AUNet: Learning Relations Between Action Units for Face Forgery Detection

Weiming Bai[1,2*] Yufan Liu[1,2*] Zhipeng Zhang[3*] Bing Li[1,4†] Weiming Hu[1,2,5]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]DiDiChuxing [4]People AI, Inc.
[5]CAS Center for Excellence in Brain Science and Intelligence Technology
{baiweiming2019,yufan.liu}@ia.ac.cn zhipeng.zhang.cv@outlook.com {bli,wmhu}@nlpr.ia.ac.cn

Abstract

Face forgery detection becomes increasingly crucial due to the serious security issues caused by face manipulation techniques. Recent studies in deepfake detection have yielded promising results when the training and testing face forgeries are from the same domain. However, the problem remains challenging when one tries to generalize the detector to forgeries created by unseen methods during training. Observing that face manipulation may alter the relation between different facial action units (AU), we propose the Action-Units Relation Learning framework to improve the generality of forgery detection. In specific, it consists of the Action Units Relation Transformer (ART) and the Tampered AU Prediction (TAP). The ART constructs the relation between different AUs with AU-agnostic Branch and AU-specific Branch, which complement each other and work together to exploit forgery clues. In the Tampered AU Prediction, we tamper AU-related regions at the image level and develop challenging pseudo samples at the feature level. The model is then trained to predict the tampered AU regions with the generated location-specific supervision. Experimental results demonstrate that our method can achieve state-of-the-art performance in both the in-dataset and cross-dataset evaluations.

1. Introduction

The success of the generative model, *e.g.,* Generative Adversarial Networks (GAN) [21], rapidly improves the quality of face forgery, which provokes researchers to pursue antithetical counter-detection methods to deal with potential social security issues. Though recent works have demonstrated their effectiveness in identifying forgery images from known forging methods that are used in training [4,10,29,39,43], the generalization on unknown forgery methods is not guaranteed [9, 17, 27, 52].

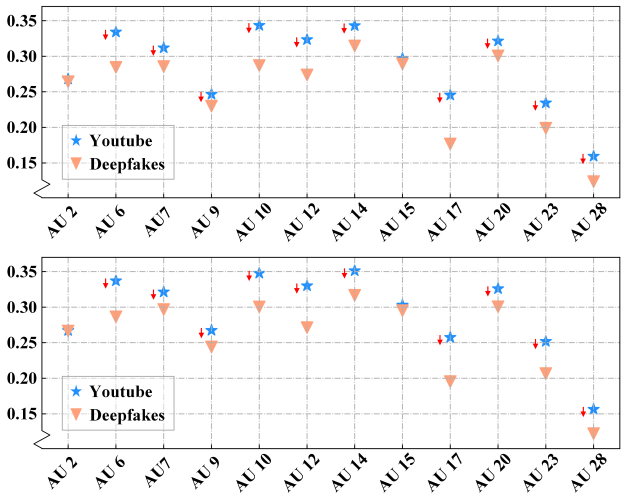

Figure 1. **The average correlation intensity (y-axis) between an AU (x-axis) and other AUs under different data volume (top-20% and bottom-80% samples from FF++ [42], respectively).** More details and results can be found in the supplementary. It can be observed that the average correlation intensity between an AU and other AUs is weakening after manipulation.

Some recent works [6, 31, 45, 49, 50, 55] have noticed this imminent problem and attempted to capture more intrinsic forgery clues to improve the generalization of identification methods. In particular, these works can be roughly categorized into two branches, 1) **data modification** which applies carefully selected augmentations [50] or manually generates forgery images [31, 45, 55] with only real images to enlarge training data diversity meanwhile avoiding overfitting to specific defects, 2) **auxiliary task integrating** which defines an affinitive loss to help the model learn underlying differences between real and fake faces [6, 55]. Despite their success, it is noticed that the relations between face units that are general in biology research for understanding human facial characteristics are less explored, in-

*Equal contribution
†Corresponding Author

hibiting the further improvement of model generalization.

In this work, we aim to construct a face forgery detection framework that unifies the data modification and auxiliary task integrating schemes with relation clues of facial action units. Our insight is motivated by the theory in the Facial Action Coding System [18], which represents human faces through a set of facial muscle movements called facial Action Units (AU). More specifically, in facial morphology, a muscle controls different AUs. When showing a certain emotion, a group of AUs would be activated simultaneously, indicating that there are underlying relations between AUs. Therefore, a direct intuition is that relations among AUs may be different in real and manipulated faces, since de facto methods modify the entire face using graphical methods, or generate the individual facial regions by GAN. This hypothesis is evidenced by our experiments in Fig. 1, which shows that an AU in real face presents stronger average correlation intensity to other AUs (the stars), whereas that in forgery face are weaker (the triangles).

