IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 August 2025, accepted 18 September 2025, date of publication 25 September 2025, date of current version 2 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3614455

RESEARCH ARTICLE

EVDO: An Enhanced Framework for Deepfake Detection in Videos Through Optical Flow and Temporal-Spatial Analysis

MOHAMMED MOUAD MELOUK[1], MINGHAO SHAO[1,2], ABDUL BASIT[1], CHAIMAE ABOUZAHIR[1], RUILIN ZHOU[1], AND MUHAMMAD SHAFIQUE[1], (Senior Member, IEEE)

[1]eBRAIN Laboratory, Division of Engineering, New York University (NYU) Abu Dhabi, Abu Dhabi, United Arab Emirates
[2]New York University Tandon School of Engineering, New York University, New York, NY 10012, USA

Corresponding author: Minghao Shao (shao.minghao@nyu.edu)

This work was supported in part by the NYU Abu Dhabi (NYUAD) Center for CyberSecurity (CCS), funded by Tamkeen through the NYUAD Research Institute under Award G1104; and in part by the NYUAD High Performance Computing (HPC) Center for providing Necessary Compute Resources for the Experiments.

ABSTRACT As generative AI advances, the realism of synthetic media has sparked serious concerns in security, privacy, and misinformation. This issue is particularly concerning with the rise of deepfake technologies that manipulate facial imagery, undermining media authenticity. While existing research has largely focused on image-based deepfake detection with specialized feature extractors, detecting deepfakes in video, especially with broad generalization across varied manipulation techniques, remains a substantial challenge. This paper introduces EVDO, a novel framework designed to detect deepfake videos by integrating temporal and spatial analysis for enhanced detection accuracy. Our approach leverages optical flow techniques to capture subtle temporal manipulation artifacts between video frames overlooked in spatial analysis. Using a FlowFormer++ model for temporal analysis, frame pairs are sampled to produce cost volumes that highlight essential motion regions and capture manipulation-specific artifacts through optical flow images which encode temporal dependencies. A flow-finetuned detector then extracts flow-level features indicative of deepfake manipulation. Complementing this, Xception-based spatial detectors analyze each frame individually, generating high-dimensional embeddings that capture frame-specific anomalies. Fusing these temporal and spatial embeddings enables comprehensive binary classification of deepfakes. Validated on the FaceForensics++ benchmark, EVDO significantly improves generalization. Our proposed temporal path contributes correct classifications to around 12% of datapoints. EVDO achieves a video AUC of 99.13% (99.43% of end-to-end SOTA methods) while enabling forensically verifiable manipulation detection.

INDEX TERMS Deepfake detection, temporal-spatial analysis, optical flow, deep-learning.

I. INTRODUCTION

The rapid advancements in generative AI technologies have enabled the creation of highly realistic synthetic media, notably deepfake videos. These developments, rooted in sophisticated generative models like GANs [1], [2], [3] and advancements in neural network architectures [4], [5], [6], have spurred serious concerns across security, privacy, and misinformation domains. For instance, in 2023, approximately 500,000 video and voice deepfakes were shared on social media worldwide, highlighting the widespread dissemination of such content [7]. This widespread dissemination brings increased potential for misuse; recent data shows a 3,000% rise in attempts to use deepfakes to bypass security verifications in 2024 [8]. Recent statistics reveal that 26% of smaller and 38% of large companies experienced deepfake fraud in the past year, with losses reaching up to $480,000. [9]. The exponential growth in the use of deepfakes

The associate editor coordinating the review of this manuscript and approving it for publication was Rongbo Zhu.

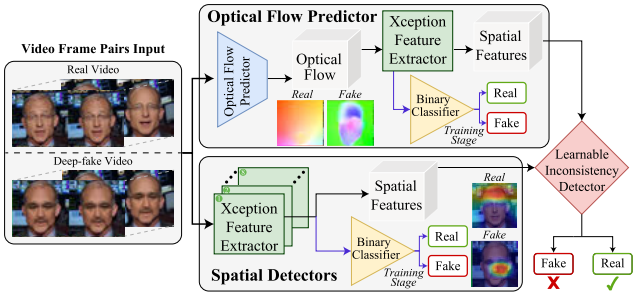

FIGURE 1. The EVDO framework processes videos through two branches: one extracts spatial features with Xception detectors across frames, and the other captures temporal dynamics via optical flow; both branches are then fused with a learnable inconsistency detector to enhance accuracy in deepfake identification. For each branch gradient-weighted class activation maps are shown.

underscores the urgent need for reliable detection methods to prevent their adverse impacts.

A. RESEARCH CHALLENGES

The key research problem addressed in this paper is the development of a robust and generalized method for detecting deepfake videos, specifically those manipulating facial imagery to produce highly realistic yet fraudulent content. Deepfakes pose significant threats to the authenticity of digital media, as they can deceive viewers with realistic facial expressions and emotions [10], [11]. The proliferation of manipulated content on social media platforms amplifies this problem, leading to potential misuse in cybercriminal activities, reputation attacks, and disinformation campaigns. Therefore, creating effective deepfake detection methods is crucial for safeguarding digital media integrity and preventing malicious misuse [12].

Current state-of-the-art deepfake detection methods can be broadly categorized into spatial analysis models and temporal analysis models:

- **Spatial Analysis Models:** Techniques like XceptionNet [13] and MesoNet [14] focus on identifying inconsistencies within individual frames by utilizing specialized feature extractors. These models excel at capturing spatial artifacts introduced during the manipulation process, achieving high accuracy on static images. However, they fail to consider temporal features, not optimally generalizing to video data.
- **Temporal Analysis Models:** Approaches such as CNN+LSTM architectures [15] and recurrent neural networks attempt to capture motion patterns across frames to detect temporal anomalies. While these models address some limitations of spatial-only methods, they struggle with generalizing across various manipulation techniques, such as GAN-based and autoencoder-based methods [16]. Additionally, they may overlook transition-specific artifacts that may be infrequent at the frame level yet essential for better detection.

Recent works like AltFreezing [17], Dynamic Difference Learning [18], Style Latent Flows [19], and Temporal Coherence Fusion [20] combine spatial and temporal analysis. However, these methods suffer from critical limitations: AltFreezing [17] uses temporal attention without explicit motion modeling, Dynamic Difference Learning [18] struggles with subtle motion artifacts due to fixed spatio-temporal correlation, Style Latent Flows [19] overlook motion inconsistencies by focusing on style transfer, and Temporal Coherence Fusion [20] prioritizes global coherence over frame-specific anomalies. In contrast, our EVDO framework uniquely integrates FlowFormer++ for optical flow-based motion estimation and masked cost volume autoencoding to recover localized manipulation artifacts, implementing precise temporal analysis complementary to Xception-based spatial detection.

A major limitation of current state-of-the-art methods is their inability to capture subtle motion artifacts specific to manipulated content, which often appear as temporal inconsistencies between frames. Our preliminary experiments reveal that existing spatial-only approaches frequently miss these artifacts, especially in cases where frame-by-frame manipulations are imperceptible. This highlights a gap in the literature and stresses the need for a model that combines spatial and temporal analyses to reliably detect deepfake videos.

In this paper, we propose EVDO, a framework for detecting deepfake videos that integrates both *temporal* and *spatial* analysis. Unlike prior works, our approach uniquely leverages *optical flow techniques* with FlowFormer++ [21] to quantify object movement between frames, capturing subtle manipulation artifacts through masked cost volume autoencoding. This recovers precise motion regions corrupted by deepfakes. Complementary to this, Xception-based spatial detectors extract frame-level anomalies, while our fusion mechanism combines temporal and spatial embeddings via a learnable inconsistency detector (Fig. 1), addressing limitations of existing hybrid approaches.

Our novel contributions in this paper are as follows:

1) *Integration of Transformer-Based Optical Flow:* We introduce a temporal analysis approach using FlowFormer++ to generate optical flow images via masked tokenization, explicitly recovering motion artifacts overlooked by prior spatio-temporal methods [17], [20].
2) *Complementary Spatial Detection:* We enhance generalization by extracting high-dimensional embeddings that capture frame-specific anomalies, avoiding the over-smoothing issues of attention-based fusion in [18].
3) *Optical Flow Dataset for Deepfake Detection:* We generate a dedicated dataset of optical flow images using FlowFormer++, providing motion-specific representations critical for temporal analysis.
4) *Learnable Inconsistency Fusion:* Our fusion technique dynamically combines spatial and temporal

embeddings, and hence outperforming static aggregation methods in [19].

Our proposed framework, EVDO, has been extensively on multiple benchmark datasets: Face2Face [22], FaceSwap [23], and Neural Textures [24], from the expanded FaceForensics++ [16] dataset. Experimental results demonstrate state-of-the-art performance, with ablation studies validating the significance of temporal features (contributing 12.27% of the samples to accuracy). While cross-dataset generalization (e.g., Celeb-DF [25], DFDC [26]) is reserved for future investigations (not in the scope of this paper), our framework establishes a robust foundation for detecting diverse manipulation techniques.

