DeepShield: Fortifying Deepfake Video Detection with Local and Global Forgery Analysis

Yinqi Cai[1,2*] Jichang Li[2*] Zhaolun Li[3] Weikai Chen[†] Rushi Lan[3] Xi Xie[1]
Xiaonan Luo[3] Guanbin Li[1,2,4‡]
[1]Sun Yat-sen University [2]Pengcheng Laboratory [3]Guilin University of Electronic Technology
[4]Guangdong Key Laboratory of Big Data Analysis and Processing
caiyq27@mail2.sysu.edu.cn, li.jichang@pcl.ac.cn, liguanbin@mail.sysu.edu.cn

Abstract

Recent advances in deep generative models have made it easier to manipulate face videos, raising significant concerns about their potential misuse for fraud and misinformation. Existing detectors often perform well in in-domain scenarios but fail to generalize across diverse manipulation techniques due to their reliance on forgery-specific artifacts. In this work, we introduce DeepShield, a novel deepfake detection framework that balances local sensitivity and global generalization to improve robustness across unseen forgeries. DeepShield enhances the CLIP-ViT encoder through two key components: Local Patch Guidance (LPG) and Global Forgery Diversification (GFD). LPG applies spatiotemporal artifact modeling and patch-wise supervision to capture fine-grained inconsistencies often overlooked by global models. GFD introduces domain feature augmentation, leveraging domain-bridging and boundary-expanding feature generation to synthesize diverse forgeries, mitigating overfitting and enhancing cross-domain adaptability. Through the integration of novel local and global analysis for deepfake detection, DeepShield outperforms state-of-the-art methods in cross-dataset and cross-manipulation evaluations, achieving superior robustness against unseen deepfake attacks.

1. Introduction

Rapid advancement of deep generative models, including GANs [15, 23], VAEs [26, 27], and diffusion models [9, 51], has significantly lowered the barrier for manipulating facial videos, raising serious concerns about their misuse in fraud, misinformation, and privacy violations. As a result, robust and generalizable deepfake detection methods have become a critical research focus to safeguard digital media integrity.

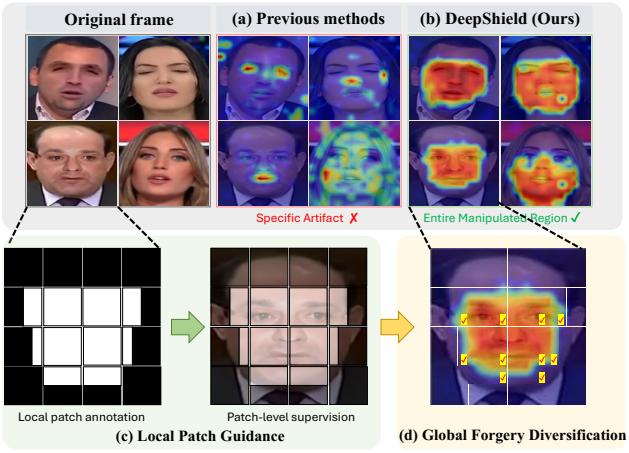


Figure 1. (a)-(b) Comparison of our DeepShield and previous methods in artifact localization using patch-based attention heatmaps. These heatmaps show attention responses from CLIP-ViT for randomly selected image patches. While previous methods emphasize only the most salient specific artifacts using global features, our DeepShield allows patch tokens to capture nuanced details across entire manipulated facial regions. (c)-(d) Illustration of DeepShield's local-to-global learning paradigm: Local Patch Guidance, and Global Forgery Diversification.

While prior deepfake detectors [22, 33, 40, 52] show strong performance in in-domain evaluations, they often fall short in cross-domain scenarios where manipulation techniques differ from those seen at training. Existing approaches [2, 28, 29, 39] attempt to address the generalization challenge by retraining the model on an augmented dataset with deepfake videos produced with a variety of manipulation techniques. However, this strategy is inherently time-consuming, computationally expensive, and limited in scalability, as it requires continuous updates to keep pace with emerging deepfake generation methods. More recently, large-scale vision-language models like CLIP [35]

*Equal Contribution.
†This paper solely reflects the author's personal research and is not associated with the author's affiliated institution.
‡Corresponding Author.

have been adopted for deepfake detection due to their strong generalization capabilities [19, 49]. However, these models are primarily optimized for image-level tasks and tend to rely heavily on global features, making them less effective at capturing fine-grained forgery traces. Moreover, these models tend to concentrate on prominent artifacts within manipulated regions, while overlooking subtle forgery signals that are crucial for identifying more sophisticated deepfakes. As illustrated in Figure 1 (a), these models exhibit strong focus on salient, globally pronounced artifacts in manipulated videos, such as exaggerated transitions in facial regions, while failing to capture subtle forgery traces such as blending boundaries. This tendency to prioritize salient features over nuanced inconsistencies can cause models to overfit to specific forgery types seen during training, further reducing their effectiveness when exposed to novel, cross-domain manipulations.

To address these challenges, we introduce *DeepShield*, a novel framework designed to enhance the model generalization while maintaining sensitivity to both subtle and pronounced manipulations. Towards this goal, DeepShield employs a local-to-global learning paradigm that combines granular patch analysis with diverse forgery representation. At the core of our method are two complementary components: Local Patch Guidance (LPG) and Global Forgery Diversification (GFD), with its conceptual pipeline as illustrated in Figure 1 (c) and (d). Our core idea is to implement deep coupling and iterative collaboration between the proposed LPG and GFD, thus creating a dynamic framework capable of precisely emphasizing local forgery details and effectively uncovering unknown global forgery patterns.

