From Specificity to Generality: Revisiting Generalizable Artifacts in Detecting Face Deepfakes

Long Ma[1] **Zhiyuan Yan**[2] **Jin Xu**[4] **Yize Chen**[3]
Qinglang Guo[1] **Zhen Bi**[5 6] **Yong Liao**[1*] **Hui Lin**[7]

[1]School of Cyber Science and Technology, University of Science and Technology of China
[2]School of Electronic and Computer Engineering, Peking University
[3]School of Data Science, The Chinese University of Hong Kong
[4]School of Information Science and Technology, University of Science and Technology of China
[5]Huzhou University [6] Banbu AI Foundation
[7]China Academy of Electronics and Information Technology
{longm@mail, yliao@}ustc.edu.cn

Abstract

Detecting deepfakes has been an increasingly important topic, especially given the rapid development of AI generation techniques. In this paper, we ask: How can we build a universal detection framework that is effective for most facial deepfakes? One significant challenge is the wide diversity of existing deepfake generators, which produced varied types of forgery artifacts (*e.g.,* lighting inconsistency, color mismatch, *etc*). But should we "teach" the detector to learn all these artifacts separately? It is impossible and impractical to elaborate on them all. So the core idea is to pinpoint the more common and general artifacts across different deepfakes. Through systematic analysis of shared technical frameworks in existing deepfake algorithms, we categorize deepfake artifacts into two distinct yet complementary types: Face Inconsistency Artifacts (FIA) and Up-Sampling Artifacts (USA). FIA arise from the challenge of generating all intricate details, inevitably causing inconsistencies between the complex facial features and relatively uniform surrounding areas. USA, on the other hand, are the inevitable traces left by the generator's decoder during the up-sampling process, with all existing deepfakes exhibiting either or both artifacts. Subsequently, we propose a novel image-level pseudo-fake creation framework that constructs fake samples with only the FIA and USA, without introducing extra less-general artifacts. Specifically, we reconstruct the target face to simulate the USA, while utilize image-level blend on diverse facial regions to create the FIA. We surprisingly found that, with this intuitive design, a standard image classifier trained only with our pseudo-fake data can non-trivially generalize well to unseen deepfakes.

1 Introduction

In recent years, A growing number of facial manipulation techniques have advanced due to the rapid development of generative models [35, 36, 58, 75], which facilitate the production of highly realistic and virtually undetectable alterations. As a result, the risks of personal privacy being violated and the spread of misinformation have increased significantly. Therefore, developing a universal deepfake detector that can be used to detect the existing diverse fake methods has become an urgent priority.

*Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

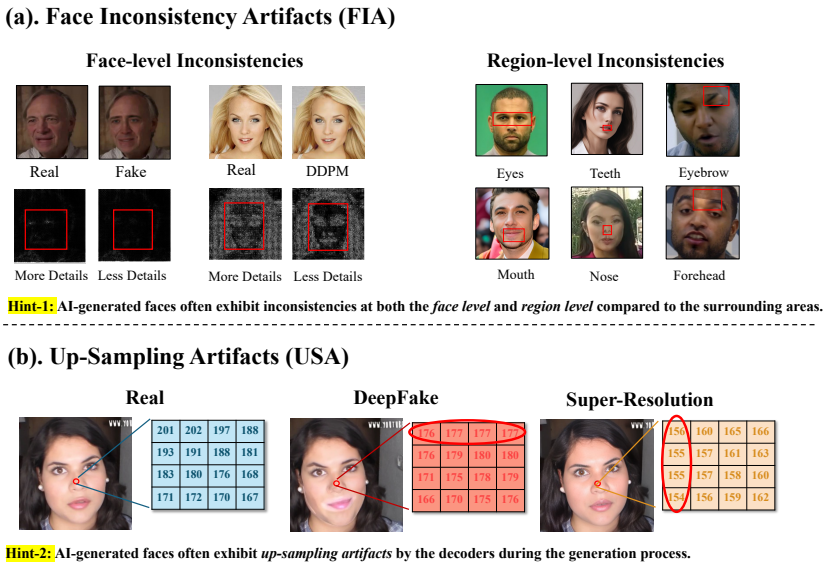


Figure 1: Illustration of the two identified general forgery artifacts across different face deepfakes: (a) Face Inconsistency Artifacts and (b) Up-Sampling Artifacts. We show that the existing face deepfakes typically exhibit both or one of these two artifacts.

In this paper, we pose the question: *How can we build a universal detection framework that is effective for most facial deepfakes?* A significant challenge lies in the wide range of deepfake generators, which lead to different types of forgery artifacts. Specifically, a large number of existing works are dedicated to creating detection algorithms that are carefully designed to identify counterfeit artifacts within specific handcrafted designs, including eye blinking [42], pupil morphology [25], and corneal specularity [32]. However, these methods are mainly effective for identifying particular forgery techniques. But should we train the detector to learn each of these artifacts separately? It is impossible and impractical to cover them all. So, *where should we start from?*
The core idea is to encourage the detection model to learn the more generalizable artifacts. As indicated in previous works [81, 63, 70], different fakes might share some common forgery patterns, and the unlimited deepfake artifacts could be generalized and summarized by the limited generalizable artifacts. Motivated by this, we conduct an in-depth preliminary exploration of the common patterns of existing deepfake artifacts, through which we identify two generalizable deepfake artifacts and categorize them into two distinct yet complementary categories: Face Inconsistency Artifacts (FIA) and Up-Sampling Artifacts (USA), as depicted in Figure 1. FIA represents an inherent limitation that deepfakes struggle to overcome, originating from the generator's incapacity to precisely reproduce facial attributes and nuances, as well as the inescapable disparities between regions induced by amalgamation processes. Unlike FIA, Up-Sampling Artifacts (USA) are not readily apparent to the unaided eye and originate from the generator's inherent up-sampling process that the generator's decoder cannot adequately substitute. Crucially, since decoding and up-sampling constitute fundamental operations in facial manipulation generators, such artifacts become an inevitable byproduct across all methods in this domain. Numerous prior investigations have substantiated this observation [22, 65]. To date, all existing deepfakes characteristically display one or both of these artifact categories.

To enable the model to simultaneously capture FIA and USA, we propose a sophisticated data augmentation method: **FIA-USA**, which leverages super-resolution models [71] and autoencoders [57] to introduce Up-Sampling Artifacts (USA) by reconstructing the face, and employs a strategy of generating multiple masks to blend the reconstructed face with the original, thereby introducing Face Inconsistency Artifacts (FIA) by creating discrepancies at both the facial and regional levels. Augmented with the aforementioned dual artifacts, our method effectively conditions the model to generalize across unseen deepfakes. Furthermore, to maximize the efficacy of FIA-USA, we introduce two complementary technique: **Automatic Forgery-aware Feature Selection (AFFS)**, which streamlines feature dimensionality and augments feature discernment through adaptive selection. **Region-aware Contrastive Regularization (RCR)**, by juxtaposing the features of the manipulated

region against those of the authentic region, RCR enhances the model's focus on the upsampling traces within the manipulated areas and the inconsistencies between different regions.

Extensive experiments conducted on seven widely used Deepfake detection datasets (**encompassing over 58 distinct forgery methods, spanning facial manipulation categories including identity swapping, expression reenactment, generative face synthesis, and attribute editing**) validate the superior efficacy of our method. Detailed ablation experiments and robustness experiments have demonstrated the effectiveness of each component in our method and the robustness of our method to disturbances.

Table 1: **Comparison of our framework and state-of-the-art(SOTA) methods using pseudo-deepfake synthesis.** Our method differs from previous methods in terms of the level at which alterations are applied (local vs global), the introduced artifacts, and the level at which facial inconsistencies are applied.

Methods	Alteration artifact		Introduced Artifacts		
	global	local	Face Inconsistency		Up-Sampling
			face-level	region-level	
Face Xray [40], PCL [89], OST [4]	✓		✓		
SLADD [3]	✓			✓	
SBI [63]	✓		✓		
SeeABLE [37]		✓		✓	
RBI [64]	✓		✓		
Plug-and-Play [83]		✓		✓	
Ours	✓	✓	✓	✓	✓

2 Related Work

Classic Deepfake Detection. Classic deepfake detection can be categorized into multiple aspects: On the one hand, typical deepfake image detection methods encompass frequency-domain analysis [56, 49], identity information [17, 9, 69], contrastive learning [23, 87], network structure improvement [88, 17], watermark [85, 72], robustness [38, 51], interpretability [16, 62] and so forth. On the other hand, deepfake video detection methods include innovative video representation formats [78], self-supervised learning [2, 27, 92], anomaly detection [19, 20], biological signal [76, 8], multi-modal based [54]. In addition to the above two aspects, the the remainder of the work involves forgery localization [51, 26], multi task learning [51, 86], adversarial attack [44], and so on.

Deepfake Detectors Based on Data Synthesis. A notable approach involves synthesizing pseudo-fake faces during training to enhance generalization capability. Early works like Face X-ray [40] pioneered blending-based augmentation by fusing facial regions from different identities to simulate forgery boundaries. Subsequent methods improved upon this paradigm through different artifact simulation strategies: SLADD [3] focused on local adversarial perturbations, SBI [63] proposed self-blending to amplify facial inconsistencies, SeeABLE [37] which proposes fine-grained region-specific blending, and Plug-and-Play [83] employs facial feature masks rather than full-face manipulation for enhanced face synthesis precision. However, as highlighted in Table 1, existing methods predominantly focus on **isolated artifact types** (Mainly focused on FIA) and **single-scale modifications** (global or local), fundamentally limiting their ability to capture the artifacts interplay inherent in real deepfakes. The paradigm of this data augmentation method can be represented as:

$$I_F = M \odot I_t + (1 - M) \odot I_s, \tag{1}$$

where, I_t represents the target face, I_s represents the source face, and M represents the mask. Our proposed FIA-USA method injects new vitality into this paradigm by generating multiple types of masks, introducing USA to I_t, as well as random combination of artifacts.

3 Method

In this section, we begin by outlining two distinct and prevalent types of forgery artifacts found in deepfakes: **Face Inconsistency Artifacts (FIA)** and **Up-Sampling Artifacts (USA)**. We then detail the construction of data augmentation techniques (**FIA-USA**) designed to enable the model to learn to

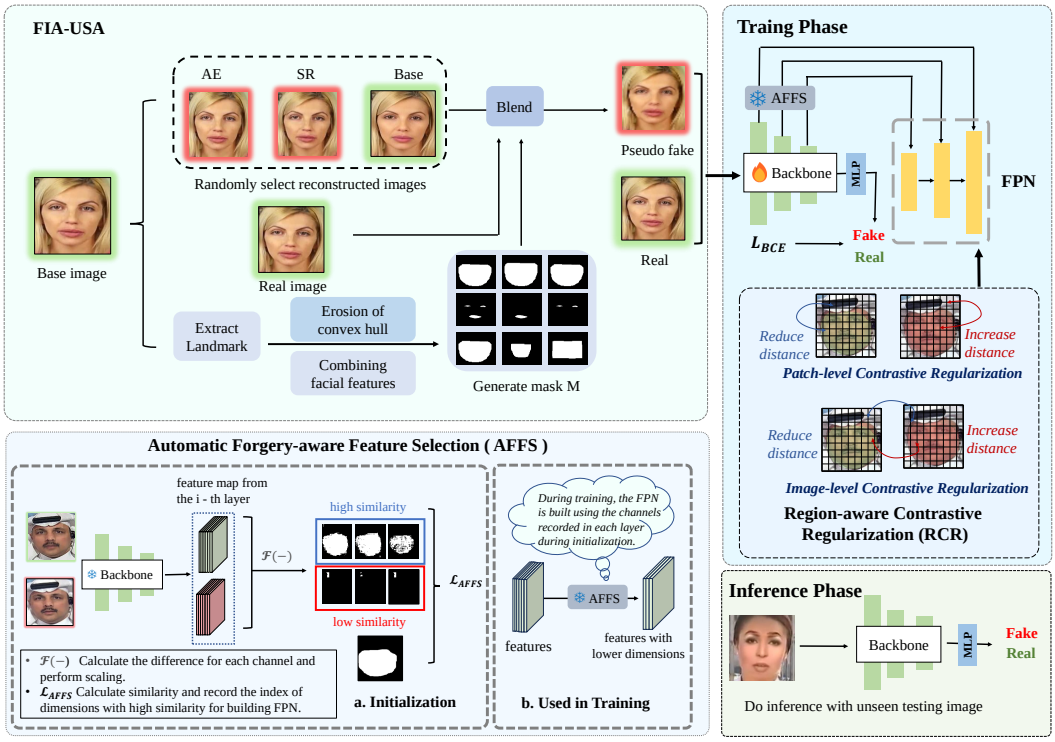


Figure 2: The overall framework of our proposed method consists of: 1) **FIA-USA** enhances the capability of standard classifiers to effectively detect and generalize to unknown deepfake techniques by generating pseudo-fake data containing Face Inconsistency Artifacts (FIA) and Up-Sampling Artifacts (USA). 2) **Region-aware Contrastive Regularization (RCR)** enables the model to focus simultaneously on forged boundaries and spatial artifacts through the contrastive regularization of features from forged and real regions. 3) **Automatic Forgery-aware Feature Selection (AFFS)** achieves efficient feature selection by assessing the similarity between the feature maps of each channel in each layer and the mask.

detect these artifacts concurrently. Subsequently, we present the **Automatic Forgery-aware Feature Selection (AFFS)** and **Region-aware Contrastive Regularization (RCR)**.

3.1 FIA-USA.

Face Inconsistency Artifacts (FIA). As depicted in Figure 1, Face Inconsistency Artifacts (FIA) frequently occur in deepfakes due to the inherent difficulty in rendering intricate details, which often results in discrepancies between intricate facial features and the more uniform surrounding regions. In essence, until deepfake technology advances to the point where it can flawlessly replicate real faces, these artifacts are likely to persist. We explore the origins of Face Inconsistency Artifacts by categorizing deepfake techniques into two domains: *Full Face Synthesis* and *Face Swapping*. In deepfake-based full face synthesis, Face Inconsistency Artifacts inherently persist regardless of the generator's capabilities, while their severity directly correlates with the model's performance. For face swapping, these artifacts arise from two primary sources: **1).** the efficacy of the generator. **2).** the blending boundary created during the face swapping's blending procedure. Furthermore, we can categorize face swapping into two main types based on the forged area: *a. macro-editing* realize face swapping through 4 or 81 facial keypoints. and *b. micro-editing*, which enables granular control over individual facial features. **In section B of the appendix, we provide a detailed description of the classification of forgery techniques and their differences.**

Up-Sampling Artifacts (USA). These artifacts consistently present in synthetic facial regions originate from upsampling operations within the generator's decoder architecture, where interpolation processes inevitably introduce characteristic distortion patterns, with the extent of USA (Up-Sampling Artifacts) introduced varying significantly across different generator implementations. To investigate

more universal up-sampling artifacts, we categorize forgery techniques into two types: One category generates images of forged regions that are as clear and free of blur as the original image. And another type result in blurred tampered areas, requiring additional facial super-resolution models (SR) to process the blurred areas.