To explore these clues, we focus on relations between local regions associated with AUs and propose the Action Units Relation Learning framework to improve the robustness and generalization of the forgery detector. The proposed framework is comprised of **A**U **R**elation **T**ransformer (**ART**) and **T**ampered **A**U **P**rediction (**TAP**). ART, explicitly learning the relations between AUs, works as a dual network. In specific, it consists of an AU-specific Branch for learning relations among AU-aligned regions and an AU-agnostic Branch for learning relations among image-patch regions. The AU-specific Branch extracts embeddings aligned with individual AU and builds their relations by attention mechanism. The AU-agnostic Branch is a standard Vision-Transformer block [16] that is designed to construct relations between different image patches. These two branches complement each other for building a detailed and global view of the input face image. From another perspective, TAP formulates an auxiliary task to enhance the ability of the model to sense local forgery defects. In particular, it constructs a Partial Face Mask by randomly removing AU-related regions from the facial area. The mask is utilized to modify data in the remaining regions at both image and feature levels to generate challenging fake counterparts. The model is then trained to predict the manipulated regions with the help of Local Tampering Supervision. By doing so, the networks are more sensitive to AU-related regions that have been manipulated, which is beneficial for identifying forgery images. Notably, the proposed Action Units Relation Learning framework absorbs the advantages of both data modification and auxiliary task integrating schemes.

We evaluate our framework following cross-dataset protocol and cross-manipulation protocol. In the cross-dataset evaluation, our approach performs favorably against other state-of-the-art detectors and achieves the AUC scores of 92.77%, 99.22%, 73.82%, 86.16%, 81.45% on CDF [36], DFD [1], DFDC [13], DFDCP [14], and FFIW [57] datasets, respectively. In the cross-manipulation evaluation, our approach achieves the AUC of 99.98%, 99.60%, 99.89%, and 98.38% on DF [2], F2F [48], FS [3] and NT [47], respectively. Experimental results demonstrate the effectiveness and generalization of our framework.

Our contributions can be summarized as follows:

- We propose the Action Units Relation Transformer (ART) to effectively build correlations among different AU-related regions and eventually improve the performance of the forgery detection.
- We propose the Tampered AU Prediction (TAP) as an auxiliary task to strengthen the model capability of sensing local forgery regions.
- Experimental results on in-dataset and cross-dataset evaluation protocols demonstrate the effectiveness and generalization of our framework.

2. Related Work

2.1. Conventional Deepfake Detection

Face forgery detection is a classical topic in computer vision. Earlier studies concentrate on hand-crafted features, *e.g.,* eye blinking [26, 34], inconsistent head poses [53] and visual artifacts [5, 35]. With the rapid progress of deep learning, recent methods based on deep neural networks achieved better performance. [10, 54] apply the attention mechanism to highlight the manipulated regions. In addition to focusing on the spatial domain, [20] notices the artifacts information hidden in the frequency domain. Subsequently, many works [30, 37, 38, 40] leverage frequency clues as the supplement to RGB information. Recently, due to the remarkable representation capability of Transformer [16], FTCN-TT [56], ICT [15] have also extended transformer to deepfake detection tasks. Although the aforementioned methods achieve promising results in the intra-domain, the performance of them suffer considerable drops in the cross-dataset scenario.

2.2. General Deepfake Detection

Recent works focusing on general face forgery detection have been proposed. FWA [35] focuses on a quality difference between GAN-generated faces and natural faces, and reproduces it by blurring facial regions on real images. [49] argues that the lack of generalizability is a result of overfitting to visual artifacts, and proposes a dynamic data argumentation scheme. [50] found that, with careful pre- and post-processing and data augmentation, a standard image classifier is able to generalize surprisingly well to unseen datasets. Face X-ray [31] generates training data using only real images and focuses on predicting the blending boundaries in fake faces. PCL [55] also creates blended faces from pairs of two pristine images and performs pair-wise self-consistency learning on generated data. SBI [45] synthesizes fake images by blending pseudo source and target images from a single image. Despite the promising improvements, it is observed that these methods mainly focus on

data manipulation, lacking a module to explore the general and intrinsic facial representation. To this end, we construct a framework that absorbs advantages of data modification meanwhile learns AU relations to model facial information.

3. Methods

3.1. Overview

As depicted in Fig. 2, our proposed Action Units Relation Learning framework consists of Action Units Relation Transformer (ART) and Tampered AU Prediction (TAP). The ART explores clues about relations between AU-related regions to boost forgery identification. The TAP tampers AU-related regions and provides Local Tampering Supervision to improve the generalization ability.