In particular, LPG enhances the model's sensitivity to local forgery inconsistencies by enforcing patch-level supervision, enabling it to capture subtle manipulation patterns that global feature-based approaches often overlook. By establishing stronger semantic correlations between patches, LPG refines the model's ability to detect fine-grained anomalies, significantly improving detection accuracy. As illustrated in Figure 1 (b), LPG successfully identifies minute inconsistencies, such as blending boundaries, across local patches. This balanced attention to both local and global features improves prediction accuracy and finally leads to robust detection of the entire manipulated regions in the deepfake videos. Additionally, we introduce Spatiotemporal Artifact Modeling (SAM) to blend novel deepfake videos with spatial artifacts and temporal inconsistencies. This addresses the lack of local annotations in original training videos while enriching the model with rich and diverse training samples, further improving its generalization capability.

On the other hand, GFD addresses forgery-specific overfitting and enhances the model's generalization across diverse deepfake domains. Specifically, GFD synthesizes a broad spectrum of forgery data, enhancing the model's capabilities through targeted training on this enriched dataset. This data synthesis is achieved by means of domain bridging and boundary expanding. As illustrated in Figure 4 (a), Domain-Bridging Feature Generation (DFG) bridges domain gaps by blending distributions from different forgery types, helping the model learn domain-invariant features and improving its adaptability across various manipulation styles. In the meanwhile, Boundary-Expanding Feature Generation (BFG) further broadens the model's scope by generating features at the boundary of existing forgery domains (See Figure 4 (b)). By scaling the standard deviation of each feature channel, BFG extends the model's detection range to include more subtle and complex forgery types without overlapping with real data clusters. We additionally design a dedicated training objective applied to original and generated samples, combining standard cross-entropy loss with supervised contrastive loss [24]. This objective enables the model to distinguish between real and fake features while reinforcing domain invariance among diverse forgery techniques.

Experimental results demonstrate that LPG and GFD work in tandem to equip DeepShield with the ability to address both local and global challenges in deepfake video detection, resulting in substantial performance improvements, particularly in cross-domain detection scenarios. We summarize our contributions as follows:

- We propose DeepShield, a novel framework for deepfake video detection that combines local and global analysis to improve generalization across diverse manipulation techniques.
- We introduce Local Patch Guidance (LPG), which enhances the model's ability to capture local and fine-grained forgery inconsistencies by explicitly constructing spatiotemporal patches masks to provide structured supervision signals.
- We present Global Forgery Diversification (GFD), which addresses forgery-specific overfitting and enhances cross-domain generalization by synthesizing diverse forgery representations and applying a novel training strategy with supervised contrastive loss.
- Extensive experiments have demonstrated that the combination of LPG and GFD significantly improves the performance of detecting deepfake videos over existing state-of-the-art algorithms, especially in cross-domain settings.

2. Related work

Deepfake Detection Deepfake detection methods are typically categorized into two main types based on the input modality: image-level and video-level detectors. Image-level methods focus on identifying spatial forgery traces, such as subtle facial artifacts [1, 33, 36, 52], inconsistencies in identity information [12, 20], and frequency-

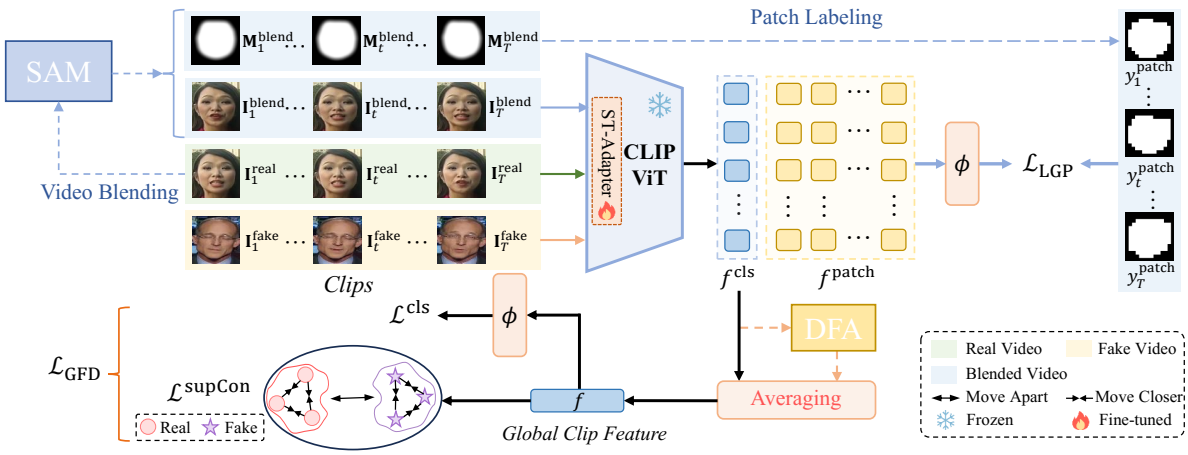


Figure 2. An overview of the DeepShield framework for deepfake video detection. The framework integrates Local Patch Guidance (LPG) and Global Forgery Diversification (GFD) to enhance generalization across diverse manipulation techniques. LPG improves sensitivity to subtle forgery inconsistencies by applying supervised learning on individual video patches, with the proposed Spatiotemporal Artifact Modeling (SAM) to blend deepfake videos. GFD addresses forgery-specific overfitting and strengthens cross-domain generalization by synthesizing diverse forgery representations through Domain Feature Augmentation (DFA). A dedicated training objective, combining standard cross-entropy loss with supervised contrastive loss, is employed to optimize detection performance. This unified approach enables the model to effectively capture both local and global manipulation traces, ensuring robust detection across various deepfake domains.

based anomalies [22, 40]. In contrast, video-level detectors leverage temporal information, analyzing multiple frames to capture dynamic inconsistencies essential for deepfake detection. For instance, STIL [16] integrates two modules to detect both spatial and temporal inconsistencies, while LipForensics [17] focuses on temporal inconsistencies in mouth movements. AltFreezing [46] alternates training spatial and temporal convolutional layers to comprehensively capture spatiotemporal information. These approaches highlight the importance of combining spatial and temporal cues but typically rely on training from scratch with limited labeled data, restricting their generalization to novel forgery types. Unlike existing methods, our proposed DeepShield leverages pre-trained vision representations and synthesizes generalized spatiotemporal artifacts to generate richer supervision signals for local patch learning, significantly boosting detection performance in cross-domain.