FIA-USA. While FIA and USA represent two fundamental artifact categories in deepfakes, conventional data augmentation methods predominantly focus on simulating isolated forgery traces, thereby hindering detection models from effectively learning cross-domain forgery patterns. To bridge this critical gap, we propose FIA-USA—a dual-artifact collaborative enhancement framework engineered for universal detection. Our framework systematically emulates the forgery process through a three-pronged methodology:

1. Multi-Type Mask Synthesis(MTMS): The Multi-Type Mask Synthesis module aims to comprehensively model Face Inconsistency Artifacts (FIA) through a hierarchical mask generation strategy, covering both macro-editing boundaries and micro-editing traces. This process involves two complementary approaches:
a. Macro-Editing Mask. We generate facial contours through two configurations: (1) a high-precision mode using 81 facial keypoints to compute detailed convex hull boundaries, and (2) a rectangular approximation mode derived from 4 keypoints. The initial mask $\mathbb{M}$ is formulated as:

$$\mathbb{M} = \mathcal{H}(K_n), n \in \{4, 81\} \tag{2}$$

where $\mathcal{H}(\cdot)$ denotes convex hull computation and K_n represents the keypoints set. As same as [63], the initial mask $\mathbb{M}$ subsequently undergoes erosion or dilation by applying two Gaussian filters with distinct kernel sizes, where erosion occurs when the kernel of the first filter is larger than that of the second, while dilation manifests when the first kernel is comparatively smaller.
b. Micro-Editing Masks. Given a set of 81 facial keypoints $K_{81} = \{k_1, k_2, \ldots, k_{81}\}$, we first extract two subsets: eyes-specific keypoints $K_{eyes} \subset K$ and mouth-specific keypoints $K_{mouth} \subset K$. Using convex hull computation $\mathcal{H}(\cdot)$, we derive two binary masks: the eye mask $M_e = \mathcal{H}(K_{eyes})$ and the mouth mask $M_m = \mathcal{H}(K_{mouth})$. These masks are then logically combined through union operations $\cup$ to generate three distinct facial regions: (1) $M = M_e$, (2) $M = M_m$, and (3) $M = M_e \cup M_m$ (combined mouth and eyes). Since other facial regions such as eyebrows and nose are rarely targeted in localized facial manipulations (e.g., attribute editing, expression synthesis), these features were intentionally excluded from our mask generation framework to prioritize regions most susceptible to forgery (eyes and mouth).

2. Multi-Modal Reconstruction(MMR): We employ two complementary reconstruction paradigms to reconstruct the target face(I_t) in Formula 1, thereby systematically simulating the Up-Sampling Artifacts (USA) inherent in diverse deepfake generators: *a. Autoencoder (AE) Reconstruction:* Reconstruct I_t through AE [57] to simulate the excessive dependency relationship between adjacent pixels [65] caused by upsampling in GAN / Diffusion models. *b. Super-Resolution (SR) Reconstruction:* Applies SR models [71] to upsample I_t, thereby replicating the characteristic artifacts inherent in deepfake post-processing pipelines. Specifically, given an input face image I, we generate reconstructed versions $I_{AE} = AE(I)$ and $I_{SR} = SR(I)$.

3. Random Artifact Combination(RAC): To ensure comprehensive coverage of artifact interactions, we blend original images(source face) with reconstructed variants(target face) (I_{AE}, I_{SR}, I) using mask M, based on Equation 1. More specifically, given a facial imageI_s, first MTMS generates multiple masks M with equal probability, then MMR generates a reconstructed images (I_{AE}, I_{SR}), and finally RAC blend the I with randomly selected reconstruction variants I_t from (I_{AE}, I_{SR}, I) using a randomly selected mask M. Before blending, I_s and I_t will undergo data augmentation(change color tone, saturation, contrast, blur, etc) to enhance the inconsistency of statistical information.
As shown in Table 1, compared to existing methods, FIA-USA has achieved collaborative artifact enhancement across local / global, facial / regional levels. Please refer to Algorithm 1 for the complete algorithmic process.

3.2 Loss Function

3.2.1 Automatic Forgery-aware Feature Selection

While FIA-USA generates discriminative training samples through dual-artifact augmentation, the inherent feature redundancy in Feature Pyramid Networks (FPN) [45] may drive models to focus on

Algorithm 1 Pseudocode for FIA-USA

Input: Base image I of size $(H, W, 3)$, facial landmarks L of size $(81, 2)$
Output: Image I with FIA and USA

def $\mathcal{T}(I)$ **:** ▷ Source-Target Augmentation
    $I \leftarrow \text{ColorTransform}(I)$
    $I \leftarrow \text{FrequencyTransform}(I)$
    return I
def $Recon(I)$ **:** ▷ Reconstruct the source image
    if $\text{Uniform}(\min = 0, \max = 1) \in [0, 0.3]$ **:**
        $I_t \leftarrow AE(I)$
    else if $\text{Uniform}(\min = 0, \max = 1) \in (0.3, 0.5]$ **:**
        $I_t \leftarrow SR(I)$
    else
        $I_t \leftarrow I$
    return I
$I_t, I_s \leftarrow Recon(I), I$
if $\text{Uniform}(\min = 0, \max = 1) < 0.5$ **:**
    $I_s, I_t \leftarrow \mathcal{T}(I_s), I_t$
else :
    $I_s, I_t \leftarrow I_s, \mathcal{T}(I_t)$
$M \leftarrow \text{CombineFacialFeatures}(L)$ or $\text{ConvexHull}(L)$
$I_{PF} \leftarrow I_s \odot M + I_t \odot (1 - M)$

non-critical forgery patterns, thereby compromising detection generalization. This limitation stems from FPN's naive aggregation of all channel-wise features without adaptively emphasizing FIA/USA-correlated representations. To address this limitation, we propose Automatic Forgery-aware Feature Selection (AFFS), a statistically-driven feature compression paradigm that quantifies channel-wise sensitivity to forgery regions, thereby dynamically constructing lightweight yet discriminative feature pyramids.

Given a pretrained backbone network ϕ, let $f_i = \phi_i(I) \in \mathbb{R}^{W_i * H_i * C_i}$ denote the feature map from the i-th layer, where $I \in \mathbb{R}^{H*W*3}$ represents the input image. To identify artifact-sensitive feature dimensions, we leverage pseudo-fake pairs $\{R_n, F_n, M_n\}_{n=1}^{N}$ generated by FIA-USA and compute the normalized response discrepancy between real and forged regions. For the k-th channel $f_i^k \in \mathbb{R}^{W_i * H_i}$ in layer i, the selection criterion is formulated as:

$$\mathcal{L}_{AFFS}^{i,k} = \frac{1}{N} \sum_{n=1}^{N} \|\mathcal{F}(f_i^k(R_n) - (f_i^k(F_n)) - M_n\|_2^2 \tag{3}$$

where $\mathcal{F}$ denotes feature normalization and spatial interpolation to align f_i^k with mask M_n. Channels are ranked by ascending $\mathcal{L}_{AFFS}^{i,k}$ and the top-m_i channels with minimal errors are retained to form a compressed feature map $f_i' \in \mathbb{R}^{W_i * H_i * m_i}$ for FPN construction. **AFFS leverages a pre-trained model as initialization, operating as a supervised feature ranking mechanism that reduces feature redundancy.**

3.2.2 Region-aware Contrastive Regularization

Building upon the artifact-sensitive features selected by AFFS, we design Region-aware Contrastive Regularization (RCR) to explicitly model the divergence between forged and authentic regions through multi-granularity contrastive learning. Given the Feature Pyramid Network (FPN) P, we denote the feature maps of real and fake faces as $H_R = P(\phi(R))$, $H_F = P(\phi(F))$. For H_R, the areas inside and outside M are real, while for H_F, the areas inside M are manipulated and the areas outside M are real. We define the region within M as H_R^{in}, H_F^{in} and the region outside M as H_R^{out}, H_F^{out}. RCR operates on the refined feature pyramid from AFFS, implementing multi-granularity contrast through:

Patch-level Contrastive Regularization. For real faces, the features f of H_R^{in} and H_R^{out} are consistent, thus we aim to minimize the distance between them. Conversely, for fake faces, the features f of H_F^{in} and H_F^{out} are distinct, thus we aim to maximize the distance between them. Therefore, the loss function of **Patch-level Contrastive Regularization (PCR)** can be formulated as:

$$\mathcal{L}_1 = -\log \frac{\sum\limits_{f_R \in H_R} e^{\delta(f_R^{in}, f_R^{out})/\tau}}{\sum\limits_{f_R \in H_R} e^{\delta(f_R^{in}, f_R^{out})/\tau} + \sum\limits_{f_F \in H_F} e^{\delta(f_F^{in}, f_F^{out})/\tau}}, \tag{4}$$

where f^i represents a pixel feature, τ is a temperature parameter and $\delta(\cdot)$ is the normalized cosine similarity between two features, as:

$$\delta(f_1, f_2) = \sum_{f_1^i \in f_1} \sum_{f_2^i \in f_2} \frac{f_1^i}{\|f_1^i\|} \cdot \frac{f_2^i}{\|f_2^i\|}, \tag{5}$$

Image-level Contrastive Regularization. Since both the region of real face and fake face outside M are real, we aim to minimize the distance between them and maximize the distance between the fake face and the real face in the area inside M. Therefore, the loss function for **Image-level Contrastive Regularization (ICR)** can be formulated as:

$$\mathcal{L}_2 = -\log \frac{\sum\limits_{f_{out} \in H^{out}} e^{\delta(f_R^{out}, f_F^{out})/\tau}}{\sum\limits_{f^{out} \in H_{out}} e^{\delta(f_R^{out}, f_F^{out})/\tau} + \sum\limits_{f^{in} \in H^{in}} e^{\delta(f_R^{in}, f_F^{in})/\tau}}. \tag{6}$$

Overall Loss. The network is optimized using the following loss:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{BCE} + \lambda_2 \mathcal{L}_1 + \lambda_3 \mathcal{L}_2, \tag{7}$$

where $\mathcal{L}_{BCE}$ denotes the cross-entropy classification loss. $\mathcal{L}_{BCE}$, $\mathcal{L}_1$ and $\mathcal{L}_2$ are weighted by the hyperparameters λ_1 ,λ_2 and λ_3, respectively.

4 Experiments

4.1 Settings

Datasets. To evaluate the effectiveness of our proposed method, we conducted extensive experiments on **seven widely-adopted benchmark datasets** spanning both classical facial manipulation paradigms and emerging generative deepfake architectures.**a. Traditional Deepfake Datasets:** 1. FaceForensics++ (FF++) [59], 2. Deepfake Detection (DFD) [12], 3. Deepfake Detection Challenge (DFDC) [15], 4. preview version of DFDC (DFDCP) [14], and 5. CelebDF (CDF) [43]. FF++ comprises 1,000 original videos and 4,000 fake videos forged by four manipulation methods, namely, Deepfakes(DF) [13], Face2Face(F2F) [68], FaceSwap(FS) [18], and NeuralTextures(NT) [67]. FF++ offers three levels of compression: raw, lightly compressed (c23), and heavily compressed (c40), **under storage constraints, our implementation adopts the FF++_c23 (including DFD) with training conducted following the SBI protocol [63], employing exclusively real facial samples from FF++_c23 subset.** Although many previous studies have utilized the same dataset for both training and testing, the preprocessing and experimental configurations can differ, making fair comparisons difficult. Therefore, in addition to testing on the raw data of the aforementioned datasets, we also performed generalization assessments on the unified new benchmark for traditional deepfakes, (DeepfakeBench) [82]. **b. Generative Deepfake Datasets.** 6. Diffusion Facial Forgery (DiFF) [5] contains over 500,000 images synthesized using 13 different generation methods under four conditions: Text to Image (T2I), Image to Image (I2I), Face Swapping (FS) and Face Editing (FE). **Text to Image** encompasses four generation methods: *Midjourney* [50], *Stable Diffusion XL (SDXL)* [55], *Free-DoM_T* [84], and *HPS* [77]. **Image to Image** comprises *Low Rank Adaptation(LoRA)* [31], *DreamBooth* [60], *SDXL Refiner* [55], and *Free-DoM_I*. **Face Swapping** includes *DiffFace* [36] and *DCFace* [36]. **Face Editing** comprises *Imagic* [35], *Cycle Diffusion(CycleDiff)* [75], and *Collaborative Diffusion(CoDiff)* [34]. **7. DF40 [80] encompasses 40 state-of-the-art deepfake generation methodologies spanning four core categories: face swapping, facial reenactment, full-face synthesis, and advanced facial manipulation.** Significantly surpassing existing benchmarks in both scope and volume, DF40 integrates cutting-edge generative architectures from 2024 alongside widely adopted commercial software solutions.

Evaluation Metrics. We report the Frame-Level Area Under Curve (AUC) metric on the DeepfakeBench and DIFF dataset, report the Video-Level Area Under Curve (AUC) metric on the traditional deepfake datasets, to compare our proposed method with prior works.