3.2. Action Units Relation Transformer

In the Action Units Relation Transformer (ART), the input images are processed through backbone to extract features. Three ART encoders are stacked to fully exploit relation between AU-related regions. Each encoder consists of an AU-specific Branch and an AU-agnostic Branch. The AU-specific Branch extracts features aligned with specific AU and builds their relations by attention mechanism. The AU-agnostic Branch is designed to construct relations between image patches, where a patch contains both AU-related regions and other potentially useful face clues. The details of the ART encoder are shown in Fig. 3.

AU-specific Branch. In this branch, the attention maps related to AUs are first generated. Let $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ denote the output feature maps of the backbone. An attention module is utilized to estimate AU regions $\mathbf{M}_{AU} \in \mathbb{R}^{K\times H\times W}$, which corresponds to K AU-related regions. The attention module consists of two 3×3 convolutional layers, a 1×1 convolution layer with K filters, and a sigmoid activation function. To target relevant AU regions more precisely, we perform supervision on the predicted attention maps as follows:

$$\mathcal{L}_{Att} = \mathcal{L}_{\delta}(\mathbf{M}_{AU} - \mathbf{Y}_{Att}), \tag{1}$$

where $\mathcal{L}_{Att}$ is a Huber loss function with the parameter $\delta = 0.5$, and $\mathbf{Y}_{Att} \in \mathbb{R}^{K\times H\times W}$ is the ground-truth attention map for K AU-related regions, which is estimated during pre-processing.

We then extract the AU embeddings $\mathbf{E}^{AU} \in \mathbb{R}^{K\times C}$, which are aligned with individual AU as follows:

$$\mathbf{E}_i^{AU} = \mathrm{GAP}(\mathrm{Conv}(\mathbf{F} \odot \mathbf{M}_i + \mathbf{F})),\ i = 1, \cdots, K, \tag{2}$$

where Conv represents a 3×3 convolutional layer, GAP is a global average pooling layer, and $\odot$ is the Hadamard product. After acquiring AU embeddings, we obtain the correlation matrix $\mathbf{A} \in \mathbb{R}^{K\times K}$ by calculating the correlation between the different AU embeddings as follows:

$$a_{i,j} = f(\mathbf{E}_i^{AU}, \mathbf{E}_j^{AU}),\ i, j = 1, \cdots, K, \tag{3}$$

where f is the dot-product similarity function. We then obtain the importance factor $\mathbf{W} \in \mathbb{R}^{K\times 1\times 1}$ based on the association of one embedding with other embeddings:

$$W_i = \sum_j a_{ij} \Big/ \sum_i \sum_j a_{ij} \tag{4}$$

With the learned factor $\mathbf{W}$, we aggregate $\mathbf{M}_{AU}$ and strengthen the important AU-related regions in the original feature map. Concretely, the AU regions are reweighted and aggregated to generate the important AU region $\mathbf{M}_I \in \mathbb{R}^{1\times H\times W}$ by an element-wise maximum operation:

$$\mathbf{M}_I = MAX\{\mathbf{W} \odot \mathbf{M}_{AU}\}. \tag{5}$$

Finally, we multiply $\mathbf{M}_I$ with the original feature map $\mathbf{F}$ to gain enhanced facial action unit features $\mathbf{F}_S \in \mathbb{R}^{C\times H\times W}$ using an element-wise production.

AU-agnostic Branch. In this branch, the Transformer encoder [16] is utilized to capture rich relations among local image patches. A projection module is developed to transform 2D sequence input to 1D. Concretely, a 1×1 convolution layer is first applied to project the input feature maps $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ to $\mathbf{F}_p^i \in \mathbb{R}^{C'\times H\times W}$. Note that every $c' \times 1$ vector can be considered as a representation of a corresponding patch in the input image. We then slice the feature maps $\mathbf{F}_p^i$ along the channel dimension and realign them as a sequence of feature vectors $\mathbf{x}_p^i \in \mathbb{R}^{C'\times(H\cdot W)}$. A learnable class token $\mathbf{x}_c^i \in \mathbb{R}^{C'\times 1}$ is also appended to get the input sequences $\mathbf{x}^i \in \mathbb{R}^{C'\times(1+H\cdot W)}$.