Deepfake Detection Using Synthetic Data Drawing inspiration from the blending operations used in deepfake generation, several methods aim to synthesize new training frames with generalized spatial artifacts, such as blending traces and statistical inconsistencies. For example, Face X-Ray [29] blends two frames with similar facial landmarks to reproduce blending artifacts. SBI [39] enhances this by blending a frame with its augmented variant, introducing more realistic blending to improve the model's generalization and make detection more challenging. SeeABLE [28] employs a more subtle approach by blending only small portions of the face, helping the model detect finer inconsistencies. However, these methods tend to overlook temporal discontinuities that arise between frames during deepfake generation, as well as deepfakes that lack prominent spatial artifacts. To address these limitations, AltFreezing [46] generates fake clips with temporal inconsistencies by randomly dropping or repeating frames, but this does not accurately mimic the actual deepfake creation process. Moreover, some studies [30, 32, 49] attempt to blend video frames using landmark misalignment or adjacent-frame inconsistencies, but these methods often introduce unrealistic artifacts significantly different from real scenarios.

Different from these previous algorithms, we propose to synthesize videos with more natural spatiotemporal manipulation traces through mimicking the generation patterns of deepfake videos, aiming to achieve better detection of generalized artifacts. Additionally, recent works [3, 4] attempt to diversify forgery samples in feature space but utilize simple linear interpolation between features from different domains, restricting the richness of feature representation. In contrast, this study introduces a distribution-based feature mixing and expanding, enabling more complex nonlinear augmentation to further enhance the diversity of forgery samples in feature space and improve the generalization capability of detection models.

3. Methodology

3.1. Preliminaries

Overview In this work, we propose *DeepShield*, a novel framework for deepfake video detection that combines local and global analysis to improve generalization. DeepShield consists of two key components: Local Patch Guidance (LPG) and Global Forgery Diversification (GFD). Specifically, LPG applies supervised learning to individual video

patches, improving the model's sensitivity to local subtle forgery traces. In the meanwhile, GFD synthesizes diverse forgery representations and performs global feature learning, addressing forgery-specific overfitting and thus leading to enhanced cross-domain generalization. An overview of DeepShield can be found in Figure 2.

Input and Output Configuration Let $\mathbf{V} \in \mathbb{R}^{T\times H\times W\times 3}$ represent a video clip consisting of T consecutive cropped face frames, where H and W denote the height and width of each frame, respectively. We feed $\mathbf{V}$ into the video encoder E, obtaining the class embedding $f_t^{\text{cls}} \in \mathbb{R}^C$ for each frame $t = 1, \cdots, T$ and their corresponding patch embeddings $f_{t,p}^{\text{patch}} \in \mathbb{R}^C$ for $p = 1, \cdots, P$, where P is the number of patch tokens per frame, and C is the embedding dimension. In this study, we utilize CLIP's pre-trained image encoder, referred to as CLIP-ViT, as the video encoder E. To capture spatial manipulations and temporal inconsistencies efficiently, we fine-tune CLIP-ViT with the parameter-efficient ST-Adapter [34]. The ST-Adapter is inserted before the Multi-Head Self-Attention and Feed-Forward Network in each Transformer block of CLIP-ViT.

3.2. Local Patch Guidance

To improve the model's sensitivity to local forgery anomalies, we propose Local Patch Guidance (LPG), which leverages patch-level supervised learning to strengthen the semantic correlations among patches. A key challenge is that both real and deepfake videos typically lack precise annotations of manipulated areas, making direct patch-level supervision impractical. To address this, we introduce *Spatiotemporal Artifact Modeling (SAM)*, which generates deepfake videos with manipulated masks by blending real videos. This scheme creates precisely labeled training data that highlights generalized spatiotemporal artifacts, facilitating patch-level learning to scale to novel forgery data.

Generating Patch-Level Annotations After obtaining blended fake videos via SAM and retaining original real videos, we divide each video frame and its corresponding manipulation mask into P non-overlapping flattened patches. Let a single frame be of size $H \times W$. Each patch is then of size $H_p = \frac{H}{\sqrt{P}}$ and $W_p = \frac{W}{\sqrt{P}}$. This Patch Division step transforms every frame (and mask) into a collection of smaller local regions, allowing us to assign patch-level forgery labels.

To determine whether a given patch p is "real" or "fake", we define a *Patch Scoring Function* that calculates the number of manipulated pixels within the corresponding mask:

$$\text{PatchMaskScore}(\mathbf{M}_{t,p}) = \sum_{h=1}^{H_p}\sum_{w=1}^{W_p} \mathbb{I}\big[\mathbf{M}_{t,p}^{(h,w)} > 0\big], \quad (1)$$

where $\mathbf{M}_{t,p}$ is the manipulation mask of patch p in frame t,

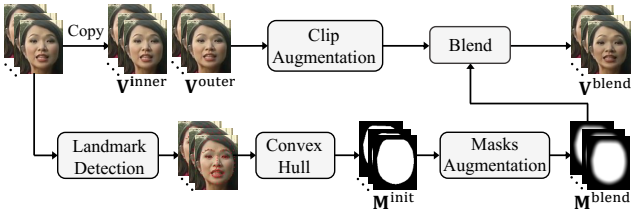


Figure 3. An illustration of the Spatiotemporal Artifact Modeling (SAM), which synthesizes deepfake video clips with generalized spatiotemporal artifacts to facilitate local feature learning.

$\mathbf{M}_{t,p}^{(h,w)}$ is the pixel value at the h-th row and w-th column of $\mathbf{M}_{t,p}$, and $\mathbb{I}[\cdot]$ is an indicator function returning 1 if the value is greater than 0 (manipulated region), and 0 otherwise. Based on a threshold θ, we assign the patch label as:

$$\mathrm{y}_{t,p}^{\text{patch}} = \begin{cases} 1, & \text{if PatchMaskScore}(\mathbf{M}_{t,p}) \geq \theta, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Thus, patches in real videos are labeled 0, while patches in blended deepfake videos are labeled 1 only if they contain enough manipulated pixels. This scheme provides patch-level labels for training without needing manual local annotations.