Implementation Details. We adopt EfficientNetB4 [66] as the backbone network architecture (We also

Table 2: **Cross-dataset Evaluations:** Cross-dataset evaluations were conducted using the **Frame-level AUC**. All experiments were trained on the **c23 version of FF++** and tested on other datasets. *** indicates the results are cited from [10] and † indicates our reproduction results using the checkpoints provided by the authors, otherwise, the results are from DeepfakeBench [82].**

Detector	Backbone	Venues	CDF-v1	CDF-v2	DFD	DFDCP	Avg.
FWA [41]	Xception	CVPRW'18	0.790	0.668	0.740	0.638	0.714
CapsuleNet [52]	Capsule	ICASSP'19	0.791	0.747	0.684	0.657	0.720
CNN-Aug [29]	ResNet	CVPR'20	0.742	0.703	0.646	0.617	0.677
Face X-ray [40]	HRNet	CVPR'20	0.709	0.679	0.766	0.694	0.712
FFD [11]	Xception	CVPR'20	0.784	0.744	0.802	0.743	0.768
F3Net [56]	Xception	ECCV'20	0.777	0.798	0.702	0.735	0.749
SPSL [47]	Xception	CVPR'20	0.815	0.765	0.812	0.741	0.783
SRM [49]	Xception	CVPR'21	0.793	0.755	0.812	0.741	0.775
CORE [53]	Xception	CVPRW'22	0.780	0.743	0.802	0.734	0.764
Recce [2]	Designed	CVPR'22	0.768	0.732	0.812	0.734	0.761
SBI* [62]	EfficientNet-B4	CVPR'22	-	0.813	0.774	0.799	-
UCF [81]	Xception	ICCV'23	0.779	0.753	0.807	0.759	0.774
ED* [1]	ResNet-34	AAAI'24	0.818	<u>0.864</u>	-	**0.851**	-
ProDet† [6]	EfficientNet-B4	NeurIPS'24	**0.909**	0.842	<u>0.848</u>	0.774	0.843
LSDA* [79]	EfficientNet-B4	CVPR'24	0.867	0.830	**0.880**	0.815	<u>0.848</u>
Forensic-Adapter* [10]	CLIP (ViT-B/16)	CVPR'25	-	0.837	-	0.799	-
Ours	EfficientNet-B4	-	<u>0.901</u>	**0.867**	0.821	<u>0.818</u>	**0.852**

investigate alternative network architectures and their respective outcomes), trained for 50 epochs using the SAM optimizer [21] with a batch size of 12 and initial learning rate of 0.001. For video processing, each input is uniformly sampled into 32 frames during both training and inference phases. Our data augmentation pipeline combines the proposed FIA-USA strategy with conventional techniques including RandomHorizontalFlip, RandomCutOut, and AddGaussianNoise. The loss coefficients $\lambda_1, \lambda_2, \lambda_3$ are empirically set to 1, 2.5, and 0.25 respectively (We also explored the impact of other variants on the detection results), with the temperature parameter τ fixed at 0.7. All experiments were conducted on a single NVIDIA 3090.

4.2 Generalization Evaluation

4.2.1 Traditional Deepfake Datasets.

We first compare our method with previous work at the **frame level**. To ensure the fairness of the experiment, we conducted the experiment on the DeepfakeBench, and adhered to the data preprocessing and experimental settings provided by them. For previous work, we utilized the experiment data provided by DeepfakeBench. As shown in Table 2, our method outperforms other methods on the majority of deepfake datasets and competes with state-of-the-art methods on the CDF-v1 dataset. In addition to comparing at the frame-level, we also compared our method at the **video-level** with state-of-the-art detection algorithms, including various data augmentation methods. The comparison results, as shown in Table 3, strongly demonstrate the effective generalization of our method.

Table 3: **Comparison with SOTA methods using the Video-Level AUC.**

Model	Venues	CDF-v2	DFDC	DFDCP
Face X-Ray [40]	CVPR'20	-	-	0.711
LipForensics [28]	CVPR'21	0.824	0.735	-
FTCN [90]	ICCV'21	0.869	-	0.740
PCL + I2G [89]	ICCV'21	0.900	0.675	0.743
HCIL [24]	ECCV'22	0.790	0.692	-
ICT [17]	CVPR'22	0.857	-	-
SBI [63]	CVPR'22	<u>0.928</u>	0.719	<u>0.855</u>
AltFreezing [74]	CVPR'23	0.895	-	-
SAM [7]	CVPR'24	0.890	-	-
LSDA [79]	CVPR'24	0.911	**0.770**	-
CFM [48]	TIFS'24	0.897	-	0.802
LAA-Net† [51]	CVPR'24	0.840	-	0.741
Ours	-	**0.941**	<u>0.732</u>	**0.866**

Table 4: **Comparison with universal deepfake detection methods using the Frame-Level AUC on the DiFF dataset.** † indicates models that are designed for deepfake detection, while ‡ signifies models intended for generated image detection.

Method	Test Subset			
	T2I	I2I	FS	FE
Xception† [59]	62.43	56.83	85.97	58.64
F^3-Net† [56]	66.87	67.64	81.01	60.60
EfficientNet† [66]	74.12	57.27	82.11	57.20
DIRE‡ [73]	44.22	64.64	84.98	57.72
SBI† [63]	<u>80.20</u>	<u>80.40</u>	<u>85.08</u>	<u>68.79</u>
Ours	**86.05**	**84.95**	**89.42**	**72.73**

4.2.2 Generative Deepfake Datasets

Our evaluation on DIFF and DF40 datasets demonstrates the superior generalization capabilities of our architecture compared to state-of-the-art detection methods. Comprehensive metrics are detailed in Table 4 and Table 5. As shown in Table 5, our method achieves an average AUC of **87.8%**, surpassing RECCE by **9.7%** and outperforming SBI by **23.4%**, and enables **10%** higher AUC than LSDA's latent space augmentation on

Table 5: **Comparative Analysis of Detection Performance Between the Proposed Method and State-of-the-Art Approaches on Six Representative Face Swapping Forgery Types in DF40 [80].**

Method	Venues	uniface	e4s	facedancer	fsgan	inswap	simswap	Avg.
RECCE [2]	CVPR 2022	84.2	65.2	78.3	88.4	79.5	73.0	78.1
SBI [63]	CVPR 2022	64.4	69.0	44.7	87.9	63.3	56.8	64.4
CORE [53]	CVPRW 2022	81.7	63.4	71.7	**91.1**	79.4	69.3	76.1
IID [33]	CVPR 2023	79.5	71.0	79.0	86.4	74.4	64.0	75.7
UCF [81]	ICCV 2023	78.7	69.2	80.0	88.1	76.8	64.9	77.5
LSDA [79]	CVPR 2024	85.4	68.4	75.9	83.2	81.0	72.7	77.8
CDFA [46]	ECCV 2024	76.5	67.4	75.4	84.8	72.0	76.1	75.9
ProgressiveDet [6]	NeurIPS 2024	84.5	71.0	73.6	86.5	78.8	77.8	78.7
Ours	-	**91.8**	**87.5**	**83.0**	86.3	**87.4**	**91.0**	**87.8**

(a) **Ablation Study of Method Components.** We report the Video-Level AUC for traditional deepfake datasets and the Frame-Level AUC for generative deepfake datasets. SR signifies the process of super-resolution, and AE refers to the reconstruction achieved through autoencoders.

Setting	Test Dataset				Avg.
	Traditional		Generative		
	CDF-v2	DFDCP	FS	FE	
w/o MTMS	91.99	87.06	88.00	71.93	84.74
w/o SR	93.25	86.81	80.73	61.07	80.46
w/o AE	92.53	86.37	88.67	70.31	84.47
w/o FIA-USA	93.18	89.45	80.68	62.99	81.57
w/o PCR	90.12	86.13	89.61	66.87	83.18
w/o ICR	90.85	85.00	89.95	75.01	85.20
w/o RCR	90.14	86.13	89.61	66.87	83.18
w/o AFFS	94.70	86.30	88.33	70.62	84.98
Ours	94.10	86.66	89.42	72.73	**85.72**

(b) **Ablation Study on Model Architectures.**

Model	Test Dataset				Avg.
	Traditional		Generative		
	CDF-v2	DFDCP	FS	FE	
Res50	82.45	79.14	68.82	58.59	72.25
Res101	85.51	82.35	71.26	60.44	74.89
Effb1	91.88	82.30	83.73	70.69	82.15
Effb4	94.10	86.66	89.42	72.73	**85.72**

(c) **Experiment on Hyperparameter Configuration.**

$\lambda_1, \lambda_2, \lambda_3$	Test Dataset				Avg.
	Traditional		Generative		
	CDF-v2	DFDCP	FS	FE	
1, 1, 1	90.86	85.01	89.91	70.19	83.99
1, 1, 0.25	91.88	86.01	89.91	72.19	85.00
1, 2.5, 1	91.82	86.00	89.95	71.01	84.70
1, 0.25, 2.5	91.11	85.59	89.61	70.92	84.30
1, 2.5, 0.25	94.10	86.66	89.42	72.73	**85.72**

Table 6: **Ablation experiment.** We conducted ablation experiments on method components, model frameworks, and hyperparameter configuration. The best results are presented in bold.

average. The results notably contradict recent findings [79] about RGB-based methods' limitations against generative deepfakes, proving that properly designed RGB-level artifacts retain critical discriminative signals. **For comprehensive experimental results on forgery techniques supported by the DF40 and DIFF, please refer to the extended analysis in Appendix Section G.**

4.2.3 Robustness

In Figure 3, we assessed the influence of different levels of random perturbations on the detection performance. We quantified the effects of different levels of Gaussian Blur, Block Wise, Change Contrast, and JPEG Compression on the detectors. Notably, to gauge accuracy, we employed the reduction rate of Video-Level AUC to assess the robustness of the detector to different degrees of disturbance, which could be expressed as $R_{AUC} = (AUC - AUC_{raw})/AUC_{raw}$, where AUC_{raw} refers to no perturbations. As shown in the Figure 3, our method exhibits superior robustness compared to other methods.

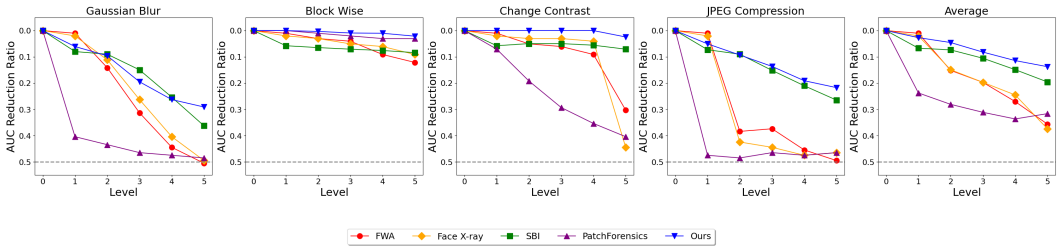


Figure 3: **Robustness to Unseen Perturbations.** We reported the Video-Level **AUC Reduction Ratio** for four specific types of perturbations at five different levels.

4.3 Ablation Study

The impact of method components. To systematically evaluate the contribution of each component in our framework, we conducted comprehensive ablation studies across seven benchmark datasets. As detailed in Table 6a, our analysis focuses on three core innovations: (1) the FIA-USA augmentation mechanism, (2) Automatic Forgery-aware Feature Selection (AFFS), and (3) Region-aware Contrastive Regularization (RCR). We adopt video-level AUC for traditional deepfakes and frame-level AUC for generative deepfakes. The empirical results demonstrate progressive performance degradation when removing individual components, with the complete framework achieving optimal detection accuracy . The combination of FIA-USA, AFFS, and RCR yields optimal performance. As intended in our design, AFFS and RCR are crucial for fully leveraging the potential of FIA-USA. Removing any component degrades performance, with the absence of FIA-USA having the most significant impact (a 4.15% decrease). It is important to clarify that "w/o FIA-USA" specifically refers to using only the data augmentation proposed by SBI. Furthermore, RCR exerts a stronger influence on FIA-USA's effectiveness compared to AFFS.

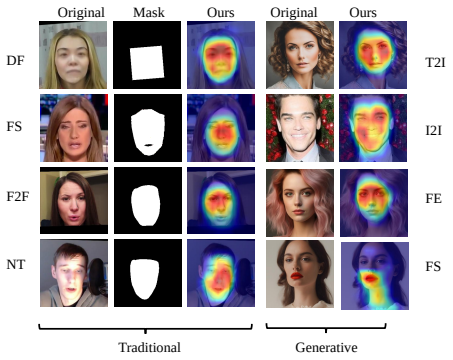


Figure 4: **GradCAM visualization of fake samples across various deepfake datasets.**

The ablation specifically targeting FIA-USA (components within it) reveals: Various data augmentation techniques contribute differently. SR and AE are shown to be effective for detecting generative forgeries. Conversely, MTMS appears to suppress performance on this specific forgery type. This aligns with our theoretical analysis: generative forgeries constitute global manipulations that produce fewer Face Inconsistency Artifacts (FIA) but primarily exhibit Up-Sampling Artifacts (USA). As seen from MTMS's performance on traditional forgeries, increasing the diversity of FIA proves beneficial for detecting face swapping forgery.

Regarding the RCR ablation: The results align well with our chosen hyperparameters. They indicate that the PCR contributes substantially more to the model's performance than the ICR. This observation is consistent with the hyperparameter sensitivity analysis presented in Table 6c: increasing the loss weight for PCR while decreasing that for ICR leads to significant performance gains. Conversely, increasing the ICR weight while reducing the PCR weight results in only marginal improvements.

The impact of model framework. Our architectural analysis in Table 6b systematically benchmarks detection performance across backbone networks, quantitatively comparing the effects of EfficientNet [66], ResNet [29], and different depth configurations on detection performance, demonstrates the compatibility of our method with various network frameworks, while also illustrating the critical impact of model parameter size and architectural design on detector performance. This highlights the ongoing importance of developing more robust and specialized deep learning architectures for counterfeit detection tasks.
Impact of hyperparameters. We also examined the effect of hyperparameters of the loss function on model performance, as detailed in the Table 6c. Our analysis of the weights revealed that increasing the weight of λ_2 relative to the weights of cross entropy enhances model performance, whereas decreasing the weight of λ_3 relative to cross entropy also improves performance, which is consistent with the ablation experiment results for RCR components in Table 6a. However, conversely, model performance will not be significantly improved.

5 Visualizations

Our GradCAM [91] analysis in Figure 4 demonstrates precise localization of facial forgery artifacts across both conventional and emerging generative architectures. Notably, the visualization framework not only pinpoints manipulation traces in traditional deepfakes but also effectively in state-of-the-art synthetic media, validating our method's generalization capability.

6 Conclusion

In conclusion, this study presents a universal framework for deepfake detection by focusing on common artifacts that span various types of face forgeries. By categorizing deepfake artifacts into Face Inconsistency Artifacts (FIA) and Up-Sampling Artifacts (USA), we enhance the generalization capability of detection models. This targeted approach allows the model to focus on detecting fundamental artifact patterns, and potentially improving its performance across diverse deepfake variations. In doing so, our framework may help address some challenges in current detection methods and could provide a promising foundation for developing more adaptable deepfake detection systems in future research.