B. PAPER ORGANIZATION

Section II discusses related work. Section III presents the EVDO framework in further detail. Section IV details the experimental setup and Section V details the results, followed by limitations of our work in Section VI. We conclude in Section VII.

II. RELATED WORK

A. DEEP FAKE TECHNOLOGY

The generation of deepfakes primarily utilizes advanced machine learning models to synthesize hyper-realistic images and videos. Key among these are Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and more recently, Diffusion Models, each contributing uniquely to the field of synthetic media creation.

1) DEEPFAKE GENERATION

VAEs [34] are foundational in deepfake generation, employing a probabilistic framework to encode input images into a latent space which can then be sampled to generate new images. The Conditional VAE (CVAE) [35] and Vector Quantized VAE (VQ-VAE) [36] are notable enhancements improving the diversity and clarity of generated outputs. GANs [1] have significantly influenced deepfake technology by pitting two neural networks against each other: a generator that creates images and a discriminator that evaluates them. This setup helps in refining the generated images to a high degree of realism. CVAE-GANs [37] and multi-domain models like StarGAN [38] further tailor the GAN framework for specific applications such as fine-grained image generation [39] and cross-domain image translation [40]. Earlier works such as [41] use convolutional neural networks for deepfake generation. Emerging as a powerful alternative, diffusion models generate data by learning to reverse a gradual process of adding noise to real images. These models have shown promising results in generating high-quality, detailed images by effectively modeling the data distribution [42], [43]. Reference [44] aims at generating a talking video with a character in the target image engaging in a conversation based on an arbitrary driving source.

2) MANIPULATION APPROACHES

Manipulation techniques in deepfake technology involve altering facial features and expressions to create convincing fake media that can be nearly indistinguishable from genuine content. Techniques include face swapping, facial reenactment, and attribute editing, each catering to specific aspects of face manipulation. Face Swapping [45], [46] involves replacing the face of a person in a video with another face, for realistic features and expressions. Advances in face swapping have been driven by neural networks that can efficiently learn mappings between different facial structures, maintaining the realism of the swapped faces. Another approach used for manipulation-based deepfake is face reenactment [22], [47], [48]. Real-time systems like Face2Face transfer the expressions of a source actor to a target video in real-time, preserving the identity and photorealism of the target. This technology has broad implications for entertainment and communications. Facial Attribute Editing such as AttGAN [49] and SMGAN [50] focus on editing specific facial attributes like age, gender, or expression while retaining the overall coherence of the original face, allowing for controlled manipulation of facial characteristics.

3) SECURITY AND PRIVACY CONCERNS OF DEEPFAKE

Deepfake technology poses serious risks, including the erosion of media trust, privacy threats, and legal challenges. Deepfakes blur the line between real and synthetic media, leading to ''reality apathy'' and reduced public engagement in societal issues [51]. This technology enables identity theft, privacy violations, and unauthorized use of likenesses, causing reputational harm and personal distress [52]. Additionally, legal frameworks must adapt to address defamation, consent, and intellectual property issues associated with AI-generated content [53].

B. DEEP FAKE DETECTION

Deepfake detection methods have evolved alongside the technology. Early approaches exploited physical inconsistencies—such as facial discrepancies, face-swapping artifacts [54], user stream descriptors [55], and irregular eye-blinking patterns [27]—but as handcrafted features proved inadequate, deep learning techniques emerged to harness spatial and temporal cues. For example, CNN-RNN models capture irregular eye blinking [56], while other methods extract temporal inconsistencies across frames [28].

Recent innovations include physiological-based detection (e.g., predicting heart rate variations [57]) and basic CNN classifiers for binary deepfake detection [13], [14]. Spatial detectors now target manipulated image regions using techniques like capsule networks [29] and blending artifact analysis [30]. Additionally, methods leveraging universal manipulation features [31], [32] and frequency-based analyses that reveal high-frequency artifacts [33], [58] further enhance robustness.

Hybrid approaches combine spatial and temporal analyses. For instance, AltFreezing [17] employs temporal

TABLE 1. Comparison to other SOTA methods. Where available, performance metrics are reported after training and testing on the benchmark FaceForensics++ HQ (c23) [16].

Method	Model Type	Spatial Aware	Temporal Aware	Frame Acc	Video Acc	Video AUC	Attributes
[27]	Handcrafted features	Yes	Partial	-	87.5	-	eye-blinking patterns; irregular blinking.
[28]	Temporal analysis	Yes	Partial	-	98.26	-	inconsistencies across successive frames.
[13]	Xception	Yes	No	-	96.37	-	binary frame classification.
[14]	Meso4	Yes	No	89.1	96.9	-	binary mesoscopic frame classification.
[29]	Capsule Networks	Yes	No	95.93	99.23	-	hierarchical spatial relationships.
[30]	CNN-based	Yes	No	97.73	-	-	face-to-background transitions.
[31]	Feature-based	Yes	No	88.44	-	-	universal features of GANs, VAEs.
[32]	Feature-based	Yes	No	97.05	-	-	universal features of GANs, VAEs.
[33]	Frequency CNN	No	No	93.02	-	-	Fourier spectrum analysis.
[17]	Hybrid CNN-Transformer	Yes	Yes	-	-	96.7	temporal attention; no explicit motion modeling.
[20]	FTCN & Transformer	Yes	Yes	-	-	99.7	end-to-end; no explainability.
[18]	Spatio-temporal CNN	Yes	Yes	95.03	95.2	96.77	spatio-temporal differences; difficulty with subtle motion artifacts.
[19]	Style-based CNN	Yes	Yes	-	-	98.4	style flow for cross-frame consistency; no motion artifact modeling.
EVDO (ours)	Hybrid CNN-Transformer	Yes	Yes	95.45	98.97	99.13	CNNs for frequency, spatial and naïve features; optical flow to achieve verifiable deepfake detection; balanced, generalized metrics

attention without explicit motion modeling, Zheng et al. [20] focus on temporal coherence while overlooking frame-specific anomalies, dynamic difference learning [18] captures spatio-temporal correlations but struggles with subtle motion artifacts, and Choi et al. [19] use style flows yet ignore motion inconsistencies. In contrast, EVDO uniquely integrates transformer-based optical flow (FlowFormer++ [21]) with masked cost volume autoencoding to recover motion artifacts, fusing spatial and temporal embeddings via a learnable inconsistency detector (Fig. 2).

1) OPTICAL FLOW vs. FRAME DIFFERENCING

Unlike [17], [20] that use frame differencing or attention, EVDO leverages FlowFormer++ [21] to explicitly model motion via optical flow, capturing subtle artifacts in corrupted motion regions.

2) ADAPTIVE FUSION

Prior works [18], [19] use static concatenation or style-based fusion, whereas EVDO employs a learnable inconsistency detector to dynamically weight spatial-temporal discrepancies.

3) ARTIFACT LOCALIZATION

Methods like [18] focus on global temporal coherence, while EVDO's masked cost volume autoencoding localizes manipulation-specific motion artifacts.

C. OPTICAL FLOW

Optical flow [59] is vital for understanding motion in image sequences, underpinning tasks like object tracking and scene reconstruction. The computational model in [60] estimates apparent motion by minimizing a global energy function that enforces brightness constancy and smoothness, while Farnebäck's algorithm [61] employs polynomial expansion for efficient real-time motion estimation.

Integrating deep learning has significantly advanced motion estimation. FlowNet [62] pioneered the use of deep convolutional networks for optical flow, moving beyond handcrafted features. LiteFlowNet [63] further reduced computational complexity with a lightweight architecture featuring multi-scale flow inference and feature warping, making it suitable for devices with limited resources. Transformer-based models such as FlowFormer [64] leverage self-attention to handle large displacements and occlusions by effectively modeling global context. Building on FlowFormer++, EVDO uses masked cost volume autoencoding to recover motion regions, specifically addressing deepfake-induced corruptions—a novel application not explored in previous optical flow frameworks.

D. BENCHMARKING DEEPFAKE

To accurately assess the capabilities and limitations of deepfake detection methods, it is essential to establish robust benchmarks. The development of modern benchmarks to evaluate models' ability to detect deepfakes relies on several aspects. Works such as those based on face-swapping algorithms— [25], [26], [65], [66], [67], [68] form the foundation of this research area. Additionally, models like those developed by Thies et al. [22], [24], also integrated into the standard face-reenactment based deepfake benchmark dataset [16], are pivotal. Yan et al. [69] proposed a comprehensive benchmark for deepfake detection that evaluates approaches based on naïve, spatial, and frequency methods. While EVDO is validated on FaceForensics++ and related benchmarks, future work will extend evaluations to cross-dataset settings (e.g., Celeb-DF [25], DFDC [26]) to assess generalization. These benchmarks are vital for evaluating the current state of the art and guiding future developments towards creating more effective and resilient deepfake detection systems.