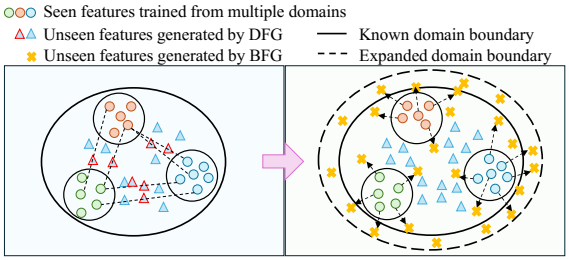
Learning Forgery Features at the Patch Level Once patches are labeled, each patch token serves as an independent training sample in a binary classification task. Concretely, we feed the patch feature $f_{t,p}^{\text{patch}}$ (extracted by the video encoder) into a binary classifier ϕ to estimate $\text{Prob}_{t,p}^{\text{patch}}$, the probability of that patch being fake. The training objective is a binary cross-entropy loss over all patches:

$$\begin{aligned}\mathcal{L}_{\text{LGP}} = -\frac{1}{T\cdot P}\sum_{t,p}\Big[&\mathrm{y}_{t,p}^{\text{patch}}\cdot\log\big(\text{Prob}_{t,p}^{\text{patch}}\big) \\ &+\big(1-\mathrm{y}_{t,p}^{\text{patch}}\big)\cdot\log\big(1-\text{Prob}_{t,p}^{\text{patch}}\big)\Big].\end{aligned} \quad (3)$$

By separately supervising each patch token, the model is guided to cluster patch embeddings of the same category (real or fake) more closely and to separate different ones. This enhances semantic correlations between patches, because the model must learn how local features interact and differ within the broader video context.

In the model's self-attention mechanism, the class token attends to every patch token [13]. When patch embeddings become more discriminative, the class token can aggregate richer local semantics to detect subtle forgeries. As a result, local forgery traces are more effectively perceived, boosting the overall detection performance.

Spatiotemporal Artifact Modeling (SAM) Deepfake videos often exhibit both spatial and temporal inconsistencies. To emulate these characteristics and create manipulated masks for patch-level supervision, we propose



(a) **DFG** (b) **BFG**

Figure 4. An illustration to exemplify our proposed DFA for Domain-Bridging Feature Generation (DFG) and Boundary-Expanding Feature Generation (BFG).

SAM, illustrated in Figure 3. SAM extends the SBI technique [39] (originally designed for image-level blending) to video-level blending, introducing spatial artifacts and temporal inconsistencies in two ways:

(1) Spatial Artifact Generating. For each frame $\mathbf{I}_t$ in a video clip $\{\mathbf{I}_t \mid t = 1, \cdots, T\}$, we produce an inner frame $\mathbf{I}_t^{\text{inner}}$ (providing the face) and an outer frame $\mathbf{I}_t^{\text{outer}}$ (providing the background) from the same original video frame by applying different enhancements (e.g., color, brightness, or sharpness) to one of them, thereby creating statistical discrepancies. We then blend the face region from $\mathbf{I}_t^{\text{inner}}$ into $\mathbf{I}_t^{\text{outer}}$ via a mask $\mathbf{M}_t^{\text{blend}}$ derived from the convex hull of predicted facial landmarks [25] with random deformations and blurring. This yields a blended frame:

$$\mathbf{I}_t^{\text{blend}} = \mathbf{M}_t^{\text{blend}} \odot \mathbf{I}_t^{\text{inner}} + (1 - \mathbf{M}_t^{\text{blend}}) \odot \mathbf{I}_t^{\text{outer}}, \quad (4)$$

where $\odot$ denotes the Hadamard product. By blending statistically inconsistent face and background regions, spatial artifacts are introduced.

(2) Temporal Artifact Generating. To model temporal inconsistencies across frames, SAM applies this blending procedure across all T frames by mimicking the generation patterns of deepfake videos. Key considerations include: 1) one version (either $\mathbf{I}_t^{\text{inner}}$ or $\mathbf{I}_t^{\text{outer}}$) undergoes consistent enhancement across T frames to prevent large disparities that could cause the model to overfit to unnatural forgeries; and 2) the random deformation and blurring of $\mathbf{M}_t^{\text{blend}}$ are consistently applied across T frames, preserving slight differences across frames.

In sum, SAM produces deepfake video clips that contain more generalized spatiotemporal artifacts, offering high-quality manipulated masks for patch-level supervision. By training with these blended videos under our Local Patch Guidance, the model gains a finer-scale perception of forgery traces, significantly improving its capacity to detect local manipulations that may otherwise be overlooked.

3.3. Global Forgery Diversification

To enhance the model's generalization capability, we propose Global Forgery Diversification (GFD). This module synthesizes diverse and domain-invariant forgery representations through *Domain Feature Augmentation (DFA)* and employs a novel training strategy that combines cross-entropy loss with supervised contrastive loss. By mitigating forgery-specific overfitting, GFD ultimately improves the model's ability to detect previously unseen forgery types.

Domain Feature Augmentation (DFA) To diversify forgery features, we introduce Domain Feature Augmentation, which generates novel forgery embeddings using two synthesis strategies: Domain-Bridging Feature Generation (DFG) and Boundary-Expanding Feature Generation (BFG). In contrast to the previous approach [48] that typically relies on linear feature augmentation, DFA exploits distributional characteristics to expose the model to a wider range of potential forgery variations, thereby enhancing its cross-domain generalization. An illustration of the proposed DFA is provided in Figure 4.