Acknowledgments and Disclosure of Funding

This work was supported in part by Key Science & Technology Project of Anhui Province 202423l110050033, Zhejiang Provincial Natural Science Foundation of China under Grant (No. LQN25F020023).

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*, volume 38, pages 719–728, 2024.

[2] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022.

[3] 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.

[4] Liang Chen, Yong Zhang, Yibing Song, Jue Wang, and Lingqiao Liu. Ost: Improving generalization of deepfake detection via one-shot test-time training. *Advances in Neural Information Processing Systems*, 35:24597–24610, 2022.

[5] Harry Cheng, Yangyang Guo, Tianyi Wang, Liqiang Nie, and Mohan Kankanhalli. Diffusion facial forgery detection. *arXiv preprint arXiv:2401.15859*, 2024.

[6] 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.

[7] 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 (CVPR)*, pages 1133–1143, June 2024.

[8] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[9] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15108–15117, 2021.

[10] Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. *arXiv preprint arXiv:2411.19715*, 2024.

[11] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020.

[12] Deepfakedetection, 2021. `https://ai.googleblog.com/2019/09/contributing-data-to-deepfakedetection.html` Accessed 2021-11-13.

[13] DeepFakes, 2020. `www.github.com/deepfakes/faceswap` Accessed 2020-09-02.

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

[15] 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.

[16] Shichao Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. Explaining deepfake detection by analysing image matching. In *European conference on computer vision*, pages 18–35. Springer, 2022.

[17] 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.

[18] FaceSwap, 2021. `www.github.com/MarekKowalski/FaceSwap` Accessed 2020-09-03.

[19] Jianwei Fei, Yunshu Dai, Peipeng Yu, Tianrun Shen, Zhihua Xia, and Jian Weng. Learning second order local anomaly for general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20270–20280, 2022.

[20] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10491–10503, 2023.

[21] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[22] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020.

[23] Sheldon Fung, Xuequan Lu, Chao Zhang, and Chang-Tsun Li. Deepfakeucl: Deepfake detection via unsupervised contrastive learning. In *2021 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2021.

[24] Zhihao Gu, Taiping Yao, Yang Chen, Shouhong Ding, and Lizhuang Ma. Hierarchical contrastive inconsistency learning for deepfake video detection. In *European Conference on Computer Vision*, pages 596–613. Springer, 2022.

[25] Hui Guo, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. Eyes tell all: Irregular pupil shapes reveal gan-generated faces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2904–2908. IEEE, 2022.

[26] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023.

[27] 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.

[28] 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.

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

[30] Hillobar, 2023. `https://github.com/Hillobar/Rope`, Accessed 2023.

[31] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[32] Shu Hu, Yuezun Li, and Siwei Lyu. Exposing gan-generated faces using inconsistent corneal specular highlights. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2500–2504. IEEE, 2021.

[33] 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.

[34] Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6080–6090, 2023.

[35] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023.

[36] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Diffface: Diffusion-based face swapping with facial guidance. *arXiv preprint arXiv:2212.13344*, 2022.

[37] 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.

[38] Binh M Le and Simon S Woo. Quality-agnostic deepfake detection with intra-model collaborative learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22378–22389, 2023.

[39] Hanzhe Li, Jiaran Zhou, Bin Li, Junyu Dong, and Yuezun Li. Freqblender: Enhancing deepfake detection by blending frequency knowledge. *arXiv preprint arXiv:2404.13872*, 2024.

[40] 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.

[41] Y Li. Exposing deepfake videos by detecting face warping artif acts. *arXiv preprint arXiv:1811.00656*, 2018.

[42] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. Ieee, 2018.

[43] 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.

[44] Jiawei Liang, Siyuan Liang, Aishan Liu, Xiaojun Jia, Junhao Kuang, and Xiaochun Cao. Poisoned forgery face: Towards backdoor attacks on face forgery detection. *arXiv preprint arXiv:2402.11473*, 2024.

[45] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[46] 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.

[47] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021.

[48] Anwei Luo, Chenqi Kong, Jiwu Huang, Yongjian Hu, Xiangui Kang, and Alex C. Kot. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19:1168–1182, 2024.

[49] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16326, 2021.

[50] Midjourney, 2022. `https://www.midjourney.com`, Accessed 2022.

[51] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17395–17405, 2024.

[52] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2307–2311. IEEE, 2019.

[53] Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12–21, 2022.

[54] Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27102–27112, 2024.

[55] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[56] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020.

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[58] Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, and Cristofer Englund. Facedancer: Pose-and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3454–3463, 2023.

[59] 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.

[60] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023.

[61] Somdev Sangwan, 2023. `https://github.com/s0md3v/roop`, Accessed 2023.

[62] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and recovering sequential deepfake manipulation. In *European Conference on Computer Vision*, pages 712–728. Springer, 2022.

[63] 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.

[64] Yuyang Sun, Huy H Nguyen, Chun-Shien Lu, ZhiYong Zhang, Lu Sun, and Isao Echizen. Generalized deepfakes detection with reconstructed-blended images and multi-scale feature reconstruction network. *arXiv preprint arXiv:2312.08020*, 2023.

[65] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the upsampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024.

[66] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[67] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Journal of ACM Transactions on Graphics*, 38(4):1–12, 2019.

[68] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[69] Patrick Tinsley, Adam Czajka, and Patrick Flynn. This face does not exist... but it might be yours! identity leakage in generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1320–1328, 2021.

[70] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.

[71] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[72] Xueyu Wang, Jiajun Huang, Siqi Ma, Surya Nepal, and Chang Xu. Deepfake disrupter: The detector of deepfake is my friend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14920–14929, 2022.

[73] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023.

[74] 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.

[75] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7378–7387, 2023.

[76] Jiahui Wu, Yu Zhu, Xiaoben Jiang, Yatong Liu, and Jiajun Lin. Local attention and long-distance interaction of rppg for deepfake detection. *The Visual Computer*, 40(2):1083–1094, 2024.

[77] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 1(3), 2023.

[78] 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.

[79] 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.

[80] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Li Yuan, Chengjie Wang, Shouhong Ding, et al. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*, 2024.

[81] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423, 2023.

[82] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. *arXiv preprint arXiv:2307.01426*, 2023.

[83] Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, and Li Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. *arXiv preprint arXiv:2408.17065*, 2024.

[84] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184, 2023.

[85] Ning Yu, Vladislav Skripniuk, Dingfan Chen, Larry Davis, and Mario Fritz. Responsible disclosure of generative models using scalable fingerprinting. *arXiv preprint arXiv:2012.08726*, 2020.

[86] Yike Yuan, Xinghe Fu, Gaoang Wang, Qiming Li, and Xi Li. Forgery-domain-supervised deepfake detection with non-negative constraint. *IEEE Signal Processing Letters*, 29:2512–2516, 2022.

[87] Cong Zhang, Honggang Qi, Yuezun Li, and Siwei Lyu. Contrastive multi-faceforensics: An end-to-end bi-grained contrastive learning approach for multi-face forgery detection. *arXiv preprint arXiv:2308.01520*, 2023.

[88] 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.

[89] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15023–15033, 2021.

[90] 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.

[91] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the central claims articulated in both the abstract and introduction rigorously align with the paper's technical contributions and methodological scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, the limitations of our methodology are systematically analyzed in Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Although our paper focuses on deepfake detection rather than theoretical analysis, all architectural propositions receive rigorous validation through quantitative experiments and visual demonstrations.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient information in multiple key aspects to support experimental reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code is currently not open source.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, the paper provides comprehensive details regarding the training and testing setup to ensure reproducibility and clarity.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, the paper explicitly specifies the type of computational hardware used (RTX3090). However, resource requirements are minimal, detailed information about execution time or memory consumption is not provided, as these aspects are unlikely to affect reproducibility for experiments.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully read the NeurIPS Code of Ethics, but this paper does not involve any issues that violate ethical standards.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, the impacts of our methodology are systematically analyzed in Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Appendix

This supplementary material provides:

- Sec. B: We discussed the classification and basic process of forgery techniques.
- Sec. C: We reviewed the application of data augmentation methods in forgery detection and discussed more details about FIA-USA.
- Sec. D: We introduced the construction process of FPN.
- Sec. E: We provided a detailed introduction to AFFS was provided, along with visual evidence.
- Sec. F: We provide a visualization example of which boundary pixels were discarded by RCR.
- Sec. G: We provided test results on more forgery techniques.
- Sec. H: We provide experimental results combining FIA-USA with other state-of-the-art detection methods.
- Sec. I: We discussed the limitations of this paper.
- Sec. J: We discussed the impacts of this paper.
- Sec. K: We provided more visual examples.

B Face forgery techniques

In this section, we discuss the classification of facial forgery technology and the basic process of forgery technology, and at which nodes FIA and USA will be introduced.

B.1 Classfication of facial forgery technology

According to different processing procedures and technical principles, we first divide forgery techniques into two categories:
1). *Full Face Synthesis:* Generate the entire image/video directly through the image / video generation model, without involving facial cropping and stitching in the basic process. Therefore, the artifacts involved in such forgery techniques mainly include USA and a small amount of FIA, where USA is the dependency between adjacent pixels in the forgery area due to the performance of the generator, and FIA is the inconsistency in the generator's ability to generate face areas with dense details and background areas with sparse details.
2). *Face Swapping*. Crop the original face, generate fake regions through a generator (optional), and then stitch it with the target face. Therefore, this forgery technique can be divided into two situations: a. It includes FIA and USA: FIA is mainly caused by cropping and stitching, as well as generator performance, while USA is mainly determined by generator performance. b. It only includes FIA: FIA is mainly caused by ropping and stitching. Furthermore, based on the forged area, we can divide face swapping into two parts: *Macro-editing* and *Micro-editing*. Macro-editing can be further classified into two subtypes: Editing based on 4 facial keypoints and Editing based on 81 keypoints. Micro-editing allows for precise adjustments to each facial feature.
Figure 5 provides a detailed overview of our classification concept.

B.2 Basic Process

In Figure 6, we depict the basic process of each type of forgery technique. We can see from this that the main difference between full face synthesis and face swapping lies in whether they include face cropping and stitching operations, which leads to the emergence of a large number of FIA, and whether the background is generated by the generator. The main difference between macro-editing and micro-editing in face swapping is whether the forged area is the entire face or can finely fabricate facial features. Further, macro editing based on 81 key points is more detailed, and the edges of the forged area are more similar to the edges of a real face, while 4 key points result in the forged area being a square. This discovery is mainly due to a detailed observation of the widely used forgery technique ROPE [30] and roop [61], which provides two types of face cropping techniques.

B.3 At what stage will FIA and USA be introduced

In Figure 7, we demonstrate which operations lead to the emergence of FIA and USA in the basic process. The observation of a phenomenon led to the refinement of our FIA-USA design.

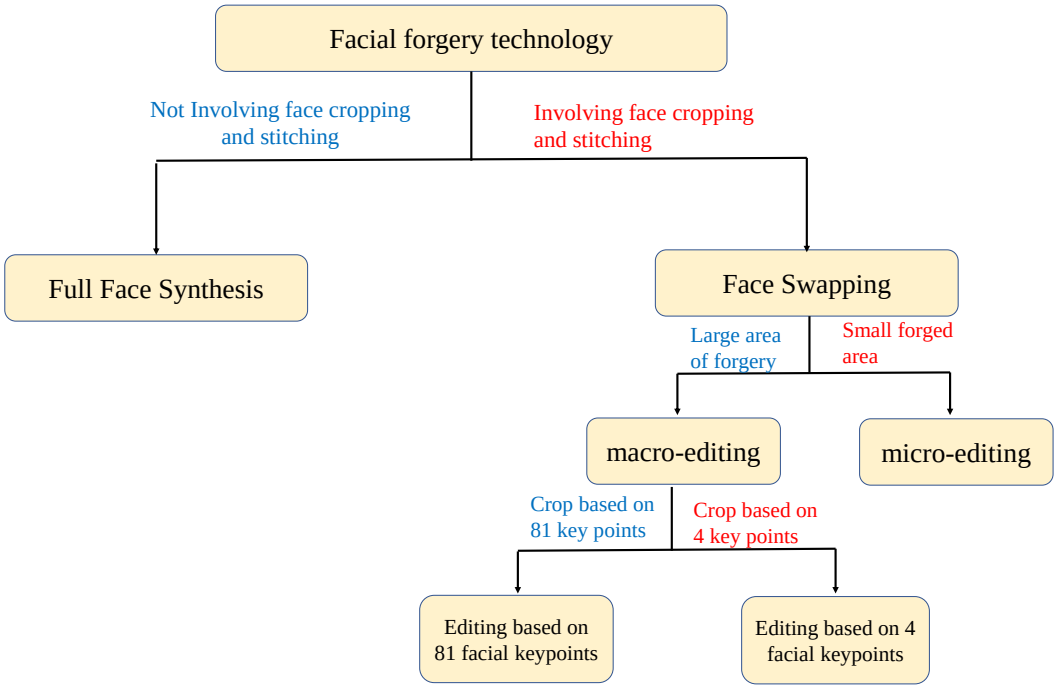


Figure 5: **We categorize forgery techniques into two main types: *Full face synthesis* and *Face swapping*.** Face swapping is further divided into *macro-editing* (based on 4 or 81 keypoints) and *micro-editing*.