III. EVDO METHODOLOGY

This section provides the outline of our proposed EVDO framework, detailing the stages of our approach from data preparation to model training and evaluation. The methodology is designed to leverage both spatial and temporal features by generating optical flow and spatial embeddings

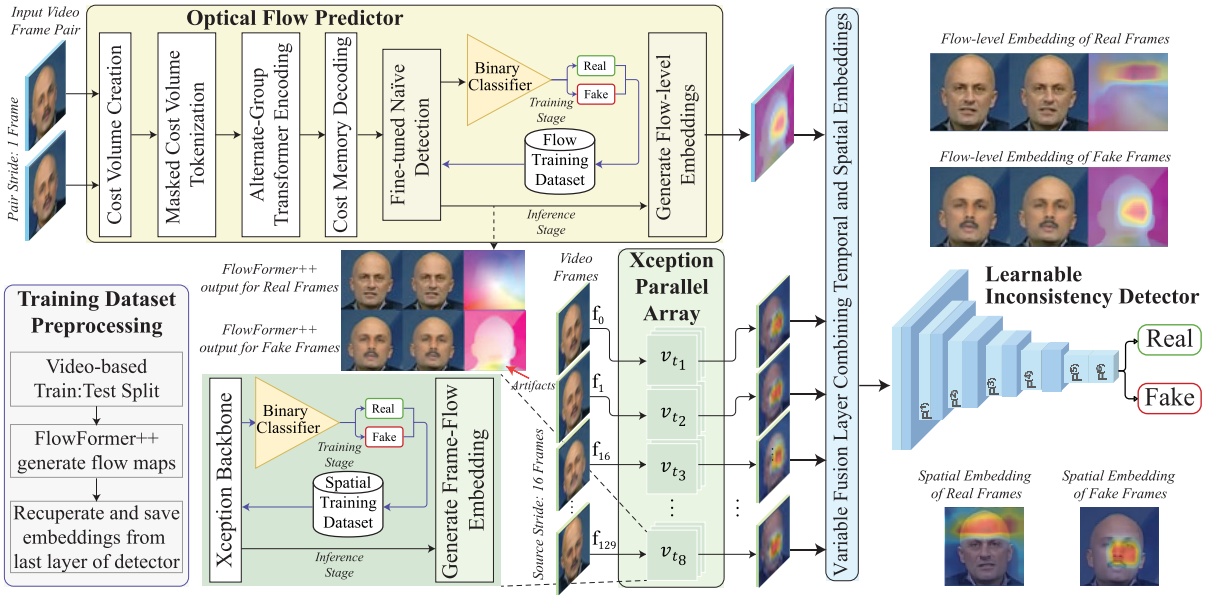

FIGURE 2. Overview of the EVDO methodology: The optical flow predictor processes video frame pairs to capture temporal manipulation artifacts, while the Xception-based spatial detector analyzes individual frames for anomalies. Flow-level and frame-level embeddings are generated, fused, and classified through a fully connected classifier to distinguish real from fake videos. The outputs from both branches are integrated in a feature fusion module, which includes a learnable inconsistency detector. This module leverages both the spatial and temporal data to generate the final predictions, effectively enhancing the system's ability to identify discrepancies and improve overall accuracy. Training dataset preprocessing ensures optimized feature extraction and classifier accuracy.

to capture manipulation-specific artifacts. Although existing works [17], [20], [70], [71], [72] have explored using optical flow in deepfake detection, our approach uniquely integrates (i) *transformer-based optical flow estimation* (FlowFormer++) rather than a simpler flow computation pipeline, (ii) *multiple specialized detectors* (Xception, UCF, and SPSL) trained on each of frames and flow, and (iii) a *learnable inconsistency detector* to dynamically fuse the resulting embeddings. This design moves beyond the simplistic temporal attention or frame differencing used in prior hybrid approaches.

A. OPTICAL FLOW GENERATION

To capture subtle temporal inconsistencies, we sample frame pairs from each video at set strides and pass them through FlowFormer++ [21] to generate optical flow images that encode motion and manipulative artifacts. FlowFormer++ uses *masked cost volume autoencoding* (MCVA) to precisely recover motion regions corrupted by deepfakes—outperforming traditional methods [62], [63] and other flow-based detectors [70], [71], [72]. These optical flow images form the basis for our temporal analysis, tracking atypical movements and exposing inconsistencies that static spatial analysis might miss, especially subtle, often imperceptible, between-frame artifacts.

Given two consecutive frames, I_t and I_{t+1}, FlowFormer++ first extracts dense feature maps via a pre-trained encoder E, yielding $F_t = E(I_t)$ and $F_{t+1} = E(I_{t+1})$. A 4D cost volume C is then constructed, where each element measures feature similarity between pixels:

$$C_{ijkl} = \text{similarity}\big(F_t(j,k),\ F_{t+1}(l,m)\big),$$

with C_{ijkl} representing the similarity score between pixel (j,k) in F_t and pixel (l,m) in F_{t+1}. MCVA then recovers masked regions in C, predicting the optical flow as $\hat{C} = MCVA(C, C_{mask})$.

In EVDO, the Sintel-pretrained parameters of FlowFormer++ are frozen. A CNN-based detector $D(\cdot)$ is subsequently applied to the predicted optical flow to produce a flow-level embedding. This embedding is generated via:

$$\hat{\phi}_{FFpp} = \mathcal{L}\big(D'(\mathbf{f}_{t\to t+1}),\ y_{I_t}\big),$$

where y_{I_t} is the ground truth for frame I_t, $D'(\mathbf{f}_{t\to t+1})$ is the detector output, and $\mathcal{L}(\cdot)$ denotes the binary cross entropy loss.

B. DETECTOR MODEL DESIGN AND PREPROCESSING

1) XCEPTION DETECTOR

The Xception network comprises three flows: the Entry Flow extracts low-level features and reduces spatial dimensions; the Middle Flow refines features via depthwise separable convolutions; and the Exit Flow generates high-level features through global average pooling (GAP). The GAP layer yields a compact, spatially invariant embedding:

$$\hat{\phi}_{Xcep} = \text{GAP}\big(f_{\text{Exit Flow}}\big).$$

Before GAP, the Exit Flow computes:

$$y_{\text{exit}} = \text{DWConv}(x_{\text{mid}}) + \text{MaxPool}(x_{\text{mid}}).$$

This embedding is employed for downstream deepfake detection [13].

2) UCF DETECTOR

The UCF Detector extracts three types of features from images: content (c), forgery-specific (f_s), and common (f_c). Its decoder reconstructs images by combining c and f via Adaptive Instance Normalization (AdaIN), while classification heads predict forgery types (f_s) and real/fake labels (f_c) using cross-entropy loss. A reconstruction loss ensures pixel-level consistency, and a contrastive loss maximizes intra-class similarity while enhancing inter-class disparity. The overall loss is:

$$\mathcal{L}_{\text{UCF}} = \lambda_1 \mathcal{L}_{\text{classification}} + \lambda_2 \mathcal{L}_{\text{reconstruction}} + \lambda_3 \mathcal{L}_{\text{contrastive}}.$$

We extract the embedding $\hat{\phi}_{UCF}$ from the common feature extractor (f_c), capturing generalizable cues across diverse manipulations. With an Xception backbone, this detector provides rich high-dimensional spatial embeddings [32].

3) SPSL DETECTOR

The SPSL Detector leverages the phase spectrum to expose up-sampling artifacts common in deepfakes. The phase spectrum $P(u)$ is derived from the DFT:

$$P(u) = \arg\big(X(u)\big), \quad \text{with } X(u) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi un/N}.$$

To integrate spatial and frequency features, SPSL converts $P(u)$ to its spatial form via the inverse DFT:

$$p(n) = \mathcal{F}^{-1}\big(P(u)\big), \quad y(n) = \big[x(n),\, p(n)\big],$$

and applies a shallow convolution:

$$y_A(n) = x(n) * c(n),$$

where $c(n)$ is a small kernel that focuses on localized texture clues. This approach enhances micro-level artifact detection and improves cross-dataset generalizability. The embedding $\hat{\phi}_{SPSL}$ is extracted from the RGBP feature maps before classification. With an Xception backbone, SPSL yields frequency-rich, high-dimensional embeddings that capture anomalies often missed by spatial-only detectors [73].

C. EMBEDDING GENERATION AND DATASET PREPARATION

To generate frame-level embeddings, each of the above detectors is fine-tuned on the published FaceForensics++ training set. To generate flow-level embeddings, the same detectors are further finetuned on the flow images produced by FlowFormer++. Thus, each detector yields two specialized versions: one focusing on spatial embeddings, the other on temporal (flow) embeddings.