(1) *Domain-Bridging Feature Generation (DFG).* Different forgery types exhibit domain gaps that hinder cross-domain generalization. Inspired by [54], DFG mixes distributions from different forgery domains to help the model learn domain-invariant representations. Concretely, we randomly select N clips (each containing T consecutive frames) from a deepfake video in a mini-batch. For each video, we compute the mean μ_c and standard deviation σ_c of the class embeddings for each channel c:

$$\mu_c = \frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{T} f_{n,t,c}^{\text{cls}}, \quad (5)$$

$$\sigma_c = \sqrt{\frac{1}{N \cdot T} \sum_{n=1}^{N} \sum_{t=1}^{T} \left(f_{n,t,c}^{\text{cls}} - \mu_c\right)^2}. \quad (6)$$

Next, we randomly pair each deepfake video with another from a different forgery type within the mini-batch. Let $\tilde{\mu}_c$ and $\tilde{\sigma}_c$ denote the paired video's mean and standard deviation, respectively. We sample a mixing weight $\lambda \sim Beta(0.1, 0.1)$ and interpolate:

$$\begin{aligned} \mu_c^{\text{mix}} &= \lambda \mu_c + (1 - \lambda) \tilde{\mu}_c, \\ \sigma_c^{\text{mix}} &= \lambda \sigma_c + (1 - \lambda) \tilde{\sigma}_c. \end{aligned} \quad (7)$$

Finally, we apply AdaIN [21] to generate domain-bridging features:

$$\hat{f}_{n,t,c}^{\text{cls}} = \sigma_c^{\text{mix}} \cdot \left(\frac{f_{n,t,c}^{\text{cls}} - \mu_c}{\sigma_c} \right) + \mu_c^{\text{mix}}. \quad (8)$$

This procedure "bridges" domain gaps among different forgery types, enabling more robust, domain-invariant representations.

(2) *Boundary-Expanding Feature Generation (BFG).* While domain-bridging features address known forgery spaces, they do not necessarily cover all potential variations. BFG tackles this by synthesizing features that lie on the outer boundaries of existing domain clusters, thereby further en-

larging the forgery domain space. Concretely, we scale the standard deviation of each channel c to slightly shift the features outward:

$$\bar{\sigma}_c = \alpha \cdot \sigma_c, \tag{9}$$

where $\alpha > 1$ (set to $\alpha = 1.1$ in practice) offsets the features beyond the known forgery regions. The boundary-expanding features are then generated via:

$$\bar{f}^{\text{cls}}_{n,t,c} = \bar{\sigma}_c \cdot \left(\frac{f^{\text{cls}}_{n,t,c} - \mu_c}{\sigma_c} \right) + \mu_c, \tag{10}$$

which expands the forgery boundaries without overlapping real feature clusters. By synthesizing such out-of-distribution examples, BFG promotes a decision space that is more generalized against diverse forgeries.

Global Clip Feature Representation For each video clip v, we compute its global clip feature f_v by averaging the frame-level class embeddings:

$$f_v = \frac{1}{T} \sum_{t=1}^{T} f^{\text{cls}}_{v,t}. \tag{11}$$

This aggregated representation captures both spatial and temporal cues essential for generalized forgery detection.

Dedicated Training Objective The objective in GFD promotes discrimination between real and fake samples while encouraging domain invariance across diverse forgery types. Formally, the total loss $\mathcal{L}_{\text{GFD}}$ is composed of a cross-entropy loss $\mathcal{L}^{\text{cls}}$ and a supervised contrastive loss $\mathcal{L}^{\text{supCon}}$:

$$\mathcal{L}_{\text{GFD}} = \mathcal{L}^{\text{cls}} + \upsilon \mathcal{L}^{\text{supCon}}, \tag{12}$$

where υ is a weight balancing the two terms. During training, the model observes original training videos, blended videos from SAM (Section 3.2), and features synthesized by DFA. This broad range of forgery instances compels the model to learn discriminative yet general features. Minimizing $\mathcal{L}_{\text{GFD}}$ sharpens the global representations of both real and fake videos, enhancing cross-domain deepfake detection.

To be more specific, the definitions of the loss functions $\mathcal{L}^{\text{cls}}$ and $\mathcal{L}^{\text{supCon}}$ can be summarized as follows.

(1) *Cross-Entropy Loss ($\mathcal{L}^{\text{cls}}$).* This loss separates real from fake samples:

$$\mathcal{L}^{\text{cls}} = \frac{1}{B} \sum_{v=1}^{B} \text{H}(\text{Prob}_v, \text{y}_v), \tag{13}$$

where H denotes the standard cross-entropy function, B is the total number of original and generated samples in a batch, Prob_v is the predicted probability for sample v, and y_v is the ground-truth label.

(2) *Supervised Contrastive Loss ($\mathcal{L}^{\text{supCon}}$).* Following [24], this loss encourages intra-class compactness and inter-class dispersion:

$$\mathcal{L}^{\text{supCon}} = \frac{1}{B} \sum_{v=1}^{B} -\frac{1}{|\text{J}(v)|} \sum_{i \in \text{J}(v)} L(v,i), \tag{14}$$

$$L(v,i) = \log \frac{\exp(f_v \cdot f_i / \tau)}{\sum_{j \in \text{J}(v) \setminus \{v\}} \exp(f_v \cdot f_j / \tau)},$$

where $\text{J}(v)$ is the set of samples sharing the same class as sample v, f_i and f_j are the global clip features for samples i and j, and τ is a temperature parameter.

Model Optimization We combine $\mathcal{L}_{\text{LPG}}$ and $\mathcal{L}_{\text{GFD}}$ to fine-tune the CLIP-ViT model:

$$\mathcal{L}^{\text{overall}} = \omega \mathcal{L}_{\text{LPG}} + \mathcal{L}_{\text{GFD}}, \tag{15}$$

where ω balances the trade-off between local patch-level supervision and global forgery diversification. By training with this combined objective, the model gains robust detection performance across both local anomalies and diverse global forgeries.