C Details of FIA-USA

C.1 Blend-based data augmentation

In Deepfake detection, a successful approach is to manually construct negative samples for training, where data augmentation in the RGB domain is based on Blend, which can be represented as follows:

$$I_F = M \odot I_t + (1 - M) \odot I_s, \tag{8}$$

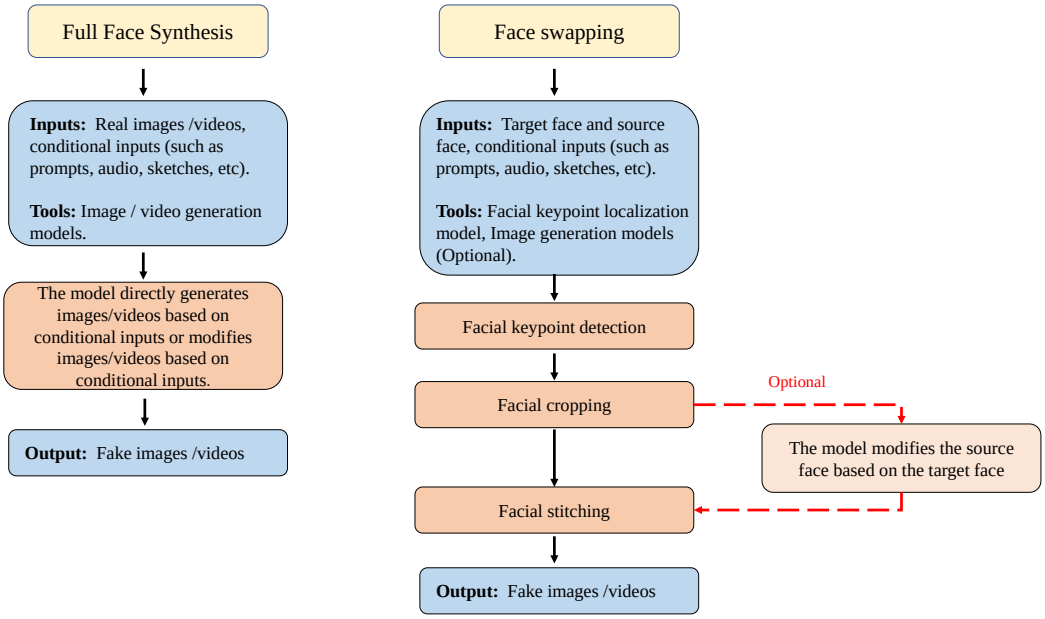
Among them, I_t represents the target face, I_s represents the source face, and I_F represents the negative sample. The earliest design [40] was to use facial similarity to find the most suitable I_s for I_t. Later, SBI [63] proposed self-blend to improve the statistical consistency of I_F. There are also works [64] that separate the background and face of I_t, add noise to the background, and reconstruct I_t, using the reconstructed I_t for blend. **However, to our knowledge, there has been no work so far that considers the introduction of both global and local artifacts, let alone the introduction of FIA at both the full face and regional levels while introducing USA.**

C.2 Other types of data augmentation

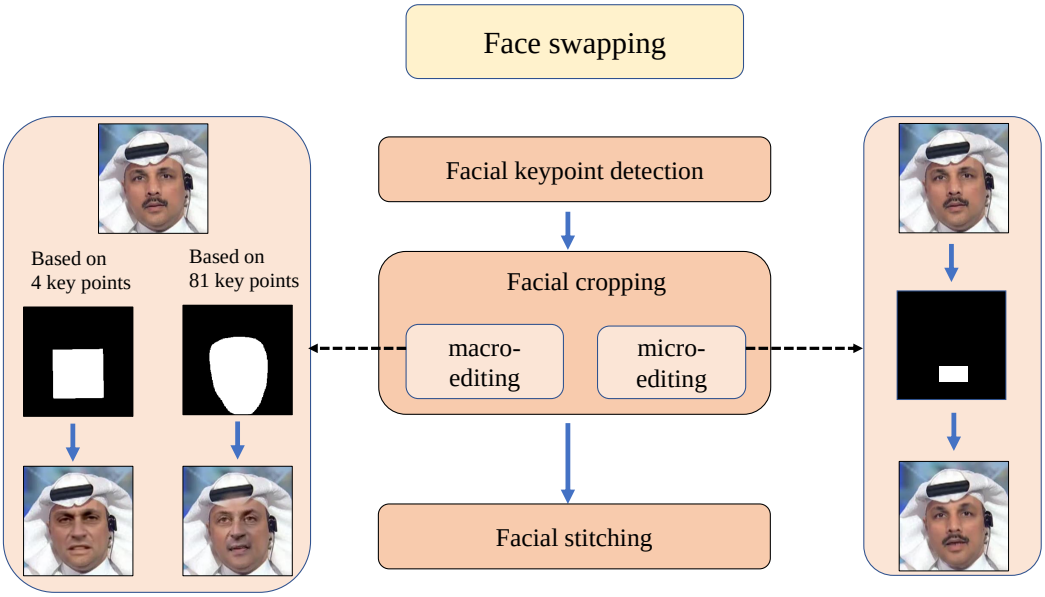
In recent years, many other data augmentation methods have been proposed, such as frequency domain [39] and latent space [79]. Many studies suggest [39, 79] that compared to data augmentation in other domains, data augmentation in the RGB domain [40, 63, 64] exhibits weak robustness and limited effectiveness. However, we have demonstrated through extensive experiments that our proposed data augmentation in the RGB domain is not weaker or even better than that in other domains. **It is worth mentioning that our proposed data augmentation belongs to the image level, and we surprisingly found that it is superior to the most advanced data augmentation methods at the video level [74].**

C.3 The detailed process of generating masks in FIA-USA

Here we have detailed the processing of generating masks in FIA-USA, as shown in Figure 8. Firstly, we divide mask generation into two types: 1) Macro editing masks and 2) Micro editing masks.
Macro editing mask: We follow the mask generation process in SBI [63], which includes the following steps: a. Calculate the convex hull of 81 facial keypoints to obtain the mask of facial contour b. Erosion or expansion of



(a) Basic processes of *full face synthesis* and *face swapping*.



(b) Basic processes of *macro-editing (4 vs 81 keypoints)* and *micro-editing*

Figure 6: We illustrates the basic processes of forgery techniques, *full face synthesis* and *face swapping* differ in face cropping and stitching, while *macro-editing (4 vs 81 keypoints)* and *micro-editing* vary in forgery scope and detail.

the mask; Besides, we also considered the case of calculating convex hull based on 4 facial keypoints, which originated from the observation of two methods for extracting raw faces provided by the open-source deepfake project that is truly available [61, 30].

Micro-editing masks: The main method is to simulate the forgery process of manipulating facial details randomly by randomly combining facial features.

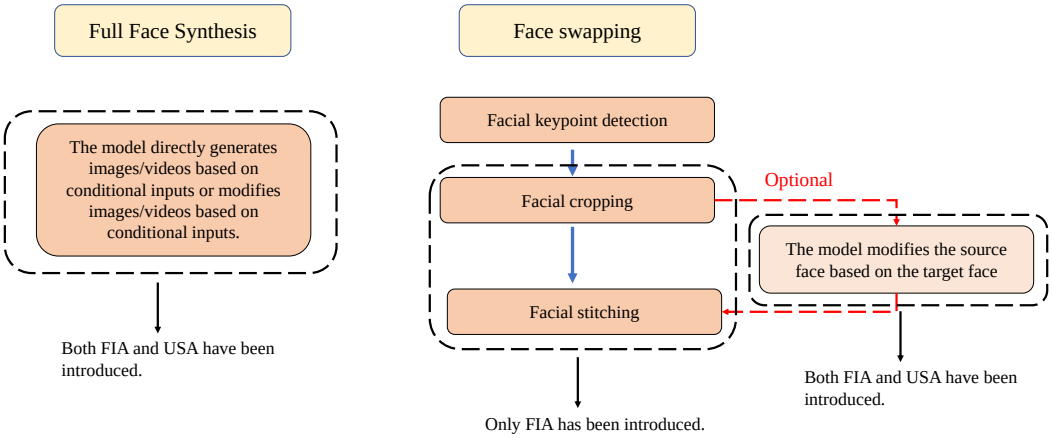


Figure 7: The operations that led to the occurrence of FIA and USA in the basic process.

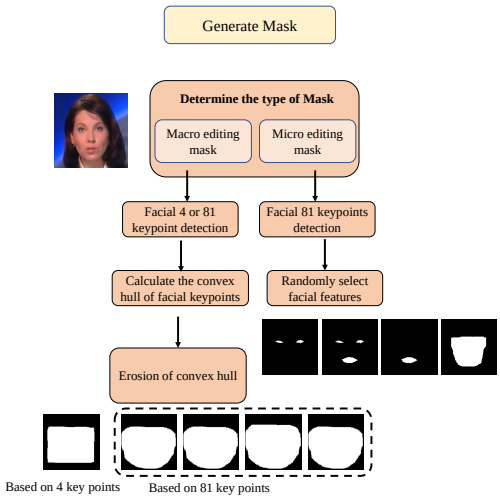


Figure 8: The generation of macro and micro-editing masks in FIA-USA, using convex hull and erosion / expansion for macro masks and random feature combination for micro masks.

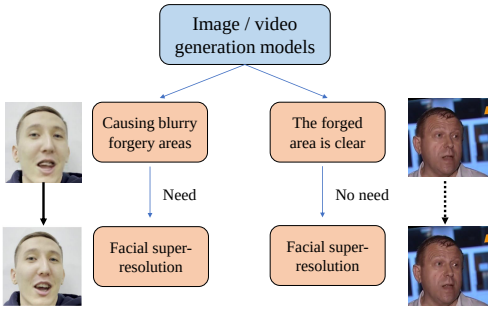


Figure 9: We divide the generation models into two categories based on whether facial super-resolution models is needed for post-processing.

C.4 Why use Autoencoder and Super-Resolution models to reconstruct images

As shown in Figure 9, we divide the processing flow of the generative model into two categories: One category generates images of forged regions that are as clear and free of blur as the original image (using Diffusion models and GANs as the main methods, and upsampling the latent code through AE to generate images). And another

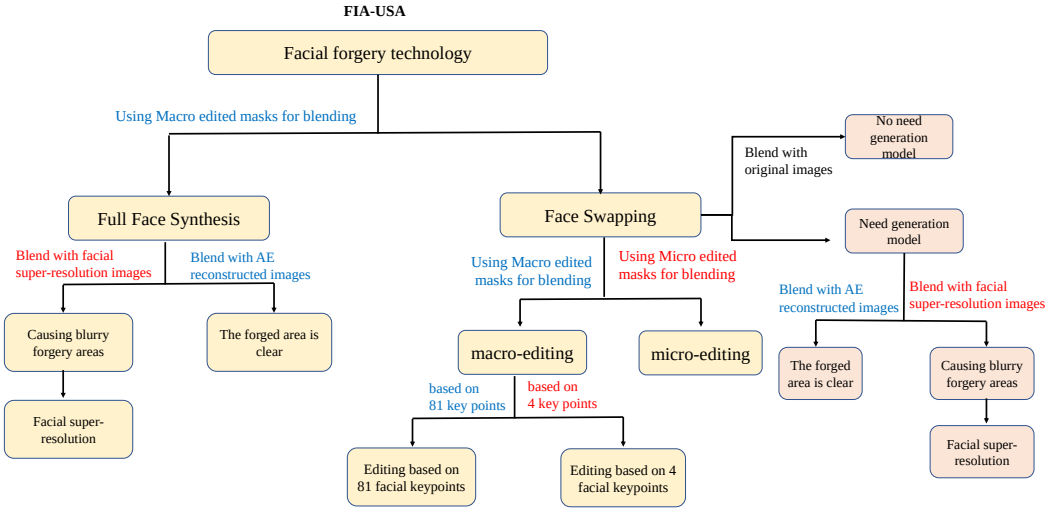


Figure 10: We demonstrate through graphical representation that our proposed FIA-USA covers all possible scenarios that may lead to the occurrence of FIA and USA.

type (especially older voice driven lip shape modification techniques) will result in blurred tampered areas, requiring additional facial super-resolution models to process the blurred areas. Therefore, we use utoencoder (AE) to reconstruct images or facial Super-Resolution (SR) to process faces, in order to approximate the USA introduced by the generator in the forged area.

C.5 FIA-USA covers all possible situations that may occur in FIA and USA

As shown in the Figure 10, our proposed FIA-USA framework covers every possible situation that may occur in FIA and USA, which is also the most intuitive explanation why our method is effective.

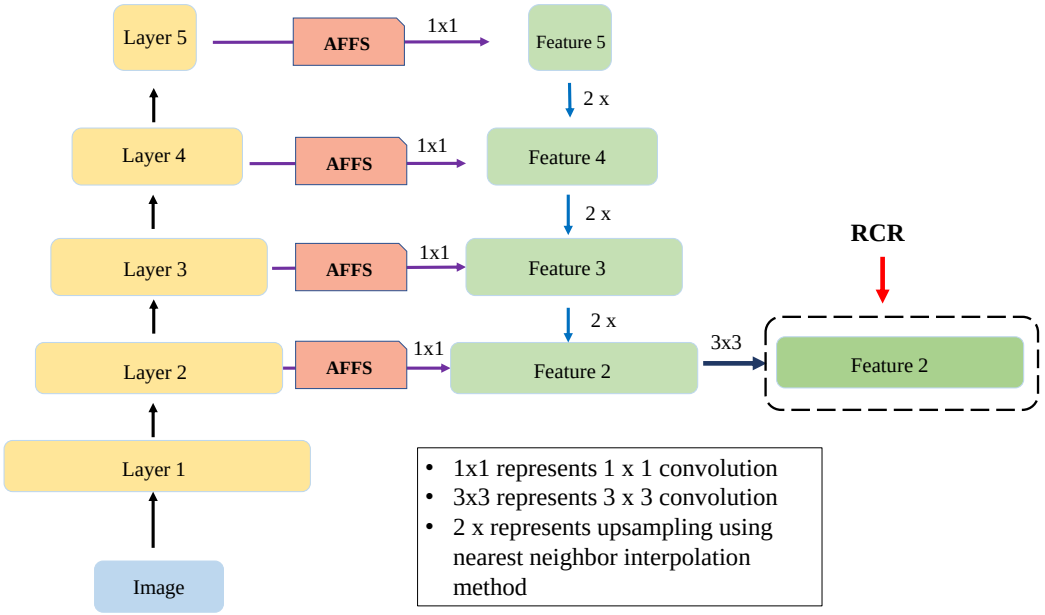


Figure 11: We illustrated the construction process of FPN.

D Details of FPN

As shown in the Figure 11, our FPN follows the framework proposed by [45], using the output features of the 2nd, 3rd, 4th, and 5th layers of the backbone to construct an FPN, In our experimental setup, the number of feature map channels is 128 and 196, corresponding to EfficientNet [66], and Resnet [29], respectively.

E Details of AFFS

Here we provide a detailed description of the process of AFFS and a visual explanation of its effectiveness.