The output sequences $\mathbf{x}^o \in \mathbb{R}^{C'\times(1+H\cdot W)}$ of the standard Transformer encoder are obtained by:

$$\mathbf{x}' = \mathrm{LN}(\mathbf{x}^i + \mathrm{MSA}(\mathbf{x}^i)),\ \mathbf{x}^o = \mathrm{LN}(\mathbf{x}' + \mathrm{FFN}(\mathbf{x}')), \tag{6}$$

where LN, MSA, and FFN are the layer normalization, multi-head self-attention, and feed-forward network, respectively. We then split output sequences $\mathbf{x}^o$ back into a class token $\mathbf{x}_c^o$ and patch tokens $\mathbf{x}_p^o$, accordingly. The patch tokens $\mathbf{x}_p^o$ are reshaped back to a 2D feature map $\mathbf{F}_p^o \in \mathbb{R}^{C'\times H\times W}$, and are then projected to the output $\mathbf{F}_A \in \mathbb{R}^{C\times H\times W}$ with a 1×1 convolution layer.

Feature Integration. After the AU-specific branch and AU-agnostic branch, we acquire features $\mathbf{F}_S$ and $\mathbf{F}_A$, which present rich relation information. On the one hand, we extract the AU-aligned embeddings and construct rich correlations between individual AUs. On the other hand, the transformer encoder extracts global relations between groups of AUs from AU-agnostic image patches. To combine these features, we integrate $\mathbf{F}_S$ and $\mathbf{F}_A$ by the element-wise addition and two 3×3 convolution layers:

$$\mathbf{F}_{out} = \mathrm{Conv}(\mathbf{F}_S + \mathbf{F}_A), \tag{7}$$

where $\mathbf{F}_{out}$ is the output of the ART encoder. After integration, these features containing relation information from the

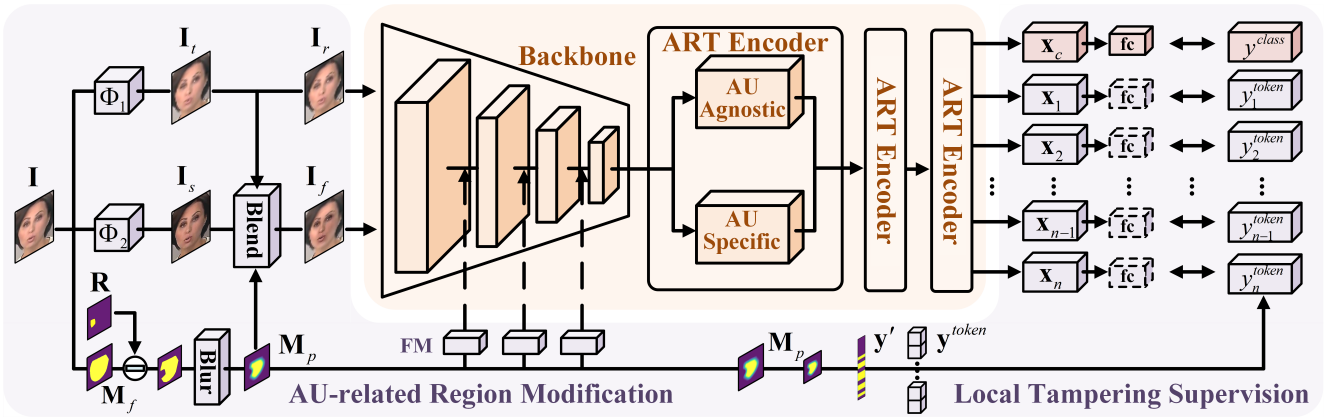

Figure 2. **The overview of the Action Units Relation Learning framework.** Our proposed framework consists of Action Units Relation Transformer (ART) and Tampered AU Prediction (TAP), shown in the orange and purple boxes, respectively

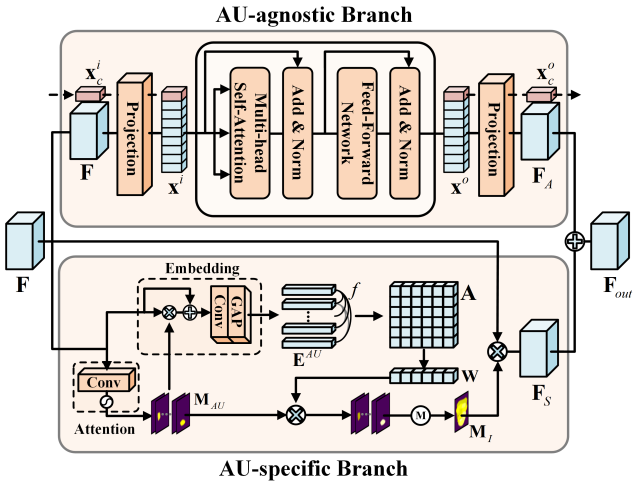

Figure 3. **The details of our proposed ART Encoder.**

detailed and global perspectives are complementary with each other to learn valuable forgery cues.