4. Experiments

4.1. Experimental Setups

Datasets To evaluate the generalization capability of the proposed framework, we conduct experiments on the following datasets: FaceForensics++ (FF++) [36], CelebDF v2 (CDF) [31], DeepFake Detection Challenge (DFDC) [11], DFDC Preview (DFDCP) [10], and Deepfake Detection (DFD) [6]. FF++ consists of 1K real videos and 4K fake videos generated using four manipulation strategies including Deepfake (DF) [7], Face2Face (F2F) [42], FaceSwap (FS) [14] and NeuralTexture (NT) [43]. In our experiments, we use the training split of FF++ (HQ, high-quality version) as the default training set, following recent studies [17, 46, 48].

Implementation In the preprocessing phase, we first utilize RetinaFace [8] to detect and crop faces in each frame, and Dlib [25] to detect facial landmarks for generating fake clips. All face crops are then resized to 224×224. For both training and testing, we sample four clips of 12 consecutive frames from each video. Next, we adopt CLIP ViT-B/16 [35] as the backbone of our framework, with the bottleneck width of ST-Adapter set to 384. For classification, we implement a binary classifier ϕ using a linear layer with softmax activation. To train the model, we employ the Adam optimizer with a cosine learning rate scheduler over 80 epochs. Specifically, we set the weight decay, initial learning rate, and batch size to 5×10^{-4}, 3×10^{-4}, and 16, respectively. Empirically, we set ω and υ to 0.5 and 0.5 in Eq. (15), while θ is set to 10 in Eq. (2).

Method	Architecture	Input Type	Testing Set AUC (%)				
			FF++	CDF	DFDCP	DFDC	DFD
Xception [36]	Xception	Frame	99.3	73.7	-	70.9	-
Face X-Ray [29]	HRNet	Frame	97.8	79.5	-	65.5	95.4
SLADD [2]	Xception	Frame	98.4	79.7	76.0	-	-
SBI* [39]	ResNet50	Frame	-	85.7	-	-	94.0
UIA-ViT [55]	ViT-B/16	Frame	99.3	82.4	75.8	-	94.7
SeeABLE [28]	EfficientNet	Frame	-	87.3	86.3	75.9	-
LSDA [48]	EfficientNet	Frame	-	91.1	81.2	77.0	95.6
CNN-GRU [37]	2D+GRU	Video	99.3	69.8	-	68.9	-
FTCN [53]	3D+TF	Video	99.8	86.9	-	74	-
LipForensics [17]	2D+MS-TCN	Video	99.7	82.4	-	73.5	-
RealForensics [18]	2D+CSN	Video	99.8	86.9	-	75.9	-
AltFreezing* [46]	3D	Video	99.7	89.0	-	-	93.7
TALL [47]	Swin-B	Video	**99.9**	90.8	-	76.8	-
StyleFlow [5]	3D+GRU+TF	Video	99.1	89.0	-	-	**96.1**
NACO [50]	2D+TF	Video	99.8	89.5	-	76.7	-
VB-StA [49]	CLIP ViT-B/16	Video	-	86.6	-	77.8	-
DeepShield (Ours)	CLIP ViT-B/16+ST-Adapter	Video	99.2	**92.2**	**93.2**	**82.8**	**96.1**

Table 1. Comparison results (video-level AUC, %) between DeepShield and existing SOTA algorithms on **cross-dataset evaluation**. The comparison methods include both frame-based and video-based approaches, employing various model architectures. * denotes the results for [39] and [46] are taken from [5].

Training Set	Method	DF	F2F	FS	NT	Avg.
DF	DCL [41]	**99.98**	77.13	61.01	75.01	78.28
	UIA-ViT [55]	99.29	74.44	53.89	70.92	74.64
	WATCHER [45]	99.68	75.50	79.51	83.66	84.59
	DeepShield (Ours)	99.94	**94.53**	**80.49**	**87.28**	**90.56**
F2F	DCL [41]	91.91	99.21	59.58	66.67	79.34
	UIA-ViT [55]	83.39	98.32	68.37	67.17	79.31
	WATCHER [45]	85.73	**99.82**	69.38	74.16	82.27
	DeepShield (Ours)	**99.76**	99.50	**83.16**	**88.63**	**92.76**
FS	DCL [41]	74.80	69.75	99.90	52.60	74.26
	UIA-ViT [55]	81.02	66.30	99.04	49.26	73.91
	WATCHER [45]	84.52	75.06	**99.95**	58.81	79.59
	DeepShield (Ours)	**99.77**	**95.78**	99.58	**87.03**	**95.54**
NT	DCL [41]	91.23	52.13	**79.31**	98.97	80.41
	UIA-ViT [55]	79.37	67.89	45.94	94.59	71.95
	WATCHER [45]	89.67	72.85	69.96	98.58	82.77
	DeepShield (Ours)	**99.93**	**97.33**	63.31	**99.03**	**89.90**

Table 2. Comparison results (video-level AUC, %) between DeepShield and existing SOTA algorithms on **cross-manipulation evaluation**. The results on the diagonal show performance in intra-domain evaluation.

Evaluation Metrics Similar to [17, 46, 48, 53], we use Area Under the Curve (AUC) as the evaluation metric and report video-level AUC for fair comparison. For frame-level methods, video-level predictions are obtained by averaging frame-level predictions.

4.2. Comparisons with the State of the Art

To assess the generalization of our framework for deepfake detection, we conduct two types of cross-domain evaluations in this study: cross-dataset evaluation and cross-manipulation evaluation. In cross-dataset evaluation, we train on the full FF++ (HQ) dataset and test on four unseen datasets (CDF, DFDCP, DFDC, and DFD), with in-dataset FF++ testing as a baseline. For cross-manipulation evaluation, we train on one of the four FF++ (HQ) manipulation types (DF, F2F, FS, NT) and test on all four for performance comparison.