Our dataset generation process:

- **Data Splitting:** We employ the original *video-based* train:test split for both the frame and flow detectors, preserving consistency with prior research.
- **Optical Flow Generation:** Using FlowFormer++ (Sintel weights), we set *Pair Stride* to 1 and *Source Stride* to 16, generating an optical flow image for each consecutive frame pair in our main experiments.
- **Frame Embeddings:** From each fine-tuned, ImageNet-pretrained frame detector, we extract and store embeddings of every frame in the train and test sets.
- **Flow Embeddings:** Similarly, from each fine-tuned flow detector, we extract and store embeddings of every flow image in the train and test sets.
- **Embedding Concatenation:** We concatenate frame and flow embeddings for each video to create a unified representation, yielding exactly one data point per video in the final dataset.
- **Classifier Training:** A fully-connected learnable inconsistency detector (LID) is then trained on these fused embeddings, following the same train:test partition and using a standard binary cross-entropy loss.

We keep the organization such that each video has a designated folder of frames, flows, and corresponding embeddings, ensuring efficient data loading during training and inference.

D. DATA LOADING AND TRAINING OF FULLY CONNECTED CLASSIFIER

We build a custom data loader to handle the two-fold input of *flow* and *frame* embeddings. The loader fetches the embeddings from their respective folders and concatenates them, producing a joint representation of spatial and temporal features. With this dual-embedding setup, we train a LID from scratch to distinguish real from fake videos, effectively fusing spatial-frequency artifacts with motion cues to capture subtle manipulations across the detectors' domains.

Mathematically, we combine the optical flow embeddings $\hat{\phi}_{FFpp}$ with the spatial embeddings y_A: $\mathcal{F} = \hat{\phi} \oplus y_A$, where $\oplus$ is channel-wise concatenation, y_A is $\hat{\phi}_{Xcep}$, $\hat{\phi}_{SPSL}$ or $\hat{\phi}_{UCF}$. This output $\mathcal{F}$ is fed into LID that adaptively weighs spatial-temporal discrepancies. Unlike the static fusion in [19], our detector dynamically captures how forged content manifests differently in RGB vs. flow domains with broad pair and source frame strides, benchmarked across 4 deepfake manipulation methods.

E. LOSS FUNCTION AND EVALUATION METRICS

We employ a binary cross-entropy loss in all classification stages, for consistent training of each detector and LID. Our core evaluation metrics include:

- **Pair-Level Accuracy:** We first assess pair-by-pair accuracy to see how reliably individual frame pairs are labeled real or fake by each model sub-path.
- **Video-Level Aggregated Accuracy:** Per-video, we evaluate an array of LIDs with 1, 2, 3 then 4 frame-pair embeddings to yield a confidence score, evaluating overall consistency of detection across the entire video as the number of frame-pairs changes.

- **Area Under Curve (AUC):** We monitor AUC to select the best checkpoint, as it measures discriminative capacity over different classification thresholds.

F. BINARY CLASSIFIER AND DECISION AGGREGATION FOR FINAL OUTPUT

Our framework ultimately involves binary classifiers for both flow and frame embeddings, with three distinct backbone variants (Xception, UCF, SPSL) trained to evaluate each embedding type. Finally, the LID block weighs all pair-level embeddings to produce a *single binary decision* (real or fake) for the video. By factoring in multiple domains (spatial and temporal), EVDO surpasses single-frame classifiers [15] and is on-par with attention-based aggregation [17], giving a robust decision even if artifacts vary across frames or appear only briefly.

IV. EXPERIMENTAL SETUP AND VALIDATION

A. EXPERIMENT SETUP

We conducted tests across four datasets: Deepfakes (FF-DF) [65], Face2Face (FF-F2F) [22], FaceSwap (FF-FS) [23] and NeuralTextures (FF-NT) [24] from FaceForensics++ [16] benchmark. We provide a basis for assessing EVDO's performance across a variety of face manipulation technologies.

While our experiments focus on FaceForensics++ subsets for fair comparison with prior works, we acknowledge the need for cross-dataset evaluation (e.g., training on FF++ and testing on Celeb-DF [25] or DFDC [26]). Future work will extend EVDO's validation to these benchmarks to rigorously assess generalization.

We assessed the performance of our Enhanced Video Deepfake Observer (EVDO) framework using Area Under Curve (AUC) for both individual frames and complete videos, which measures our system's ability to distinguish between authentic and deepfake images. The AUC is defined as $AUC = \int_0^1 F(x)\,dx = P(s > s' \mid y = 1, y' = 1)$, where F(x) represents the true positive rate, and s and s' are the prediction scores for positive and negative samples, respectively. We used conditional accuracy to evaluate how these two modules contribute to the overall accuracy, denoted as

$$Acc_c = \frac{\sum_{i=1}^{N} \mathbf{1}(P(\hat{\phi}[i]) \neq y[i] \wedge P(y_A[i]) = y[i])}{\sum_{i=1}^{N} 1(P(\hat{\phi}[i]) \neq y[i])}$$

where $P(\hat{\phi}[i])$ represents the binary classification output from a learnable detector. This detector utilizes optical flow prediction to generate its classifications. Similarly, $P(y_A[i])$ denotes the binary classification output from spatial detectors. Conditional accuracy shows the contribution of each module to the overall accuracy.

We evaluated four detectors — Xception [13], SPSL [73], UCF [32], and our comprehensive EVDO pipeline with the three variants. For each variant, we train using a variable number of pairs per video (for example, 4 individual frame-pairs, giving 8 individual frames and 4 flow images per video). During evaluation LID produces a binary outcome based on one pair per video. Consequently, the pair metrics report performance per number of pairs used during training, and the video metrics report the final evaluation outcome derived from aggregating metrics for all pairs from the respective video.

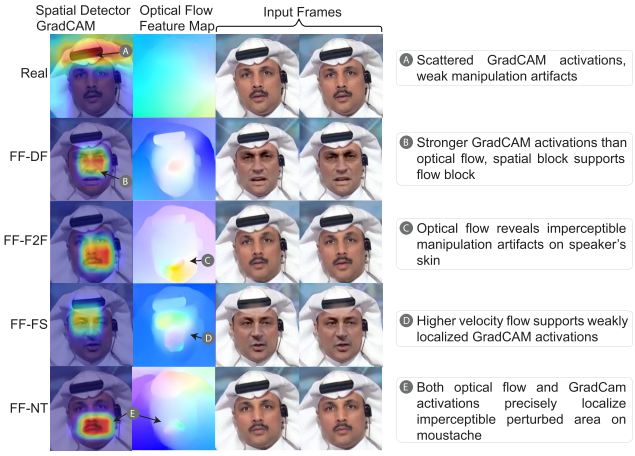

FIGURE 3. Visualization of the EVDO feature maps across four datasets and a real frame. GradCAM visualizations support the optical flow visualizations for the deepfake cases.

B. VISUAL COMPARISON ON FEATURE MAP

The visualization Figure 3 shows the feature embeddings generated by the spatial feature extractor and the optical flow feature extractor from EVDO, using two video examples from [16] across all four datasets, each at the same frame. On the left side of the feature map, the embeddings from the optical flow show a distinct face shape with notable patterns for deepfake faces, contrasting with the less defined face shape and smoother optical flow of natural faces. On the right-hand side, the heatmap from the spatial detector reveals that deepfake faces exhibit a more extensive area of gradient changes within the face area compared to real faces, suggesting a greater extent of perturbation by deepfake manipulation techniques and a higher likelihood of tampering in the original frame.

V. RESULTS AND DISCUSSION

We compare multiple spatial feature detectors with the Xception backbone to assess their performance across the four datasets we used as our baseline to compare with our approach EVDO.

Table 2 shows the conditional accuracies, which is the percentage of cases where the flow-level detector correctly classifies a video which the image-level (video frame) detector misclassified, and vice versa.

Our experiments show non-zero minimal conditional accuracies across datasets generated using different deepfake generation methods. This tells us that there is indeed additional information in the temporal embedding which spatial detectors do not have access to. Furthermore, this initial result is based on the intermediate binary classifier,

TABLE 2. Main result and comparison of different methods (Video AUC). I: Image only (Frame), F: flow only. conditional accuracies produced by our submodules: Ix, Fv: percentage of flow image classified correctly given that image (frame) misclassified, Iv, Fx: Percentage of image (frame) classified correctly given that flow image is misclassified.

Approach	Type	FF-c23	FF-DF	FF-F2F	FF-FS	FF-NT
Native SPSL	Image	0.9746	0.9896	0.9884	0.9976	0.9227
Native UCF	Image	0.9932	0.9984	0.9978	0.9977	0.9787
Native Xcep	Image	0.9894	0.9974	0.9956	0.9939	0.9707
Cond. Acc.	Ix, Fv	12.27%	5.88%	8.57%	11.11%	23.53%
	Iv, Fx	77.63%	77.97%	77.22%	77.30%	78.01%
EVDO (Ours)	I + F	0.9913	0.9982	0.9962	0.9981	0.9726

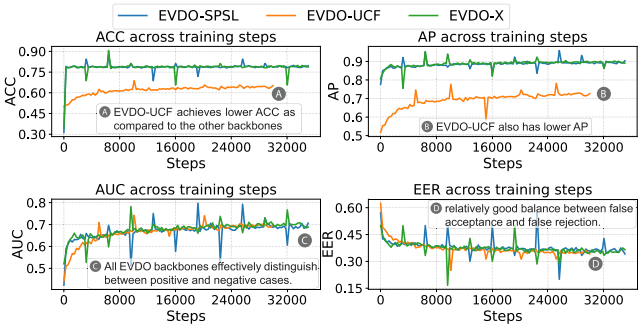

FIGURE 4. Diagram for detector training. (a). Accuracy (ACC) (b). Average precision (AP) (c). Area under curve (AUC) (d). equal error rate (ERR).

which does not aggregate information from all embeddings. This means that the reported conditional accuracies are likely to be the maximal bound for the same metric which are presented in the ablation study, which uses LID with access to both types of embeddings, for more fine-grained correlation of hidden high-dimensional temporal and spatial features.