E.1 Why is AFFS proposed

Let's consider a neural network ϕ and an FPN P, as well as the process of performing RCR on the FPN's maximum resolution layer. The output of ϕ is a set of features $\{d_1, d_2, d_3, d_4\}$, P takes this set of features as input to construct a set of feature maps with same dimensions and different resolutions, while RCR performs feature differentiation learning on the feature map with the highest resolution. One intuition is that there is a significant amount of redundancy in the output features of the ϕ, and not every dimension of the features $d_i, i \in [1, 2, 3, 4]$ is suitable for constructing FPN for RCR. Therefore, we propose AFFS for feature dimensionality reduction of each feature d_i before constructing FPN. Moreover, this feature dimensionality reduction method effectively exploits the potential advantages of our proposed FIA-USA and lays the foundation for the effective execution of RCR.

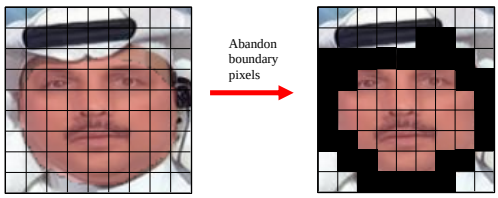


Figure 12: We provide a visual example to illustrate which pixels will be discarded.

E.2 AFFS

1) First, let's consider multiple positive and negative sample pairs generated by FIA-USA, along with their corresponding masks $\{R_n, F_n, M_n\}_{n=1}^N$. The usual method is to directly construct an FPN using a set of features f_i from a neural network ϕ as input, where i represents the i-th layer of the ϕ.
2) Next, let's discuss the initialization process of AFFS. For the output of each layer of the neural network, **we hope that the difference between the real image and the fake image after normalization is as consistent as possible with the mask M used to represent forged areas.** However, we have noticed that not every channel meets this requirement, so we would like to reserve channels that meet the requirements for each layer. Therefore, for $f_i \in \mathbb{R}^{W_i * H_i * C_i}$, we calculate the following loss for each channel on a set of FIA-USA generated samples $\{R_n, F_n, M_n\}_{n=1}^N$ to obtain those excellent dimensions,

$$\mathcal{L}_{AFFS}^{i,k} = \frac{1}{N} \sum_{n=1}^{N} \|\mathcal{F}(f_i^k(R_n) - (f_i^k(F_n)) - M_n\|_2^2 \tag{9}$$

$\mathcal{F}$ represents normalization and scaling operations, $f_i^k \in \mathbb{R}^{W_i * H_i}, k \in [0, C_i]$, represents the feature map of the $k-$th dimension of the f_i . During the training process, we only select the channels reserved for f_i during the AFFS initialization process. After AFFS processing, the original features $f_i \in \mathbb{R}^{W_i * H_i * C_i}$ will become $f_i' \in \mathbb{R}^{W_i * H_i * m_i}$.

E.3 Visual Explanation of AFFS

Here we provide a set of visual examples of AFFS in Figure 13. The experiment was conducted on the 2-th layer of EfficientNetB4. Unlike the actual situation, in order to enhance readability, we only calculated L_{AFFS} for a single epoch on a single pair of samples. And based on this, sort the 32 dimensions of the features in the 2-th layer. Here, we present the dimensions ranked in the top 14 and bottom 4. In the actual process, we iterate for 5 epochs on all samples on the training set to initialize AFFS. In Table 7, we present the changes in feature dimensions of different backbone features before and after performing AFFS.

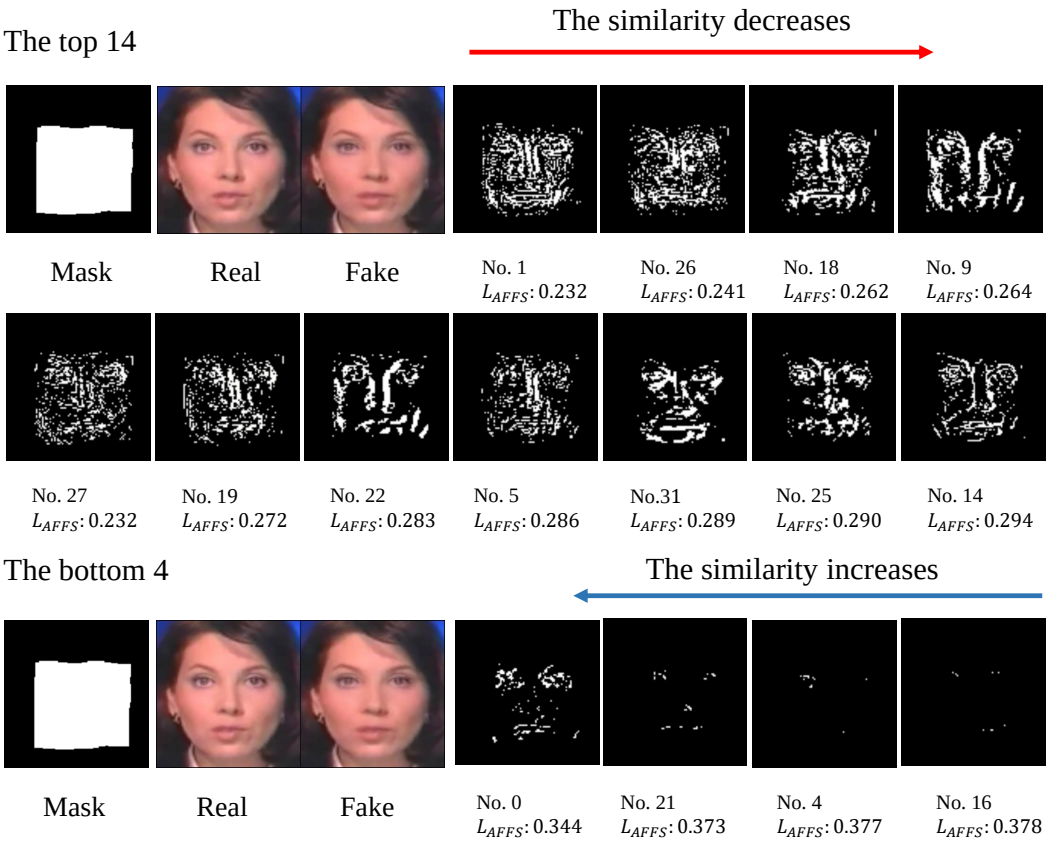


Figure 13: The result of calculating L_{AFFS} for the 32 dimensions of the 2-th layer features of EfficientNetB4 on a single sample and sorting them by similarity.

Table 7: Feature dimension parameters of AFFS.

Model	Input Dimension	Output Dimension
Res50	{256,512,1024,2048}	{196,384,768,1536}
Res101	{256,512,1024,2048}	{196,384,768,1536}
Effb1	{24,40,112,320}	{16,24,40,112}
Effb4	{32,56,160,448}	{24,32,56,160}

Table 8: We report the Video-level AUC. The combination of our data augmentation method with other SOTA detection algorithms. † represents the method used in the original paper.

Method	Training set	CDFv2	DFDCP	Avg.
LAA-Net	SBI + EPPN †	0.840	0.741	0.791
LAA-Net	FIA-USA + EFPN	0.875	0.782	0.829
Ours	FIA-USA + FPN (AFFS)	0.901	0.861	<u>0.881</u>
Ours	FIA-USA + FPN (AFFS) +RCR	0.941	0.866	**0.904**

F Details of RCR

In the LAA Net [51], a very interesting hypothesis is the pixels on the fake boundary contain both the features of the fake area and the features of the real area, so in our designed RCR, such pixels will be discarded. Figure 12 provides a visual example.

G Testing on more forgery techniques

In addition to several widely used forgery detection datasets mentioned in the paper, we also conducted experiments on other forgery techniques based on the DF40 dataset [80], where real samples were obtained from FF++ and fake samples were constructed using corresponding forgery techniques based on FF++ settings. As shown in the Table 9a, our proposed forgery detection framework has achieved excellent detection results on the vast majority of forgery methods. Meanwhile, Table 9b shows the comparison between our method and other state-of-the-art methods on various forgery techniques in DIFF.

Table 9: Test results on more forgery techniques.

(a) **We tested all forgery techniques provided on the DF40 dataset and reported Frame-Level AUC.** All results with AUC greater than 80 are highlighted in bold, while those with AUC less than 80 but greater than 70 are underlined.

Type	mobileswap	MidJourney	faceswap	styleclip	DiT	lia	ddim	mcnet	RDDM	StyleGAN3
AUC	**0.97**	**0.97**	**0.89**	**0.82**	0.66	**0.91**	**0.97**	**0.84**	<u>0.70</u>	**0.93**
Type	e4e	pixart	whichfaceisreal	deepfacelab	StyleGANXL	heygen	facedancer	MRAA	pirender	VQGAN
AUC	**0.95**	**0.93**	0.42	<u>0.77</u>	0.41	0.58	**0.83**	**0.85**	**0.81**	**0.85**
Type	StyleGAN2	facevid2vid	simswap	inswap	one_shot_free	wav2lip	e4s	starganv2	tpsm	sd2.1
AUC	**0.93**	**0.83**	**0.91**	**0.87**	**0.85**	<u>0.76</u>	**0.88**	0.48	**0.83**	**0.97**
Type	blendface	hyperreenact	uniface	CollabDiff	fsgan	danet	SiT	sadtalker	fomm	
AUC	**0.94**	**0.80**	**0.92**	**0.95**	**0.86**	**0.80**	<u>0.72</u>	<u>0.74</u>	**0.85**	

(b) **Comparison with universal deepfake detection methods using the Frame-Level AUC on the DiFF dataset.** The notation † indicates models that are designed for deepfake detection, while ‡ signifies models intended for general generated detection. All experiments were trained on the c23 version of FF++. The best result is bolded.

Method	Test Subset												
	Cycle	CoDiff	Imagic	DiFace	DCFace	Dream	SDXL_R	FD_I	LoRA	Midj	FD_T	SDXL	HPS
F^3-Net† [56]	36.14	35.00	32.56	25.61	53.29	55.45	65.04	40.90	-	-	45.00	61.64	68.96
EfficientNet† [66]	56.51	38.38	48.50	64.45	89.13	71.64	65.04	59.93	-	-	69.67	64.94	74.63
CNN_Aug‡ [70]	50.31	46.84	75.95	43.10	80.69	58.75	60.10	43.65	-	-	47.90	61.95	60.56
SBI† [63]	82.32	64.70	49.84	73.64	89.92	75.66	77.73	87.67	89.18	79.56	90.55	80.37	83.30
Ours	**84.76**	**72.61**	52.92	**76.86**	**93.63**	**83.43**	**80.54**	**92.68**	**93.48**	**82.57**	**94.85**	**82.15**	**88.40**

H Combining with SoTA detection method

In this section, we aim to validate the effectiveness of our FIA-USA data augmentation by integrating it with state-of-the-art image-level detection methods. While our comprehensive evaluation currently focuses on LAA-Net [51] (CVPR 24) due to limited availability of open-source image-level detection implementations, Table 8 reveals significant performance improvements. **Notably, the original LAA-Net implementation was trained on raw version FF++ data, whereas our experiments utilize the c23 version.** Compared with the baseline implementation, our data augmentation strategy has significantly improved the performance of the original detector. To address requests for detailed comparisons with LAA-Net, our table shows two key findings: First, our method achieves a 7.5% performance advantage over LAA-Net when using identical data augmentation conditions. Second, even without employing RCR (Robust Context Refinement), our approach substantially outperforms LAA-Net's EFPN component, demonstrating the inherent strength of our core methodology.

I Limitations

The proposed detection framework primarily focuses on **image-level data augmentation** and matching detection architectures. While we categorize image-level artifacts into FIA and USA, and have achieved satisfactory video-level detection performance, we acknowledge that video-level artifacts present more complex challenges - including temporal inconsistencies and audio-visual asynchrony - which fall outside the current research scope but represent our intended future research direction.
Second, although recent approaches increasingly incorporate large models' zero-shot capabilities for deepfake

detection, existing work has not yet explored the integration of data augmentation methods with such models due to practical constraints including variations in training methodologies and model scales. This does not mean that our method cannot be applied to large models, but currently we are still unfamiliar with large models and have not fully explored this part.
Regarding experimental validation, **we mainly compared and analyzed our proposed method with previous data augmentation methods, which is consistent with the main contribution of our paper**. Of course, we also compared state-of-the-art detection methods that are not data augmentation based. Following established practices in data augmentation research, we conducted comprehensive evaluations across four widely adopted network architectures to ensure fair comparison. However, we note that detection performance may vary across different network frameworks depending on their baseline performance and inherent compatibility with deepfake detection tasks. We cannot guarantee that all model frameworks will achieve excellent detection results after using our method.
Finally, while our method demonstrates effectiveness across a substantial majority of forgery methods (**validated on over 58 different techniques**), we acknowledge relatively weaker performance in specific edge cases. Nevertheless, the framework maintains robust detection capability for most common forgery approaches, achieving satisfactory overall performance that meets our research objectives.

J Broader impacts

The proposed framework offers positive societal impacts by enhancing detection of sophisticated face deepfakes, thereby mitigating disinformation campaigns, fake news, and fraudulent content that erode public trust in digital media. Additionally, it safeguards individual privacy by identifying forged facial manipulations that enable identity theft, unauthorized face swapping, and other privacy violations. The method could further support content authenticity initiatives through integration into social media platforms or news verification systems to prioritize legitimate visual content. However, potential negative impacts include the risk of adversaries reverse-engineering the detection framework to refine deepfake generation methods, potentially escalating the adversarial "arms race" between forgery creators and detectors. Furthermore, false positives in detection could lead to unintended censorship, where legitimate content is erroneously flagged as fake.

K More visual examples

From Figure 14 to Figure 16, we demonstrate the fake samples that FIA-USA can generate through different reconstruction methods, including autoencoderr (AE) and facial super-resolution (SR), and multiple types of masks based on three real faces. Figure 17 shows the negative samples randomly generated by FIA-USA during the real training process. Figures 18 to 22 list the images in the DF40 dataset [80] that were misclassified by our method.