3.3. Tampered AU Prediction

In addition to building an effective model, we further formulate an auxiliary task, *i.e.,* Tampered AU Prediction (TAP), to enhance the ability of the model to sense local forgery defects. As illustrated in the purple boxes in Fig. 2, this task includes AU-related Region Modification (ARM) and Local Tampering Supervision (LS). In the ARM, given an authentic face, the Image-level Tampering is first performed on the AU-related regions to obtain the authentic face and its fake counterparts. The Feature-level Mixing then develops more challenging fake features by mixing the statistical features of real features into those of fake features. With the authentic and tampered faces, the model is trained to predict the location-specific manipulated regions under the Local Tampering Supervision (LS).

AU-related Region Modification. As shown in Fig. 2, we first generate a Partial Face Mask $\mathbf{M}_p$ for the Image-level Tampering and the Feature-level Mixing. Given a real image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, a landmark detector is first applied to $\mathbf{I}$ for predicting 68 facial landmarks. The convex hull of these facial landmarks is calculated to obtain the facial mask $\mathbf{M}_f$. We randomly select several AU-related regions $\mathbf{R}$ from the pre-defined action unit masks. These selected areas are removed from the facial mask $\mathbf{M}_f$, and then we obtain the Partial Face Mask $\mathbf{M}_p$ after Gaussian Blur.

The Image-level Tampering is then performed on real images. As proposed in [45], the pseudo target and source images $\mathbf{I}_t, \mathbf{I}_s$ are created from a single real image $\mathbf{I}$ through different augmentation pipelines:

$$\mathbf{I}_t, \mathbf{I}_s = \Phi_1(\mathbf{I}), \Phi_2(\mathbf{I}), \tag{8}$$

where Φ_1, Φ_2 both randomly perform color transformations and frequency transformations. By blending the source image $\mathbf{I}_s$ and the target image $\mathbf{I}_t$ with the mask $\mathbf{M}_p$, we get the pseudo image $\mathbf{I}_f$ as

$$\mathbf{I}_f = \mathbf{I}_s \odot \mathbf{M}_p + \mathbf{I}_t \odot (1 - \mathbf{M}_p). \tag{9}$$

Note that the background and the removed AU regions $\mathbf{R}$ remain unchanged during the blending.

Moreover, we implement the Feature-level Mixing on the AU-related regions in the Partial Face Mask $\mathbf{M}_p$, which is detailed in Fig. 4. The feature space has more dimensions than the input space, so that more diverse and challenging samples can be obtained for forgery detection. Without loss of generality, we take features in layer i as an example. Let $\mathbf{F}^r, \mathbf{F}^f \in \mathbb{R}^{C^i\times H^i\times W^i}$ represent real feature and fake feature, respectively, which are corresponding to the activations of real and fake images. Let $\mathbf{M}'_p \in \mathbb{R}^{1\times H^i\times W^i}$ be the resized Partial Face Mask. We first calculate the channel-wise mean and standard deviation $\boldsymbol{\mu}^r$, $\boldsymbol{\mu}^f$, $\boldsymbol{\sigma}^r$, $\boldsymbol{\sigma}^f \in \mathbb{R}^{C^i}$:

$$\boldsymbol{\mu}^{\{r,f\}} = \frac{\sum_{h,w} (\mathbf{F}^{\{r,f\}} \odot \mathbf{M}'_p)}{\sum_{h,w} \mathbf{M}'_p}, \tag{10}$$

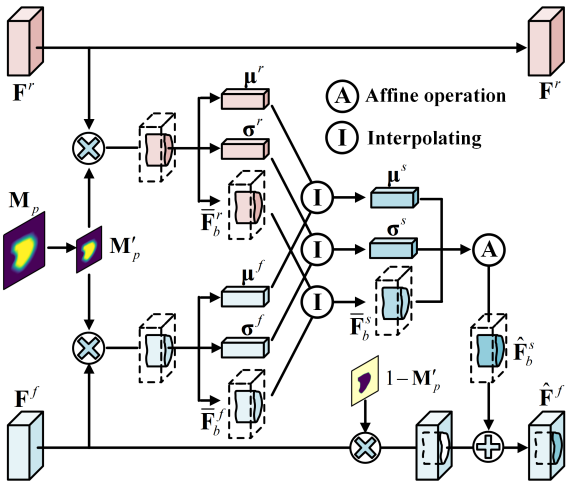

Figure 4. **The pipeline of the Feature-level Mixing (FM).**

$$\boldsymbol{\sigma}^{\{r,f\}} = \sqrt{\frac{\sum_{h,w}\left[\left(\mathbf{F}^{\{r,f\}} - \boldsymbol{\mu}^{\{r,f\}}\right)^2 \odot \mathbf{M}'_p\right]}{\sum_{h,w}\mathbf{M}'_p}} + \varepsilon. \quad (11)$$