On the other hand, these conditional accuracies could be improved by spending more effort into the choice of the flow detector's structure and cost functions. The reason is that the currently used detectors are designed to treat RGB images of people's faces, rather than a visual representation of optical flow, of the temporal inconsistencies between consecutive frames. Figure 4 shows the training performance.

A. ABLATION STUDY

In this section, we present a comprehensive ablation study to investigate the effects of two key design choices within our proposed framework. First, we examine the impact of the chosen backbone architecture by comparing three variants: a baseline configuration with the native Xception backbone (EVDO-X), a modified architecture employing SPSL with an Xception backbone (EVDO-SPSL), and a modified architecture employing UCF with an Xception backbone (EVDO-UCF). These variations enable us to assess how different learned feature representations influence the model's capability to detect manipulations across multiple datasets.

Second, we explore how the temporal context, as captured by the number of input frame pairs processed end-to-end, affects performance. To this end, we consider EVDO variations receiving 1, 2, 3, or 4 pairs of frames per video. By systematically increasing the temporal resolution in the input, we aim to quantify the benefits of incorporating richer motion cues and temporal patterns in deepfake detection.

In the following subsections, we detail the experimental setup for each of these ablations and analyze their outcomes, showing how backbone selection and the depth of temporal context jointly shape the effectiveness of our model.

B. EFFECT OF BACKBONE ARCHITECTURE

Figure 5 presents results for EVDO-X, EVDO-SPSL and EVDO-UCF, respectively. Under comparable conditions, EVDO-X and EVDO-UCF backbones consistently achieve higher accuracies, AUC, and AP values than SPSL. Both EVDO-X and EVDO-UCF show higher values across their best epochs, demonstrating superior overall correctness. AUC, which evaluates the model's ability to distinguish between genuine and manipulated inputs across thresholds, is also higher for EVDO-X and EVDO-UCF, demonstrating better discrimination capabilities. AP, derived from the precision-recall curve, is highest for EVDO-UCF, demonstrating its improved ability to prioritize manipulated inputs. EER, which captures the balance between false acceptance and false rejection rates, is lowest for EVDO-UCF, demonstrating fewer errors at the decision threshold. While EVDO-SPSL achieves competitive intermediate performance, its metrics remain below those of EVDO-X and EVDO-UCF, with EVDO-UCF achieving the highest peak values across most metrics, demonstrating more effective feature extraction and classification by the naïve and spatial backbones.

Our approach is an outlier on FF-NT with EVDO-SPSL because (1) FF-NT videos are generated via 3D face mesh reconstruction, Neural Textures, and a Deferred Neural Renderer [24], for temporal coherence without generative hallucination. This results in frequency characteristics distinct from GAN-based methods. (2) SPSL detects upsampling artifacts common in GANs, but these generative frequencies are absent by design in FF-NT, limiting SPSL's effectiveness.

C. EFFECT OF TEMPORAL CONTEXT (NUMBER OF FRAME PAIRS)

Figure 5 also presents the performance of EVDO-X, EVDO-SPSL and EVDO-UCF as the number of frame pairs processed is varied across the FF-DF, FF-F2F, FF-FS, and FF-NT subsets.

Across all subsets, incrementing the width of the input from 1 to 4 (i.e. number of training frame pairs) achieves consistent improvements in both pair-level and video-level metrics across the test set, with varying degrees of impact. Following the main tendency, pair accuracy, pair AUC, and pair AP improve as the number of pairs increases, pair EER decreases. This directly confirms our hypothesis:

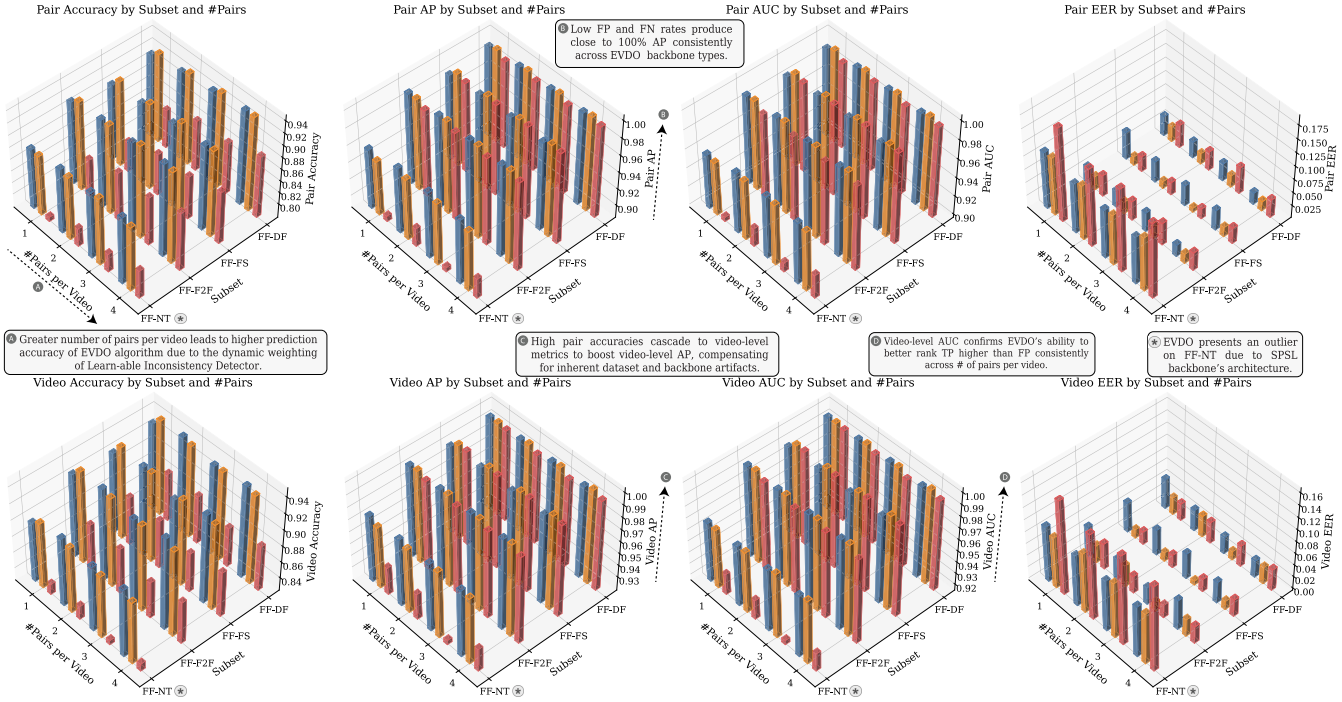

FIGURE 5. **Performance across different backbones on Xception, UCF, and SPSL: This figure evaluates our approach using different backbones across varying frame pairs per video. Results indicate that increasing frame pairs generally improves performance, with variations across datasets and architectures.**

providing more temporal context during training enhances the discriminatory weighing mechanism of LID. Video-level metrics, including Video Accuracy and Video AUC, also show improvements, though gains tend to plateau after 3 pairs, particularly in FF-DF and FF-F2F. The FF-NT subset, being the most challenging, shows the smallest improvements across metrics, reflecting its inherent difficulty. While the extent of improvement differs among subsets, the overall tendency demonstrates the benefit of leveraging temporal context for both pair-level and video-level performance, without the need to increase input width at inference time.

Finally, depending on the dataset, temporal information introduces inherent artifacts, including imperfections from the cropping process, compression-induced distortions, natural subject movements, and backbone design and architecture. As a result, increasing the number of frame pairs per video beyond a certain threshold does not yield proportional benefits. In our experiments, this point of diminishing returns is identified at three pairs for EVDO-X, three pairs for EVDO-SPSL, and two pairs for EVDO-UCF.

D. JOINT ANALYSIS OF BACKBONE AND FRAME PAIR ABLATIONS

Combining the insights from Sections V-B and V-C, we observe that while the choice of backbone architecture determines the upper performance bounds, increasing the number of frame pairs provides a complementary enhancement. EVDO-X and EVDO-UCF exhibit higher baseline metrics and more pronounced gains with up to three pairs, whereas EVDO-SPSL's overall performance remains comparatively lower but still benefits from additional temporal context. Beyond the identified points of diminishing returns—three pairs for EVDO-X and EVDO-SPSL, and two pairs for EVDO-UCF—further increases in temporal resolution do not yield proportional improvements and may lead to imbalances in the underlying training sets. These findings complement the architectural feature extraction capabilities of each backbone, with the effective integration of temporal information, guiding the selection of both backbone and temporal configuration for optimal deepfake detection performance across datasets. See Appendix for reference.