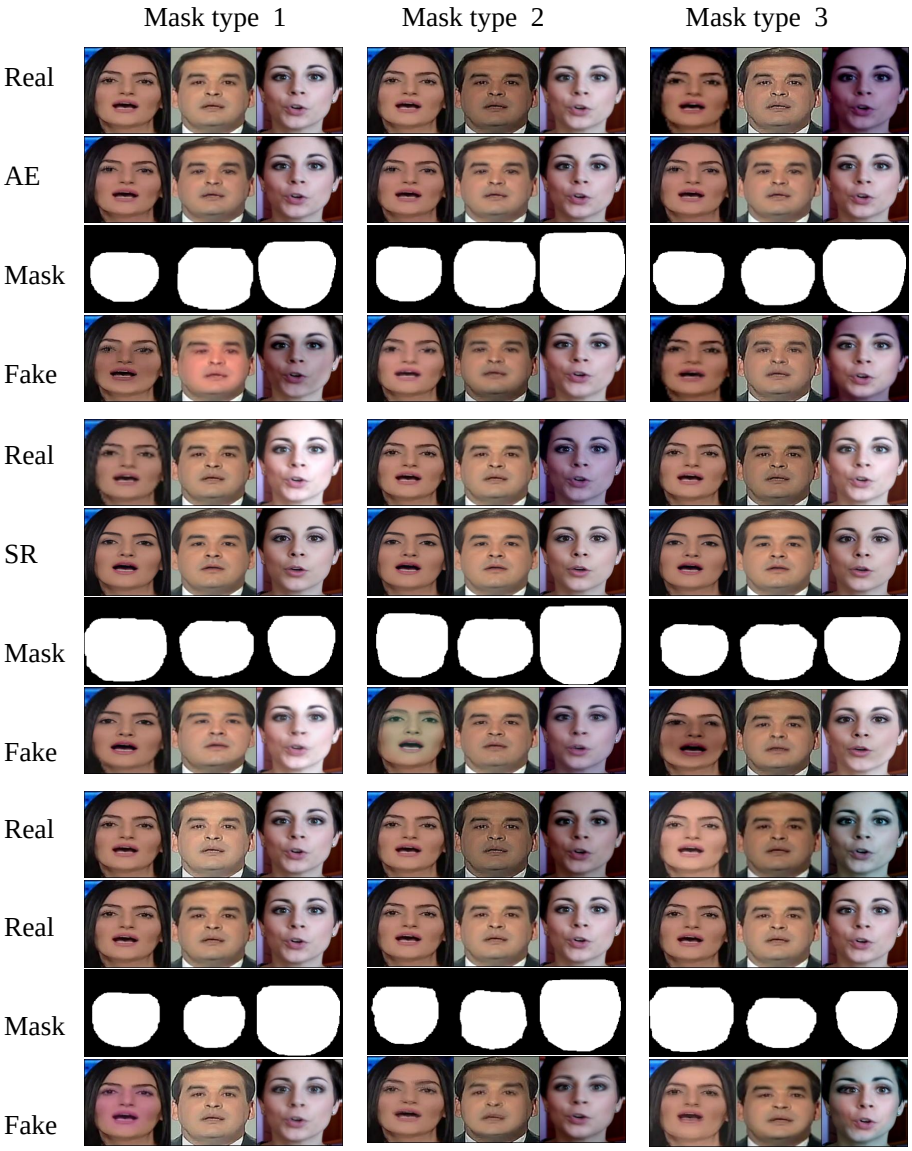


Figure 14: **More examples of FIA-USA.**

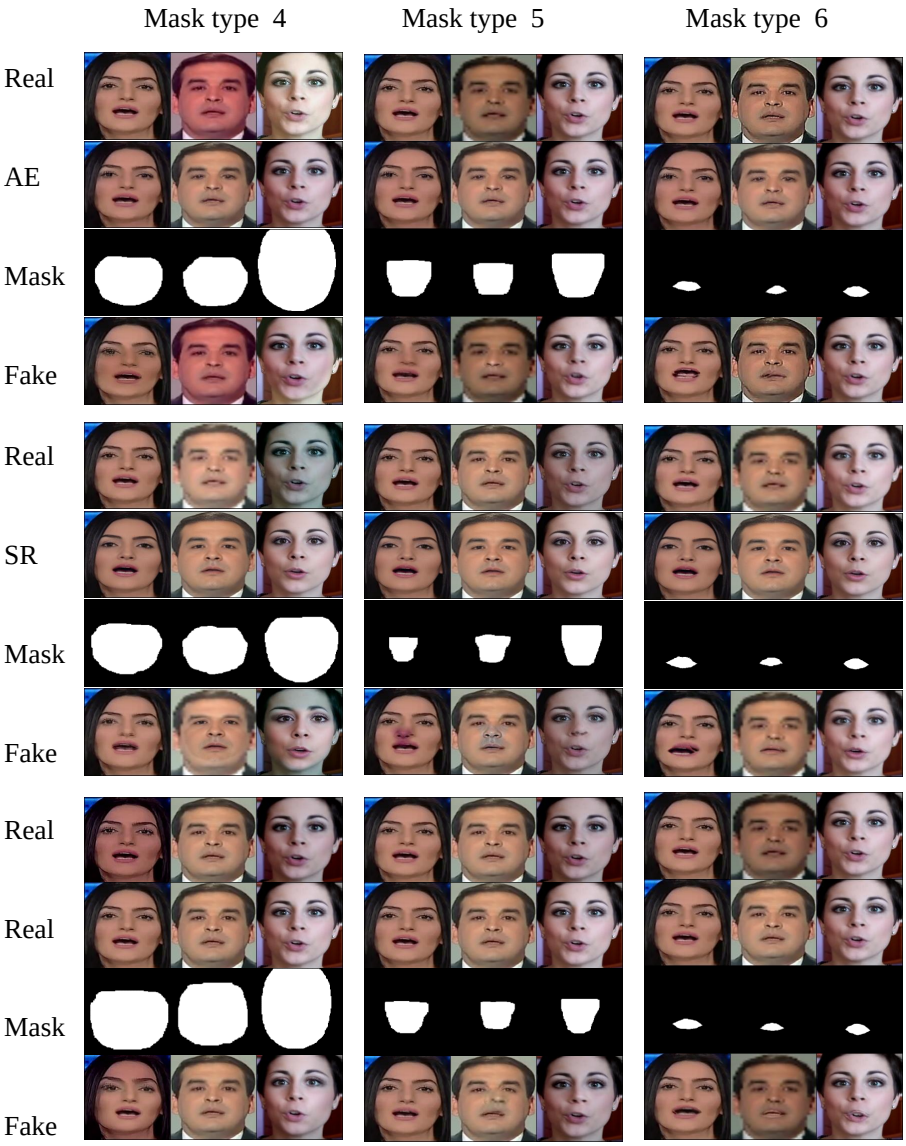


Figure 15: **More examples of FIA-USA.**

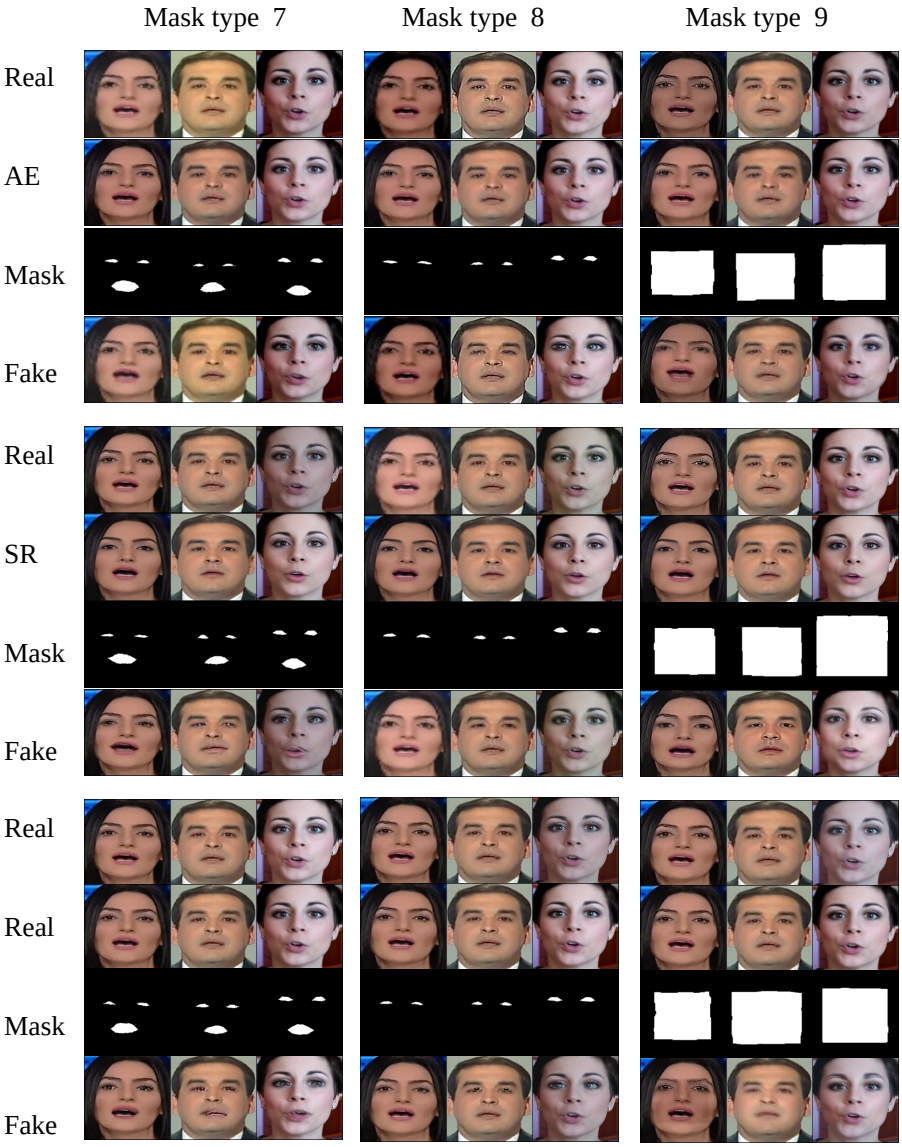


Figure 16: **More examples of FIA-USA.**

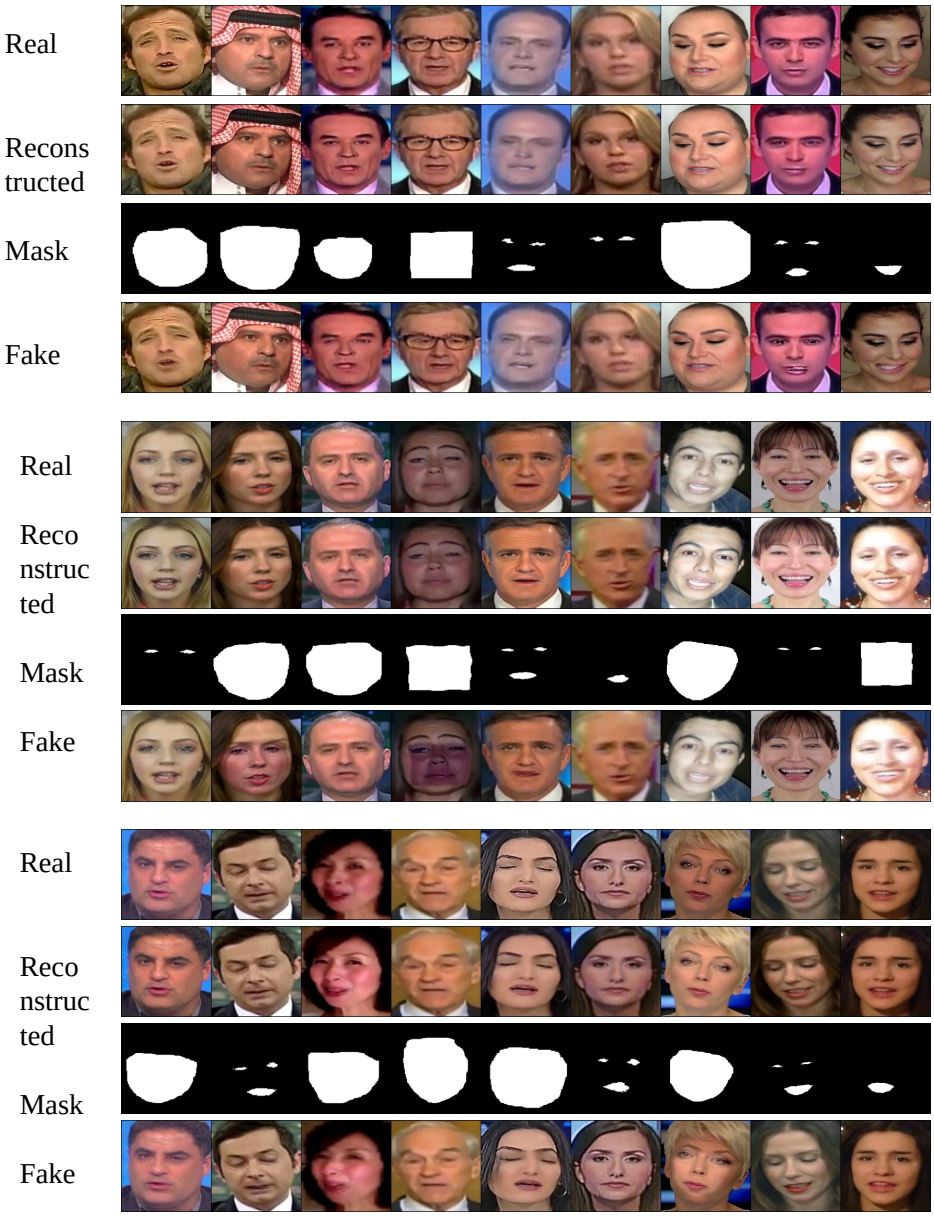


Figure 17: **More examples of FIA-USA.**

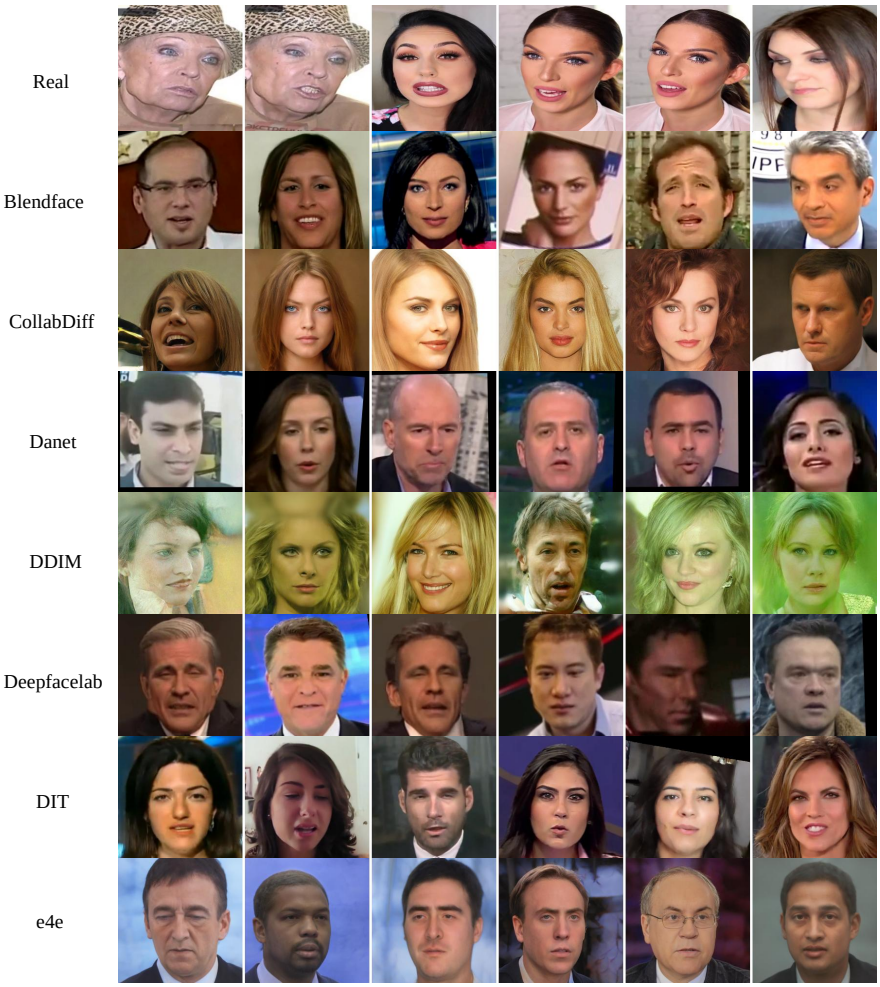


Figure 18: **Images misclassified by our method in the DF40 dataset.**

