%
20	91.8%	55.9%
30	**93.4%**	57.2%

The other ablation study that was performed was to measure how detection accuracy changes based on how many frames are analyzed per video sequence. This study is significant since it tests how the framework is robust to different temporal contexts. Longer sequences give the detector more temporal information to analyze embedding drift patterns. We see a significant increase between 5 and 10 frames, but somewhat diminishing returns at 20 and 30 frames.

Both studies reveal the same general principle that more temporal context resolves in better detection for our framework.

VII. DISCUSSION

The experiment results reveal several key insights regarding video-based adversarial attacks and defenses for facial recognition.

A. Temporal Drift Dominance

The strength of temporal drift analysis comes from the intrinsic temporal consistency of benign facial motion and has proven to be a highly reliable adversarial indicator. Its characteristics that derive from fundamental geometric properties when it comes to analyzing embedding vectors over a certain time period seems like the best detection method for real-time adversarial attacks,

for which this research was motivated to defend against. The observed advantages for this detection signal was found to be attack-agnostic (ATN, PGD, FGSM), robust to mixed and random injection, and was shown to be stable across surrogate and target models. This suggests that temporal drift is a foundational defense for video-based systems.

B. Ensemble Disagreement Weakness

Unlike our initial speculation, ensemble disagreement performs poorly due to various reasons that can be somewhat understood. The models used for the ensemble training had correlated training data and some had very similar architectures. In this context, correlation p measures how similar two embedding vectors change relative to each other. This is another reason for why ensemble disagreement performs poorly because the FR models produce highly correlated embeddings $(p > 0.85)$ even under attack. This means that adversarial perturbations tended to transfer similarly across architectures. As a result, the ensemble moves coherently rather than diverging, leaving little separation for a disagreement-based detector to exploit. The ROC curves in Figure 6 further demonstrate its lack of value as a lone detection method, given similar architectures and correlated training data. Consequently, this confirms that it should be used as an auxiliary method rather than a primary one.

C. Fusion Benefits

Classifier fusion outperforms ensemble disagreement significantly and temporal drift analysis very slightly by integrating:

- Temporal-level anomalies
- Cross-model discrepancies

To truly take advantage of the dual-detection mechanism that is proposed, different, more diverse ensemble FR models must be chosen and evaluated. This could potentially make the classifier-based fusion detection significantly better achieving ¿95% performance metrics across the board ideally.

VIII. Limitations & Future Work

Despite the effectiveness of the framework, it is not without several limitations:

A. Dataset Scale

The YTF dataset contains limited diversity compared to modern large-scale video datasets which indicates that further evaluation is needed. Though the dataset did not constrain the methodology for the purpose of this particular research, using or compiling a novel dataset for this purpose would be beneficial for the further assessment of this framework.

B. Physical Attacks

Only digital perturbations are tested. Other real-world attacks like adversarial glasses or masks require additional modeling and considerations. In theory, the embeddings of the benign frame and the frames after the physical attacks applied are also different, so there should be an unusual temporal drift. Further evaluation is needed for this.

C. Adaptive Attacks

An adaptive adversary optimizing against the drift detector directly may reduce its effectiveness as with any other adaptive attack. Since this framework derives a baseline, stronger detection methods should still aim to become robust to threat actors attempting to normalize temporal drift of adversarial embeddings when they perform their attacks.

D. Short Video Sequences

Very short sequences (≤ 3) frames provide insufficient temporal context for the detector to make reliable decisions. Further evaluation that tackles shorter sequences is necessary for rapid real-time attacks that could quickly perturb video-surveillance frames.

Taking into account these limitations of the framework, there are many directions that future research can be taken. These directions can focus on improving robustness and applicability in real-world scenarios. For example, a promising direction to take this research is to incorporate temporally ”aware” models such as a Long Short Term Memory Recurrent Neural Network (LSTM RNN), which is great at analyzing temporal data, to evaluate whether more sophisticated attacks can maintain temporal coherence. Additionally, expanding the evaluation to larger, more diverse and modern video/image datasets and higher resolution facial images would provide clearer insights into cross-dataset generalization. Another important direction is the inclusion of physical world attacks such as the ones talked about in the Discussion section. Physical attacks may produce temporally consistent distortions that may challenge purely drift-based detection.