PART (1.A): DETAILED METRICS

Figure 5 shows the performance of our approach with different backbones on **Xception**, **UCF** and **SPSL**. We calculated the accuracy, AP, AUC and EER for different numbers of frame pairs sampled in one video stream instance.

The results provide metrics across four subsets: **FF-DF**, **FF-F2F**, **FF-FS**, and **FF-NT**. Key observations are as follows:

- For **FF-DF**, the highest pair accuracy is **93.84%**, achieved with 3 pairs per video. However, the configuration with the highest Pair AUC (**99.86%**) and Video Accuracy (**93.57%**) is 4 pairs per video. All are produced by EVDO-UCF.
- For **FF-F2F**, 4 pairs per video result in the highest pair accuracy (**93.63%**), Pair AUC (**99.43%**), Video Accuracy (**93.93%**). All are produced by EVDO-X.

- The **FF-FS** subset achieves the best pair accuracy (**93.03%**) at 3 pairs per video, accompanied by a high Pair AUC (**99.75%**), and Video Accuracy (**93.93%**). All are produced by EVDO-UCF.
- In the **FF-NT** subset, which is more challenging, the highest pair accuracy is **89.34%** with 4 pairs per video, though Video EER remains relatively high at **8.57%** across all configurations.

PART (1.B): BEST METRICS

Table 2 shows the experiment results obtained by our approach on three backbones. The second table highlights the best metrics for the EVDO framework:

- Best pair accuracy was achieved with 3 pairs per video (**95.45%**), which also corresponds to the lowest Pair EER (**4.19%**). Produced by EVDO-UCF.
- The best Pair AUC was achieved with 4 pairs per video **98.97%**, with the highest Pair AP of **99.72%**. Produced by EVDO-UCF.

PART (1.C): OVERALL RESULTS

Table 2 presents the overall accuracy results across five datasets: **FF-c23**, **FF-DF**, **FF-F2F**, **FF-FS**, and **FF-NT**. Our proposed approach, EVDO, demonstrates strong performance, outperforming the state-of-the-art baselines on the **FF-FS** dataset. This highlights the effectiveness of incorporating both frame-level and optical flow features for more generalizable deepfake detection. Additionally, EVDO maintains highly competitive performance across the other datasets, achieving 99.13% video AUC on the overall set termed **FF-c23**: 99.82% on **FF-DF**, 99.62% on **FF-F2F**, and 97.26% on **FF-NT**. These results indicate that our model generalizes well across these different manipulation techniques while maintaining on-par accuracy, particularly in detecting frame-based and temporal inconsistencies. Compared to the baseline methods that rely solely on frame-based analysis, the inclusion of motion-based features in EVDO contributes to enhanced more generalizable detection capabilities, especially in cases where subtle artifacts are more difficult to capture. Overall, our approach provides a robust and effective solution for deepfake detection, contributing a benchmark to the field.

VI. LIMITATION AND FUTURE WORK

EVDO has demonstrated promising results by integrating spatial-temporal analysis on deepfake video detection. However, our current work encounters the following limitations and potential future directions: **1.** The inclusion of an additional optical flow predictor module increases the computational complexity of the framework. **2.** Further exploration with additional spatial detectors is necessary to evaluate the impact of using a diverse range of detectors on our feature aggregation in EVDO. Future directions include efforts to enhance the efficiency of the optical flow prediction module and to conduct analyses using a broader array of the selection on spatial detectors. **3.** Cross-Dataset Generalization: While EVDO achieves strong intra-dataset performance in 4 distinct deepfake manipulation techniques, cross-dataset evaluation is essential for real-world deployment. We plan to incorporate domain adaptation techniques in future work to address this gap.

VII. CONCLUSION

We proposed EVDO, an Enhanced Framework for Deepfake Detection with main object to enhance the generalizability of deepfake detection especially on face forensics. The fusion of temporal and spatial embeddings within EVDO not only improve detection accuracy but also enhance the system's generalization across the four types of deepfake generation techniques, as temporal features contribute exclusively to the detection's input feature space. The use of optical flow techniques coupled with the advanced Sintel FlowFormer++ model enables the detection of time inconsistencies between video frames, which are crucial for identifying deepfake content. The spatial detectors such as Xception increase the system's ability to analyze each frame for anomalies, for a comprehensive multi-modal approach to deepfake detection. Our approach obtained a non-zero conditional accuracy towards the exclusivity of flow features among 4 datasets generated from different sources and approaches, which achieves state-of-the-art results among existing cutting edge standalone deepfake detection methods. The contributions of this paper not only pave the way for more secure digital media landscapes but also set a scalable standard for research in the field of artificial intelligence and cybersecurity.

APPENDIX
EXPANDED ABLATION RESULTS

This appendix reports complete ablations for EVDO combined with FlowFormer++ and three spatial backbones (Xception, SPSL, UCF). For each number of temporal pairs per video $k \in \{1, 2, 3, 4\}$, Tables 3, 7, and 5 list pair-level and video-level metrics across FaceForensics++ subsets. Tables 4, 8, and 6 summarize the best-metric configurations per k. EER is lower-is-better; the other metrics are higher-is-better. ''Best Epoch'' denotes the epoch at which the highest overall accuracy was observed.

A. HIGH-LEVEL TRENDS

Across all FaceForensics++ subsets and backbones, increasing the number of temporal pairs per video k generally improves accuracy and AUC while reducing EER, with clear gains from $k = 1$ to $k = 2$, diminishing returns at $k = 3$, and occasional plateaus or slight regressions at $k = 4$. Aggregating decisions at the video level consistently yields lower EER and slightly higher AUC than pair-level evaluation (e.g., Tables 3 and 7), confirming that temporal pooling helps stabilize predictions.

B. BACKBONE COMPARISON (OVERALL)

The UCF backbone is the strongest overall in the same-dataset settings, followed by Xception, then SPSL.

TABLE 3. Ablation study (1.a): EVDO with FlowFormer++ + xception backbone results.

Subset	k	Pair Acc.	Pair AUC	Pair EER	Pair AP	Video Acc.	Video AUC	Video EER	Video AP
FF-DF	1	0.9174	0.9917	0.0427	0.9931	0.9321	0.9958	0.0500	0.9960
FF-DF	2	0.9258	0.9950	0.0366	0.9955	0.9429	0.9958	0.0357	0.9961
FF-DF	3	0.9314	0.9956	0.0391	0.9958	0.9357	0.9963	0.0286	0.9965
FF-DF	4	0.9381	0.9960	0.0279	0.9963	0.9393	0.9955	0.0286	0.9959
FF-F2F	1	0.9183	0.9920	0.0378	0.9938	0.9286	0.9951	0.0500	0.9954
FF-F2F	2	0.9277	0.9942	0.0366	0.9954	0.9393	0.9949	0.0357	0.9955
FF-F2F	3	0.9302	0.9940	0.0419	0.9953	0.9321	0.9950	0.0286	0.9956
FF-F2F	4	0.9363	0.9943	0.0279	0.9956	0.9393	0.9951	0.0500	0.9957
FF-FS	1	0.9067	0.9872	0.0575	0.9876	0.9250	0.9908	0.0500	0.9918
FF-FS	2	0.9139	0.9912	0.0453	0.9913	0.9357	0.9921	0.0429	0.9932
FF-FS	3	0.9195	0.9928	0.0447	0.9925	0.9250	0.9917	0.0429	0.9932
FF-FS	4	0.9229	0.9919	0.0438	0.9915	0.9357	0.9924	0.0357	0.9937
FF-NT	1	0.8810	0.9568	0.1125	0.9582	0.9000	0.9680	0.0929	0.9751
FF-NT	2	0.8888	0.9635	0.0993	0.9649	0.9107	0.9691	0.0857	0.9758
FF-NT	3	0.8916	0.9649	0.0978	0.9663	0.9071	0.9704	0.0857	0.9767
FF-NT	4	0.8934	0.9644	0.0956	0.9643	0.9107	0.9703	0.0857	0.9762

TABLE 4. Ablation study (1.b): EVDO with FlowFormer++ + Xception backbone best metrics.

k	Best Epoch	Best Overall Acc.	Pair Acc.	Pair AUC	Pair EER	Pair AP
1	5	0.9435	0.9435	0.9829	0.0640	0.9955
2	3	0.9489	0.9489	0.9869	0.0557	0.9966
3	1	0.9533	0.9533	0.9877	0.0587	0.9968
4	3	0.9523	0.9523	0.9878	0.0558	0.9967

TABLE 5. Ablation study (2.a): EVDO with FlowFormer++ + SPSL backbone results.