Note that the mean and standard deviation are only calculated on the tampering region with the aid of mask $\mathbf{M}'_p$. Subsequently, we normalize the features in the tampering region and gain the normalized region features $\bar{\mathbf{F}}_b^r$, $\bar{\mathbf{F}}_b^f$:

$$\bar{\mathbf{F}}_b^{\{r,f\}} = \frac{\left(\mathbf{F}^{\{r,f\}} - \boldsymbol{\mu}^{\{r,f\}}\right) \odot \mathbf{M}'_p}{\boldsymbol{\sigma}^{\{r,f\}}}. \quad (12)$$

We then mix the statistical features of real features into those of the fake features, and obtain the scale $\boldsymbol{\sigma}^s$, bias $\boldsymbol{\mu}^s$ and the synthesized normalized features $\bar{\mathbf{F}}_b^s$:

$$\boldsymbol{\mu}^s = \lambda_1 \boldsymbol{\mu}^f + (1-\lambda_1)\boldsymbol{\mu}^r, \quad (13)$$

$$\boldsymbol{\sigma}^s = \lambda_2 \boldsymbol{\sigma}^f + (1-\lambda_2)\boldsymbol{\sigma}^r, \quad (14)$$

$$\bar{\mathbf{F}}_b^s = \lambda_3 \bar{\mathbf{F}}_b^f + (1-\lambda_3)\bar{\mathbf{F}}_b^r, \quad (15)$$

where $\lambda_1, \lambda_2, \lambda_3 \in [0,1]$ are sampled from a beta distribution. The synthesized normalized feature $\bar{\mathbf{F}}_b^s$ is affined with the scale and bias to get the synthesized region feature $\hat{\mathbf{F}}_b^s$:

$$\hat{\mathbf{F}}_b^s = \boldsymbol{\sigma}^s \odot \bar{\mathbf{F}}_b^s + \boldsymbol{\mu}^s \odot \mathbf{M}'_p. \quad (16)$$

Subsequently, we add the synthesized region feature $\hat{\mathbf{F}}_b^s$ back to the original fake feature as follows:

$$\hat{\mathbf{F}}^f = \mathbf{F}^f \odot (1 - \mathbf{M}'_p) + \hat{\mathbf{F}}_b^s. \quad (17)$$

In the whole pipeline of Feature-level Mixing, only the tampering regions of the fake feature are modified, while real features as well as other regions of the fake feature remain unchanged. In each training step, the Feature-level Mixing is performed at a randomly chosen layer to enrich fake samples.

Local Tampering Supervision. Recently transformer-based face forgery detectors normally utilize class tokens that aggregate global information to identify face forgery. However, this strategy neglects the role of other tokens that encode rich information on their respective local image areas. In Local Tampering Supervision, we assign each patch token with individual location-specific supervision indicating the existence of the tampering operation on the corresponding image area.

Fig. 2 provides an intuitive interpretation. Given an image $\mathbf{I}$, we denote the output of the last transformer block as $[\mathbf{x}_c, \mathbf{x}_1, \cdots, \mathbf{x}_n]$, where $\mathbf{x}_c$ and $\mathbf{x}_1, \cdots, \mathbf{x}_n$ correspond to the class token and n patch tokens, respectively. To utilize the local information inside the patch tokens, we first generate individual location-specific supervision based on the Partial Face Mask $\mathbf{M}_p \in \mathbb{R}^{H\times W}$. Specifically, the mask $\mathbf{M}_p$ is downsampled to the size of $(\sqrt{n} \times \sqrt{n})$ and is then reshaped to get the coarse version $\mathbf{y}' \in \mathbb{R}^n$. The token label $\mathbf{y}^{token} \in \mathbb{R}^n$ is generated as follow:

$$\mathrm{y}_i^{token} = \begin{cases} 1 & \mathrm{y}'_i > t \\ 0 & \mathrm{y}'_i \le t \end{cases} \qquad i = 1, \cdots, n \quad (18)$$

with the parameter $t = 0.5$. The cross-entropy loss is calculated between each patch token and the corresponding aligned token label as an auxiliary loss at the training phase. The token labeling objective can be defined as:

$$\mathcal{L}_{token} = \frac{1}{n}\sum_{i=1}^{n} H(fc(\mathbf{x}_i), \mathrm{y}_i{}^{token}), \quad (19)$$

where H is the softmax cross-entropy loss and fc is a shared full-connected layer performed on each patch token. During inference, the prediction is calculated based on the output class token and patch tokens. More details are provided in the supplementary.