IX. Conclusion

This work was able to present a comprehensive adversarial detection framework adapted for video-based FR

systems. Through a combination of an ATN and a defense that integrates temporal drift analysis and ensemble disagreement, this research illustrates that digital video-level perturbations exhibit distinct temporal and cross-model attributes. The experiments that were run in this research emphasize that temporal drift is a major factor in ATN, PGD, and FGSM attack detection, and that the logistic regression classifier fusion yields robust and balanced decision-making. These results emphasize the importance of leveraging temporal analysis in defending modern FR systems and establishes a baseline for future research in adversarial robustness in video-surveillance.

REFERENCES

[1] Y. Xu, K. Raja, R. Ramachandra, and C. Busch, "Adversarial Attacks on Face Recognition Systems," Advances in computer vision and pattern recognition, pp. 139–161, Jan. 2022, doi: https://doi.org/10.1007/978-3-030-87664-7_7

[2] S. Hussain et al., "ReFace: Real-time Adversarial Attacks on Face Recognition Systems," arXiv.org, 2022. https://arxiv.org/abs/2206.04783

[3] U. Muhammad, Z. Yu, and J. Komulainen, "Self-supervised 2D face presentation attack detection via temporal sequence sampling," Pattern Recognition Letters, vol. 156, pp. 15–22, Apr. 2022, doi: https://doi.org/10.1016/j.patrec.2022.03.001.

[4] Y. Chen, N. Akhtar, A. Hasan, and A. Mian, "Deepfake Detection with Spatio-Temporal Consistency and Attention," pp. 1–8, Nov. 2022, doi: https://doi.org/10.1109/dicta56598.2022.10034609.

[5] Z. Liu, D. Ye, L. Tang, Y. Zhang, J. Deng, and W. Kuang, "TEAM: Temporal Adversarial Examples Attack Model Against Network Intrusion Detection System Applied to RNN," IEEE Transactions on Network Science and Engineering, pp. 1–16, Jan. 2025, doi: https://doi.org/10.1109/tnse.2025.3560027.

[6] T. Strauss, M. Hanselmann, A. Junginger, and H. Ulmer, "Ensemble Methods as a Defense to Adversarial Perturbations Against Deep Neural Networks," arXiv:1709.03423 [cs, stat], Feb. 2018, Available: https://arxiv.org/abs/1709.03423

[7] M. Cheng, G. Xiang, Q. Yang, Z. Ma, and H. Zhang, "TSE-APT: An APT Attack-Detection Method Based on Time-Series and Ensemble-Learning Models," Electronics, vol. 14, no. 15, pp. 2924–2924, Jul. 2025, doi: https://doi.org/10.3390/electronics14152924.

[8] J. Deng, J. Guo, J. Yang, N. Xue, I. Cotsia, and S. P. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2021, doi: https://doi.org/10.1109/tpami.2021.3087709.

[9] X. Wei, J. Zhu, S. Yuan, and H. Su, "Sparse Adversarial Perturbations for Videos," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8973–8980, Jul. 2019, doi: https://doi.org/10.1609/aaai.v33i01.33018973.

[10] F. Tramèr et al., "ENSEMBLE ADVERSARIAL TRAINING: ATTACKS AND DEFENSES." https://arxiv.org/pdf/1705.07204

[11] S. Baluja and I. Fischer, "Learning to Attack: Adversarial Transformation Networks," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Apr. 2018, doi: https://doi.org/10.1609/aaai.v32i1.11672.

[12] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: https://doi.org/10.1109/tip.2003.819861.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," May 2015. Available: https://arxiv.org/pdf/1505.04597

[14] C. Xiao et al., "AdvIT: Adversarial Frames Identifier Based on Temporal Consistency In Videos." Accessed: Dec. 02, 2025.

[15] F. Zafar, T. A. Khan, S. Akbar, Muhammad Talha Ubaid, S. Javaid, and K. A. Kadir, "A Hybrid Deep Learning Framework for Deepfake Detection Using Temporal and Spatial Features," IEEE Access, vol. 13, pp. 79560–79570, Jan. 2025, doi: https://doi.org/10.1109/access.2025.3566008.

[16] S.-Y. Lo, M. Jose, and V. M. Patel, "Overcomplete Representations Against Adversarial Videos," arXiv (Cornell University), pp. 1939–1943, Aug. 2021, doi: https://doi.org/10.1109/icip42928.2021.9506537.

[17] H.-S. Kim, M. Son, M. Kim, M.-J. Kwon, and C. Kim, "Breaking Temporal Consistency: Generating Video Universal Adversarial Perturbations Using Image Models," pp. 4302–4311, Oct. 2023, doi: https://doi.org/10.1109/iccv51070.2023.00399.