Subset	k	Pair Acc.	Pair AUC	Pair EER	Pair AP	Video Acc.	Video AUC	Video EER	Video AP
FF-DF	1	0.8446	0.9889	0.0493	0.9898	0.8857	0.9951	0.0286	0.9952
FF-DF	2	0.8647	0.9921	0.0418	0.9926	0.8857	0.9937	0.0357	0.9938
FF-DF	3	0.8669	0.9915	0.0559	0.9916	0.8750	0.9923	0.0357	0.9924
FF-DF	4	0.8842	0.9952	0.0398	0.9951	0.8821	0.9947	0.0357	0.9947
FF-F2F	1	0.8428	0.9861	0.0452	0.9892	0.8821	0.9921	0.0429	0.9934
FF-F2F	2	0.8615	0.9883	0.0418	0.9911	0.8821	0.9901	0.0286	0.9923
FF-F2F	3	0.8617	0.9876	0.0475	0.9906	0.8714	0.9910	0.0214	0.9922
FF-F2F	4	0.8785	0.9891	0.0398	0.9920	0.8786	0.9897	0.0286	0.9921
FF-FS	1	0.8309	0.9927	0.0345	0.9924	0.8857	0.9971	0.0214	0.9973
FF-FS	2	0.8501	0.9944	0.0296	0.9940	0.8857	0.9959	0.0214	0.9963
FF-FS	3	0.8529	0.9969	0.0251	0.9966	0.8750	0.9972	0.0214	0.9974
FF-FS	4	0.8685	0.9953	0.0359	0.9944	0.8821	0.9959	0.0286	0.9962
FF-NT	1	0.7876	0.9034	0.1806	0.8942	0.8393	0.9343	0.1571	0.9429
FF-NT	2	0.8104	0.9180	0.1516	0.9092	0.8429	0.9332	0.1429	0.9415
FF-NT	3	0.8111	0.9084	0.1620	0.8929	0.8321	0.9203	0.1500	0.9240
FF-NT	4	0.8277	0.9239	0.1474	0.9112	0.8357	0.9321	0.1429	0.9391

TABLE 6. Ablation study (2.b): EVDO with FlowFormer++ + SPSL backbone best metrics.

k	Best Epoch	Best Overall Acc.	Pair Acc.	Pair AUC	Pair EER	Pair AP
1	9	0.9131	0.9131	0.9697	0.0878	0.9918
2	8	0.9220	0.9220	0.9750	0.0801	0.9933
3	4	0.9235	0.9235	0.9731	0.0838	0.9927
4	8	0.9302	0.9302	0.9781	0.0717	0.9940

TABLE 7. Ablation study (3.a): EVDO with FlowFormer++ + UCF backbone results.

Subset	k	Pair Acc.	Pair AUC	Pair EER	Pair AP	Video Acc.	Video AUC	Video EER	Video AP
FF-DF	1	0.9280	0.9957	0.0337	0.9961	0.9429	0.9962	0.0286	0.9963
FF-DF	2	0.9328	0.9973	0.0261	0.9975	0.9393	0.9971	0.0357	0.9973
FF-DF	3	0.9384	0.9977	0.0279	0.9978	0.9357	0.9972	0.0143	0.9974
FF-DF	4	0.9381	0.9986	0.0239	0.9986	0.9357	0.9982	0.0286	0.9982
FF-F2F	1	0.9286	0.9934	0.0312	0.9947	0.9393	0.9962	0.0286	0.9966
FF-F2F	2	0.9286	0.9958	0.0279	0.9964	0.9357	0.9957	0.0357	0.9963
FF-F2F	3	0.9358	0.9957	0.0279	0.9963	0.9321	0.9957	0.0143	0.9963
FF-F2F	4	0.9343	0.9963	0.0239	0.9968	0.9321	0.9962	0.0286	0.9967
FF-FS	1	0.9220	0.9958	0.0222	0.9959	0.9393	0.9973	0.0143	0.9977
FF-FS	2	0.9226	0.9966	0.0192	0.9968	0.9357	0.9975	0.0143	0.9979
FF-FS	3	0.9303	0.9975	0.0140	0.9975	0.9321	0.9971	0.0071	0.9977
FF-FS	4	0.9252	0.9964	0.0239	0.9963	0.9321	0.9980	0.0143	0.9983
FF-NT	1	0.8805	0.9551	0.1141	0.9519	0.9071	0.9674	0.0857	0.9711
FF-NT	2	0.8859	0.9615	0.1080	0.9595	0.9071	0.9693	0.1000	0.9733
FF-NT	3	0.8932	0.9633	0.1006	0.9609	0.9036	0.9726	0.0929	0.9756
FF-NT	4	0.8889	0.9629	0.1076	0.9589	0.9036	0.9719	0.0857	0.9754

TABLE 8. Ablation study (3.b): EVDO with FlowFormer++ + UCF backbone best metrics.

k	Best Epoch	Best Overall Acc.	Pair Acc.	Pair AUC	Pair EER	Pair AP
1	4	0.9478	0.9478	0.9859	0.0599	0.9963
2	5	0.9511	0.9511	0.9888	0.0505	0.9970
3	3	0.9545	0.9545	0.9894	0.0419	0.9972
4	0	0.9532	0.9532	0.9897	0.0438	0.9972

On *FF-DF* and *FF-FS* in particular, UCF attains the highest AUCs and the lowest EERs at both the pair and video levels (Table 7). Xception is competitive and occasionally superior on *FF-NT* at $k \in \{2, 4\}$ (Table 3), while SPSL lags on all subsets, especially *FF-NT*.

C. WHERE MOTION HELPS THE MOST

The largest relative gains from adding temporal pairs are observed on *FF-DF* and *FF-FS*, where motion inconsistencies interact with photometric and warping artefacts. With UCF at $k = 3$, *FF-DF* reaches a video AUC of 0.9972 with a video EER of 0.0143, and *FF-FS* reaches a video AUC of 0.9971 with a video EER of 0.0071 (Table 7), indicating that EVDO's flow tokens capture discriminative temporal cues even with strong spatial content.

D. THE HARDEST SUBSET

FF-NT is consistently the most challenging. Pair accuracies and AUCs are lower and EERs are higher for all backbones (Tables 3–7). Interestingly, Xception benefits noticeably from more pairs on *FF-NT* (video accuracy 0.9107 at $k = 2$ and $k = 4$ in Table 3), slightly edging UCF on that subset. This suggests that when manipulations are more texture-dominant and motion artifacts are subtle, a texture-oriented backbone fused with flow can be preferable.

ACKNOWLEDGMENT

(Mohammed Mouad Melouk and Minghao Shao contributed equally to this work.)

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Sci. Robot.*, vol. 63, no. 11, pp. 139–144, 2020.

[2] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4217–4228, Dec. 2021.

[3] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[4] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. 4th Int. Conf. Learn. Represent.*, 2015.

[5] A. Razavi, A. van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates Inc., 2019, Paper 1331.

[6] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. 38th Int. Conf. Mach. Learn.* (Proceedings of Machine Learning Research), vol. 139, M. Meila and T. Zhang, Eds., PMLR, Jul. 2021, pp. 8821–8831. [Online]. Available: http://proceedings.mlr.press/v139/ramesh21a/ramesh21a.pdf

[7] D. Ramirez and C. Andrada, "70 deepfake statistics you need to know (2024)," Spiralytics, Manila, Philippines, 2024. [Online]. Available: https://www.spiralytics.com/blog/deepfake-statistics/

[8] S. Horswell, "Identity fraud report 2024," Entrust, Shakopee, MN, USA, 2024. [Online]. Available: https://onfido.com/landing/identity-fraud-report/

[9] L. Takruri, "How deepfakes are challenging identity verification," Entrust, Shakopee, MN, USA, 2023. [Online]. Available: https://onfido.com/blog/deepfakes-challenging-identity-verification/

[10] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Comput. Vis. Image Understand.*, vol. 223, Oct. 2022, Art. no. 103525.