The total loss function can be written as:

$$\mathcal{L}_{total} = \lambda_c \mathcal{L}_{cls} + \lambda_t \mathcal{L}_{token} + \lambda_a \mathcal{L}_{att} \quad (20)$$

where $\mathcal{L}_{cls}$ is the original image-level loss function, $\mathcal{L}_{token}$ is the token labeling objective, $\mathcal{L}_{att}$ is the Huber loss function mentioned in Equ. 1. λ_c, λ_t, λ_a are the balancing weights for these terms. By default, we set $\lambda_c = 1$, $\lambda_t = 30$, $\lambda_a = 0.5$ in our experiments.

4. Experiments

4.1. Implementation Details

Pre-processing. For each video frame, face crops are detected by using RetinaFace [12] and landmarks are detected by the public toolbox Dlib [28]. All face crops are resized to 224×224. To acquire ground-truth attention map

Method	Testing Set (AUC (%))				
	DF	F2F	FS	NT	FF++
MIL [51]	99.51	98.59	94.86	97.96	97.73
XN-avg [42]	99.38	99.53	99.36	97.29	98.89
Face X-ray [31]	99.12	99.31	99.09	99.27	99.20
S-MIL-T [33]	99.84	99.34	99.61	98.85	99.41
PCL+I2G [55]	**100.00**	99.57	**100.00**	99.58	99.79
SOLA [19]	**100.00**	99.67	**100.00**	**99.82**	99.87
Ours	**100.00**	**99.86**	99.98	99.71	**99.89**

Table 1. **In-dataset evaluation on FF++ (raw).** Bold and underlined values correspond to the best and second-best values, respectively. Our approach achieves competitive performance.

$\mathbf{Y}_{Att}$, landmarks specific to each action unit are defined similarly to [32, 44]. We fit ellipses to landmarks as the initial regions of each action unit, smooth the image (Gaussian with $\sigma = 3$), and then obtain 15 action unit masks. These masks are concatenated in the channel direction to obtain the ground-truth map $\mathbf{Y}_{Att}$. Note that the landmarks and the $\mathbf{Y}_{Att}$ are not needed during inference; hence we only use RetinaFace to crop faces at the inference time.

Training. We adopt the part of Xception [8] up to $block11$ as our backbone. The Xception is initialized with weights pre-trained on ImageNet [11]. Given an input image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, we first feed it into backbone, and features $\mathbf{F} \in \mathbb{R}^{C\times H'\times W'}$ after the $block11$ layer of Xception are extracted for further calculations, where $H' = H/16$ and $W' = W/16$. We train the model for 100 epochs with the SGD [41] optimizer. The batch size and learning rate are set to 32 and 0.001, respectively. We sample only eight frames per video for training. Each batch consists of real images and their corresponding generating images.

Datasets. Following the convention, we adopt the FaceForensics++ (FF++) [42] for training. FF++ is a large-scale benchmark dataset containing 1000 original videos from youtube and corresponding fake videos which are generated by four typical manipulation methods: *i.e.,* Deepfakes (DF) [2], Face2Face (F2F) [48], FaceSwap (FS) [3] and NeuralTextures (NT) [47]. We split FF++ following the official splits and use the HQ version by default or specify the version otherwise. To evaluate the generalization of our method, we also conduct experiments on the recent proposed face forensic dataset, i.e., Celeb-DF (CDF) [36], DeepfakeDetection (DFD) [1], Deepfake Detection Challenge (DFDC) [13] and DFDC Preview (DFDCP) [14], Wild-Deepfake (FFIW) [57]. We also follow the official splits to construct testing sets.

4.2. In-dataset Evaluation

In the in-dataset evaluation, we train the ART on both real and fake data from training split of FF++ [42], without the proposed TAP. The results are shown in Table 1. We provide more in-dataset evaluation results on CDF [36] and DFDCP [14] in the supplementary material. Our model achieves competitive performance on FF++ dataset. Specially, we achieve the best performance on the DF and F2F. Although the performances of previous methods (*e.g.,* SOLA [19], PCL+I2G [55]) tend to saturate on FF++, we still obtain the best performance of 99.89%. Overall, the ART obtains excellent results, proving that the AU relation learning benefits face forgery detection.

Method	Testing Set AUC (%)				
	DF	F2F	FS	NT	FF++
Face X-ray+BI [31]	99.17	98.57	98.21	98.13	98.52
PCL+I2G [55]	**100**	98.97	99.86	97.63	99.11
Xception+SBI [45]	99.99	**99.90**	98.79	98.20	99.22
Ours	99.98	99.60	**99.89**	**98.38**	**99.46**

Table 2. **Cross-manipulation evaluation on FF++ (raw).** Our method achieves the best results on FS, NT, and the whole FF++.