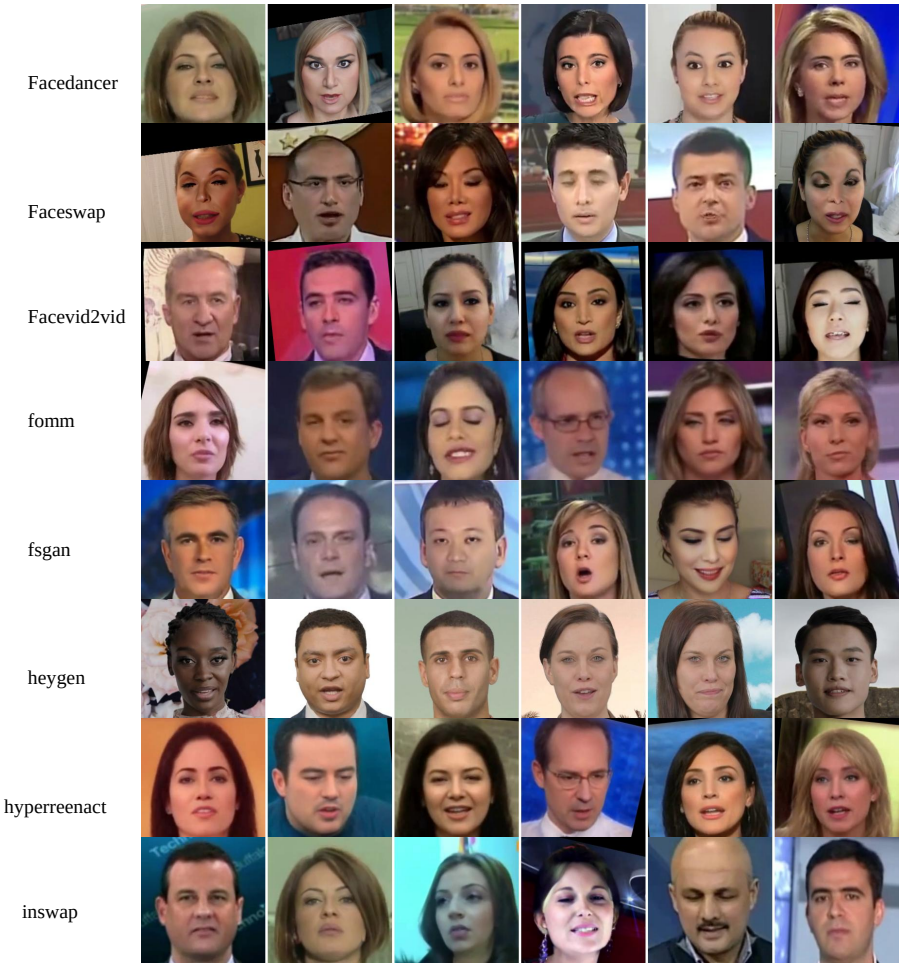


Figure 19: **Images misclassified by our method in the DF40 dataset.**

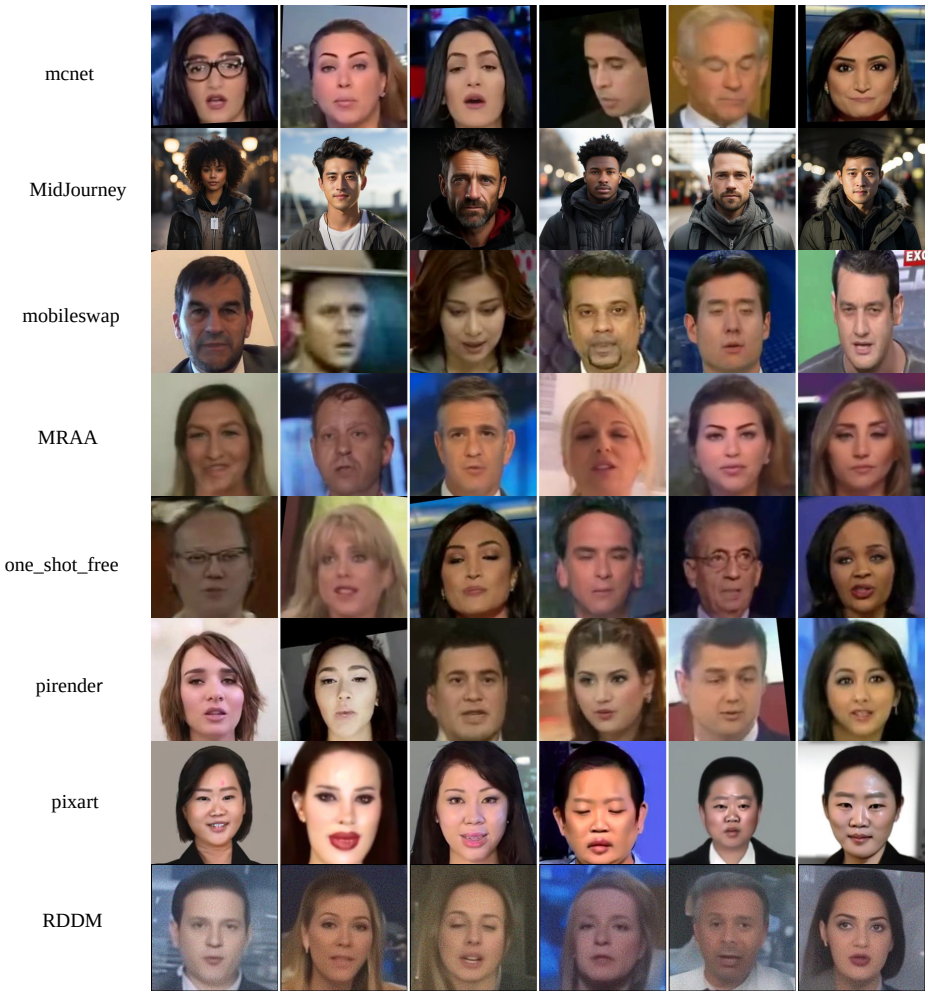


Figure 20: **Images misclassified by our method in the DF40 dataset.**

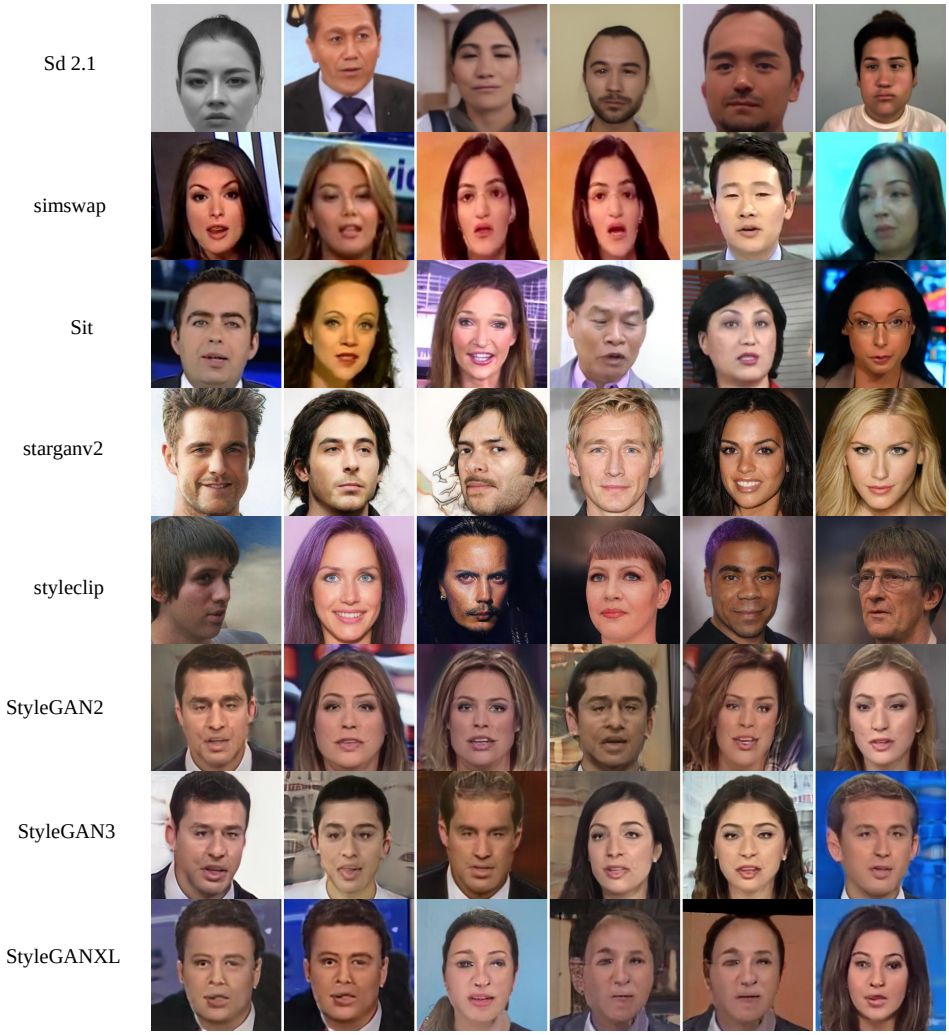


Figure 21: **Images misclassified by our method in the DF40 dataset.**

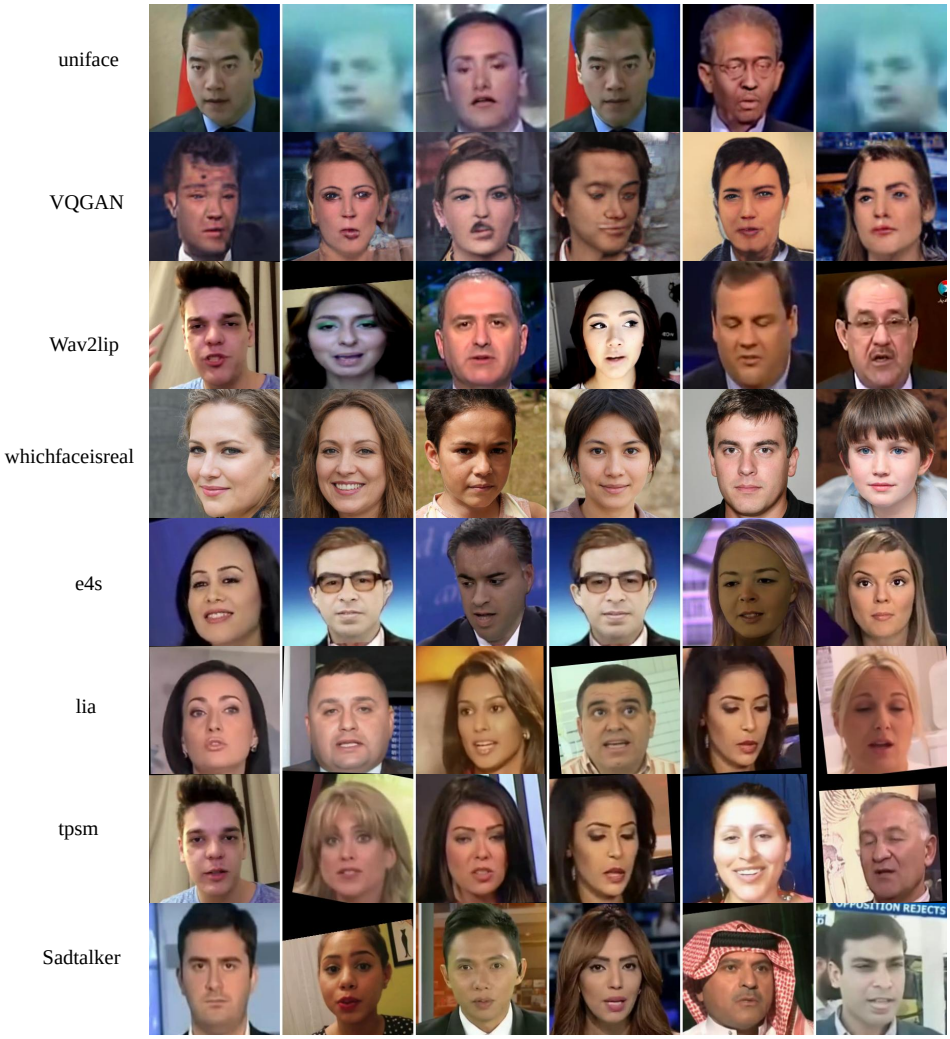


Figure 22: **Images misclassified by our method in the DF40 dataset.**