[11] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020.
[12] L. Verdoliva, "Media forensics and DeepFakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, Aug. 2020.
[13] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.
[14] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, Dec. 2018, pp. 1–7.
[15] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2019, pp. 80–87. [Online]. Available: https://api.semanticscholar.org/CorpusID:147704268
[16] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.
[17] Z. Wang, J. Bao, W. Zhou, W. Wang, and H. Li, "AltFreezing for more general video face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 4129–4138.
[18] Q. Yin, W. Lu, B. Li, and J. Huang, "Dynamic difference learning with spatio-temporal correlation for deepfake video detection," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 4046–4058, 2023.
[19] J. Choi, T. Kim, Y. Jeong, S. Baek, and J. Choi, "Exploiting style latent flows for generalizing deepfake video detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 1133–1143.
[20] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15024–15034.
[21] X. Shi, Z. Huang, D. Li, M. Zhang, K. C. Cheung, S. See, H. Qin, J. Dai, and H. Li, "FlowFormer++: Masked cost volume autoencoding for pretraining optical flow estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1599–1610.
[22] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2387–2395.
[23] M. Kowalski, "Faceswap," Microsoft, Cambridge, U.K., 2019. [Online]. Available: https://github.com/MarekKowalski/FaceSwap
[24] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, Aug. 2019.
[25] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3207–3216.
[26] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.
[27] T. Jung, S. Kim, and K. Kim, "DeepVision: Deepfakes detection using human eye blinking pattern," *IEEE Access*, vol. 8, pp. 83144–83154, 2020.
[28] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, "Deepfake detection using spatiotemporal convolutional networks," 2020, *arXiv:2006.14749*.
[29] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2307–2311.
[30] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5000–5009.
[31] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the detection of digital face manipulation," 2019, *arXiv:1910.01717*.
[32] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "UCF: Uncovering common features for generalizable deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 22412–22423.
[33] Y.-Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 86–103.
[34] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
[35] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 3483–3491.
[36] A. Van Den Oord, "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[37] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "CVAE-GAN: Fine-grained image generation through asymmetric training," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2764–2773.
[38] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.
[39] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.
[40] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7183–7192.
[41] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3697–3705.
[42] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.
[43] L. Lei, K. Gai, J. Yu, and L. Zhu, "DiffuseTrace: A transparent and flexible watermarking scheme for latent diffusion model," 2024, *arXiv:2405.02696*.
[44] Z. Peng, W. Hu, Y. Shi, X. Zhu, X. Zhang, H. Zhao, J. He, H. Liu, and Z. Fan, "SyncTalk: The devil is in the synchronization for talking head synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 666–676.
[45] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: Automatically replacing faces in photographs," in *Proc. ACM SIGGRAPH Papers*, Aug. 2008, pp. 1–8.
[46] Y. Nirkin, I. Masi, A. Tran Tuan, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 98–105.
[47] W. Wu, Y. Zhang, C. Li, C. Qian, and C. C. Loy, "ReenactGAN: Learning to reenact faces via boundary transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 622–638.
[48] A. Pumarola, A. Agudo, A. M. Martínez, A. Sanfeliu, and F. Moreno-Noguer, "GANimation: Anatomically-aware facial animation from a single image," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 818–833.
[49] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5464–5478, Nov. 2019.
[50] Z. Lin, W. Xu, X. Ma, C. Xu, and H. Xiao, "Multi-attention infused integrated facial attribute editing model: Enhancing the robustness of facial attribute manipulation," *Electronics*, vol. 12, no. 19, p. 4111, Sep. 2023.
[51] R. Chesney and D. Citron, "Deepfakes and the new disinformation war: The coming age of post-truth geopolitics," *Foreign Aff.*, vol. 98, p. 147, Jan. 2019.
[52] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley, "Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3347–3356.
[53] A. George, M. Punia, S. Agarwal, and R. Singh, "Deepfakes, deception and the law: Unraveling the legal complexities of AI-generated content," in *Proc. Int. Conf. Smart Syst., Innov. Comput.*, 2024, pp. 747–763.
[54] Y. Zhang, L. Zheng, and V. L. L. Thing, "Automated face swapping and its detection," in *Proc. IEEE 2nd Int. Conf. Signal Image Process. (ICSIP)*, Aug. 2017, pp. 15–19.
[55] D. Güera, S. Baireddy, P. Bestagini, S. Tubaro, and E. J. Delp, "We need no pixels: Video manipulation detection using stream descriptors," 2019, *arXiv:1906.08743*.
[56] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking," 2018, *arXiv:1806.02877*.
[57] S. Fernandes, S. Raj, E. Ortiz, I. Vintila, M. Salter, G. Urosevic, and S. Jha, "Predicting heart rate variations of deepfake videos using neural ODE," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1721–1729.
[58] J. Fridrich, M. Goljan, and D. Soukal, "Higher-order statistical steganalysis of palette images," *Proc. SPIE*, vol. 5020, pp. 178–190, Jun. 2003.
[59] J. J. Gibson, *The Perception of the Visual World*. Boston, MA, USA: Houghton Mifflin, 1950.
[60] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.
[61] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Image Anal., 13th Scandin. Conf.*, Jun. 2003, pp. 363–370.

[62] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.
[63] T.-W. Hui, X. Tang, and C. C. Loy, "LiteFlowNet: A lightweight convolutional neural network for optical flow estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8981–8989.
[64] Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li, "FlowFormer: A transformer architecture for optical flow," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 668–685.
[65] *Faceswap: Deepfakes Software for All*, Deepfakes, 2019. [Online]. Available: https://github.com/deepfakes/faceswap
[66] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8261–8265.
[67] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2886–2895.
[68] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu, "ForgeryNet: A versatile benchmark for comprehensive forgery analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4358–4367.
[69] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "DeepfakeBench: A comprehensive benchmark of deepfake detection," in *Proc. Adv. Neural Inf. Process. Syst.*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023, pp. 4534–4565.
[70] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1205–1207.
[71] R. Caldelli, T. Uricchio, and I. Amerini, "Optical flow based CNN for detection of unlearnt deepfake manipulations," *Pattern Recognit. Lett.*, vol. 146, pp. 31–37, Jun. 2021.
[72] A. Chintha, S. Agarwal, and H. Farid, "Leveraging edges and optical flow on faces for deepfake detection," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–9.
[73] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 772–781.

MOHAMMED MOUAD MELOUK received the B.Sc. degree in computer engineering from New York University Abu Dhabi, in May 2025. He contributes to and leads machine learning–focused research projects in ML security, agentic AI, and low-level machine learning architecture design.

MINGHAO SHAO received the B.S. degree in computer science with artificial intelligence from the University of Nottingham (UoN), U.K., in 2020, and the M.S. degree in computer science from New York University (NYU), USA, in 2023. He is currently pursuing the Ph.D. degree in computer science with New York University Abu Dhabi (NYUAD), co-advised by Prof. Muhammad Shafique and Prof. Ramesh Karri with New York University Tandon School of Engineering. His research interests include agentic systems, large language models, machine learning, and cybersecurity.

ABDUL BASIT received the B.Sc. degree in electrical engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2021. He is currently a Research Engineer with the eBRAIN Laboratory, New York University Abu Dhabi (NYUAD), United Arab Emirates. He has been awarded the Rector's Gold Medal in B.S. in EE. His research interests include machine learning, large language models, robotics systems, human–computer interaction, energy-efficient design, and emerging technologies.

CHAIMAE ABOUZAHIR received the B.S. degree in computer engineering from New York University Abu Dhabi, in 2025. Her work focuses on natural language generation and adapting language models for high-stakes domains, such as healthcare and cybersecurity, with emphasis on model tuning, optimization, domain-adaptive pretraining, and explainability methods.

RUILIN ZHOU received the B.S. degree in computer engineering from New York University Abu Dhabi, in 2025. Her research interests include applied machine learning, with a focus on language models, cybersecurity, and efficient model design.

MUHAMMAD SHAFIQUE (Senior Member, IEEE) received the Ph.D. degree in computer science from Karlsruhe Institute of Technology (KIT), Germany, in 2011. Afterwards, he established and led a highly recognized research group with KIT for several years and conducted impactful collaborative research and development activities across the globe. In October 2016, he joined the Institute of Computer Engineering, Faculty of Informatics, Technische Universität Wien (TU Wien), Vienna, Austria, as a Full Professor of computer architecture and robust, energy-efficient technologies. Since September 2020, he has been with New York University (NYU) Abu Dhabi, United Arab Emirates, where he is currently a Full Professor and the Director of the eBrain Laboratory, and a Global Network Professor with the Tandon School of Engineering, NYU-New York City, USA. He is also a CoPI/an Investigator in multiple NYUAD Centers, including the Center of Artificial Intelligence and Robotics (CAIR), the Center of Cyber Security (CCS), the Center for Interacting Urban Networks (CITIES), and the Center for Quantum and Topological Systems (CQTS). He holds one U.S. patent, and has co-authored six books, more than ten book chapters, more than 350 papers in premier journals and conferences, and more than 100 archive articles. His research interests include AI and machine learning hardware and system-level design, brain-inspired computing, machine learning security and privacy, quantum machine learning, cognitive autonomous systems, wearable healthcare, energy-efficient systems, robust computing, hardware security, emerging technologies, FPGAs, MPSoCs, and embedded systems. He has given several keynotes, invited talks, and tutorials, and organized many special sessions at premier venues. He is a Senior Member of the IEEE Signal Processing Society (SPS) and a member of the ACM, SIGARCH, SIGDA, SIGBED, and HIPEAC. He has served as the PC chair, the general chair, the track chair, and a PC member for several prestigious IEEE/ACM conferences. He received the 2015 ACM/SIGDA Outstanding New Faculty Award, the AI 2000 Chip Technology Most Influential Scholar Award, in 2020 and 2022, the AI 2000 Chip Technology Most Influential Scholar Award, in 2020, 2022, and 2023, the ASPIRE AARE Research Excellence Award, in 2021, six gold medals, and several best paper awards and nominations at prestigious conferences.