Cross-Dataset Evaluation Table 1 illustrates the comparison results between the proposed method and current state-of-the-art (SOTA) algorithms of deepfake video detection in cross-dataset evaluation. As showcased, compared to the SOTA algorithms, while our method does not exhibit a significant advantage on the FF++ test set, it achieves the best performance across four unseen datasets. Notably, our approach excels on the most challenging datasets DFDCP and DFDC, outperforming the previous best methods, SeeABLE [28] and LSDA [48], by 6.9% and 5.8% in AUC, respectively. These findings demonstrate the generalization capacity of the proposed algorithm.

Cross-Manipulation Evaluation Table 2 reports the comparison results of the cross-manipulation evaluation between DeepShield and three other classic methods for detecting deepfake videos: DCL [41], UIA-ViT [55] and WATCHER [45]. The results indicate that our model achieves the highest average performance on both seen and unseen manipulation techniques compared to other detection approaches, consistently excelling across nearly all cases. Notably, in the case where FS is used as the training set, our model outperforms the second-best method WATCHER by 15.95% in AUC on average. In other words, these results underscore the advantage of our method in cross-manipulation evaluation, confirming its strong generalization and effectiveness in deepfake video detection.

4.3. Analysis

We conduct an ablation study to evaluate the contribution of each component in DeepShield, with results presented in Table 3, Table 4, Figure 5, and Figure 6.

How do LPG and GFD Contribute to DeepShield? An ablation study is conducted to evaluate the contributions of LPG and GFD to DeepShield, as summarized in Table 3, Figure 5, and Figure 6. Table 3 first illustrates that the re-

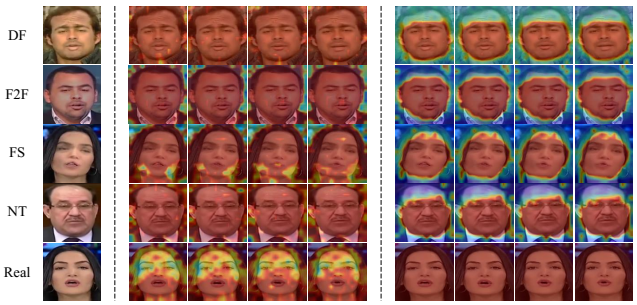


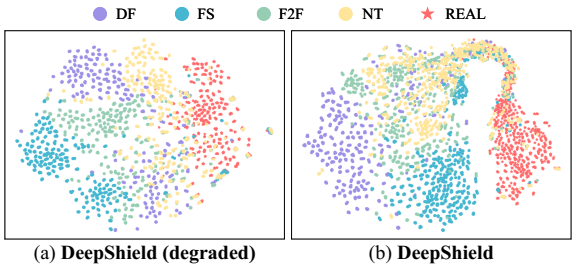
Figure 5. GradCAM [38] visualization of our proposed DeepShield and its variant "DeepShield w/o LPG". We apply Grad-CAM to identify the regions activated for detecting forgery artifacts in both real videos and deepfake videos created with various manipulation techniques. Visualization results are based on intra-dataset scenarios within FF++ (HQ), where warmer colors indicate higher model attention to specific areas.

Variant	CDF	DFDC	DFDCP	Avg.
DeepShield (Ours)	**92.2**	**82.8**	**93.2**	**89.4**
DeepShield w/o LPG	89.1	81.3	87.1	85.8
DeepShield w/o GFD †	89.0	81.9	91.9	87.6
DeepShield w/o LPG & GFD † ‡	85.4	78.4	88.9	84.2

Table 3. Ablation study results (video-level AUC, %) of DeepShield on both LPG and GFD, showing the impact of removing one or both components. Here, † indicates that GFD is removed but the model still uses the original real and deepfake video data given in the training set for fine-tuning CLIP-ViT via standard cross-entropy loss. Additionally, ‡ denotes removal of both LPG and GFD. These experiments conduct model training on FF++ (HQ) and perform cross-dataset evaluations on CFD, DFDC, and DFDCP.

moval of LPG leads to a decrease in average AUC from 89.4% to 85.8%, highlighting its critical role in improving sensitivity to local forgery traces. Specifically, LPG enables the model to focus on manipulated regions in fake frames and distribute attention more evenly in real frames, as visualized in Figure 5. This confirms that LPG enhances the model's capability to detect subtle manipulation traces. Additionally, as shown in Table 3, the exclusion of GFD causes an AUC reduction of 1.8%, demonstrating its significance in mitigating forgery-specific overfitting. By synthesizing diverse forgery representations, GFD enhances the model's adaptability to cross-domain scenarios with varying manipulation styles. Moreover, it also can be observed in Table 3 that removing both LPG and GFD results in a substantial drop of 5.2% in average AUC. As shown in Figure 6, the full model, with both LPG and GFD, achieves superior feature separation between real and manipulated data, underscoring their combined effectiveness in enhancing generalization and detection performance.

How is the Impact of DFA? Here, we investigate how DFA affects the overall detection performance of the proposed model, i.e. DeepShield. To this end, we conduct an ablation study on DFA by removing either DFG or BFG. The experimental results are presented in Table 4. The table shows that removing either DFG or BFG leads to a decrease in testing detection performance across individual datasets and overall. This highlights the importance of synthesizing diverse and domain-invariant forgery representations for generalized deepfake video detection. Notably, the performance of the two feature synthesis strategies in DFA varies across different datasets when applied individually; only their combined effect achieves optimal results in all datasets.



(a) **DeepShield (degraded)** (b) **DeepShield**

Figure 6. Feature Visualization of our proposed DeepShield and its degraded variant using t-SNE [44]. In the degraded version "DeepShield (degraded)", the original CLIP-ViT model is fine-tuned with the ST-Adapter [34] using cross-entropy loss for deepfake video detection. In this experiment, 8 video clips are sampled from each video in the FF++ test set and encoded their features.

DFG	BFG	CDF	DFDC	DFDCP	Avg.
✓	✓	**92.2**	**82.8**	**93.2**	**89.4**
✓	-	91.1	81.9	92.9	88.6
-	✓	90.3	82.1	92.5	88.3
-	-	92.0	81.3	91.6	88.3

Table 4. Ablation study results (video-level AUC, %) of DeepShield on DFA, showing the impact of removing its components, DFG and BFG. These experiments conduct model training on FF++ (HQ) and perform cross-dataset evaluations on CFD, DFDC, and DFDCP.

5. Conclusion

In this work, we propose DeepShield, a novel framework for deepfake video detection that effectively combines local and global analysis to enhance generalization across diverse manipulation techniques. The framework is built upon two complementary components: Local Patch Guidance (LPG) and Global Forgery Diversification (GFD). LPG improves the model's sensitivity to subtle forgery inconsistencies by introducing patch-level supervision, allowing it to detect fine-grained anomalies often overlooked by global methods. Meanwhile, GFD addresses forgery-specific overfitting by synthesizing diverse forgery representations and introducing a novel training strategy with supervised contrastive loss, significantly boosting cross-domain adaptability. Extensive experiments validate that the synergy of LPG and GFD enables DeepShield to outperform existing methods, particularly in challenging cross-domain settings, demonstrating its robustness for real-world deepfake detection.

Acknowledgments

This work is supported in part by the National Key R&D Program of China (2024YFB3908503), and in part by the National Natural Science Foundation of China (62322608).

References

[1] Zhongjie Ba, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren. Exposing the deception: Uncovering more forgery clues for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 719–728, 2024. 2

[2] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18710–18719, 2022. 1, 7

[3] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? *arXiv preprint arXiv:2408.17052*, 2024. 3

[4] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Li Hao, Jiaxin Ai, Qin Zou, Chen Li, and Zhongyuan Wang. Stacking brick by brick: Aligned feature isolation for incremental face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13927–13936, 2025. 3

[5] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1133–1143, 2024. 7

[6] Deepfakedetection, 2021. https://ai.googleblog.com/2019/09/contributing-data-to-deepfakedetection.html Accessed 2024-7-20. 6

[7] DeepFakes, 2020. www.github.com/deepfakes/faceswap Accessed 2024-06-01. 6

[8] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 6

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[10] B Dolhansky. The dee pfake detection challenge (dfdc) pre view dataset. *arXiv preprint arXiv:1910.08854*, 2019. 6

[11] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 6

[12] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9468–9478, 2022. 2

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4

[14] FaceSwap, 2021. www.github.com/MarekKowalski/FaceSwap Accessed 2024-06-01. 6

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[16] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3473–3481, 2021. 3

[17] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5039–5049, 2021. 3, 6, 7

[18] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14950–14962, 2022. 7

[19] Yue-Hua Han, Tai-Ming Huang, Shu-Tzu Lo, Po-Han Huang, Kai-Lung Hua, and Jun-Cheng Chen. Towards more general video-based deepfake detection through facial feature guided adaptation for foundation model. *arXiv preprint arXiv:2404.05583*, 2024. 2

[20] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499, 2023. 2

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 5

[22] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1060–1068, 2022. 1, 3

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 2, 6

[25] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 5, 6

[26] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[27] Peiwen Lai, Weizhi Zhong, Yipeng Qin, Xiaohang Ren, Baoyuan Wang, and Guanbin Li. Llm-driven multimodal and multi-identity listening head generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10656–10666, 2025. 1

[28] Nicolas Larue, Ngoc-Son Vu, Vitomir Struc, Peter Peer, and Vassilis Christophides. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21011–21021, 2023. 1, 3, 7

[29] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020. 1, 3, 7

[30] Maosen Li, Xurong Li, Kun Yu, Cheng Deng, Heng Huang, Feng Mao, Hui Xue, and Minghao Li. Spatio-temporal catcher: A self-supervised transformer for deepfake video detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8707–8718, 2023. 3

[31] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020. 6

[32] Yuzhen Lin, Wentang Song, Bin Li, Yuezun Li, Jiangqun Ni, Han Chen, and Qiushi Li. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In *European Conference on Computer Vision*, pages 104–122. Springer, 2024. 3

[33] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8060–8069, 2020. 1, 2

[34] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477, 2022. 4, 8

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 6

[36] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 2, 6, 7

[37] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3(1):80–87, 2019. 7

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[39] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022. 1, 3, 5, 7

[40] Chao Shuai, Jieming Zhong, Shuang Wu, Feng Lin, Zhibo Wang, Zhongjie Ba, Zhenguang Liu, Lorenzo Cavallaro, and Kui Ren. Locate and verify: A two-stream network for improved deepfake detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7131–7142, 2023. 1, 3

[41] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2316–2324, 2022. 7

[42] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 6

[43] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 6

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[45] Yuan Wang, Chen Chen, Ning Zhang, and Xiyuan Hu. Watcher: Wavelet-guided texture-content hierarchical relation learning for deepfake detection. *International Journal of Computer Vision*, pages 1–22, 2024. 7

[46] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4129–4138, 2023. 3, 6, 7

[47] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668, 2023. 7

[48] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8994, 2024. 5, 6, 7

[49] Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, Yunsheng Wu, and Li Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12615–12625, 2025. 2, 3, 7

[50] Daichi Zhang, Zihao Xiao, Shikun Li, Fanzhao Lin, Jianmin Li, and Shiming Ge. Learning natural consistency rep-

resentation for face forgery video detection. *arXiv preprint arXiv:2407.10550*, 2024. 7

[51] Jiacheng Zhang, Jie Wu, Yuxi Ren, Xin Xia, Huafeng Kuang, Pan Xie, Jiashi Li, Xuefeng Xiao, Weilin Huang, Shilei Wen, Lean Fu, and Guanbin Li. Unifl: Improve latent diffusion model via unified feedback learning. In *Advances in Neural Information Processing Systems*, pages 67355–67382. Curran Associates, Inc., 2024. 1

[52] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 1, 2

[53] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15044–15054, 2021. 7

[54] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021. 5

[55] Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uiavit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European conference on computer vision*, pages 391–407. Springer, 2022. 7