4.3. Generalization Ability Evaluation

In real detection situations, defenders are generally unaware of the attacker's forgery methods. For this reason, we perform evaluations to verify the model generalization to various forgery methods

Cross-Manipulation Evaluation. Following the evaluation protocol used in [45], we evaluate our model on four manipulation methods of FF++. Table 2 presents our cross-manipulation evaluation result on FF++, where we only used the real videos of FF++ for training. Our method outperforms or nearly equals the similar existing methods on four manipulations (99.98% on DF, 99.60% on F2F, 99.89% on FS, and 98.38% on NT) and achieves the best performance on the whole FF++ (99.46% vs. 99.22%). This result shows that our method works well not only on deepfakes but also on other face manipulations.

Cross-Dataset Evaluation. To show the generality of our method, we further conduct a cross-dataset evaluation where our framework is trained on the real images of FF++ and evaluated on other recently released datasets. Table 3 presents the cross-dataset evaluation results. As seen, our method outperforms other models in all the cases and achieves the overall best performance. Compared with video-level methods, Our approach outperforms the state-of-the-art transformer-based method FTCN [56] on CDF, DFD, DFDC, DFDCP and FFIW by 5.87%, 4.82%, 2.82%, 12.16% and 6.98% points, respectively, and improves the performance by 6.53% points on average (86.68% vs. 80.15%). One possible reason is that FTCN, trained on real and fake videos, still focuses on a particular forgery pattern of FF++, which hinders the performance of generalization. LipForensics [24] targets high-level semantic irregularities in the mouth region and leads to improved generalization performance. However, their method may ignore the forensic clues in other facial action units. This limitation may explain why their performance is inferior to ours.

Compared with frame-level methods trained on real and fake images (*i.e.*, LRL [7], FRDM [38], DCL [46]), our approach outperforms their methods on CDF by more

Method	Year	Input Type	Training Set		Testing Set AUC (%)				
			Real	Fake	CDF	DFD	DFDC	DFDCP	FFIW
DAM [57]	2021	Video	✓	✓	75.30	-	-	72.80	-
LipForensics [24]	2021	Video	✓	✓	82.40	-	73.50	-	-
FTCN [56]	2021	Video	✓	✓	86.90	94.40†	71.00†	74.00	74.47†
RealForensics [23]	2022	Video	✓	✓	86.90	-	-	-	-
FInfer [25]	2022	Video	✓	✓	70.60	-	-	70.39	69.46
Face X-ray+BI [31]	2020	Frame	✓	✗	-	93.47	-	71.15	-
Face X-ray+BI [31]	2020	Frame	✓	✓	-	95.40	-	80.92	-
LRL [7]	2021	Frame	✓	✓	78.26	89.24	-	76.53	-
FRDM [38]	2021	Frame	✓	✓	79.40	91.90	-	79.70	-
PCL+I2G [55]	2021	Frame	✓	✗	90.03	99.07	67.52	74.37	-
ICT‡ [15]	2022	Frame	✗	✗	85.71	84.13	-	-	-
DCL [46]	2022	Frame	✓	✓	82.30	91.66	-	76.71	71.14
Xception+SBI [45]	2022	Frame	✓	✗	90.27	96.21	70.77*	78.85	76.72
Ours	2022	Frame	✓	✗	**92.77**	**99.22**	**73.82**	**86.16**	**81.45**

Table 3. **Cross-dataset evaluation on CDF, DFD, DFDC, DFDCP, and FFIW.** The results of prior methods are directly cited from the original paper for a fair comparison. †:denotes experiments performed by [45]. ‡: ICT [15] is trained on MS-Celeb-1M [22]. *: we experiment with the official code. Our method outperforms state-of-the-art methods and presents excellent generalization abilities.

Variants	Testing Set AUC (%)			Avg
	CDF	DFDCP	FFIW	
Backbone	88.10	80.99	74.45	81.18
ART w/o SPB.	90.39	83.68	78.13	84.07
ART w/o AGB.	90.60	84.11	77.10	83.94
TAP w/o $\mathbf{M}_p$.	91.48	84.47	77.95	84.63
TAP w/o IT.	88.98	79.85	72.60	80.48
TAP w/o FM.	89.56	79.51	74.16	81.08
TAP w/o LS.	90.80	82.17	78.06	83.68
Ours	**92.77**	**86.16**	**81.45**	**86.79**

Table 4. **Effect of each branch in ART and each process in TAP.** The absence of any block causes performance degradation.

than 10% in terms of AUC. Compared with similar approaches trained on real and synthesized images (*i.e.*, Face X-ray+BI [31], PCL+I2G [55], Xception+SBIs [45]), our method still exhibits better generalization performance. Specially, we outperform the state-of-the-art Xception+SBIs on CDF, DFD, DFDC, DFDCP and FFIW, by about 2.50%, 3.01%, 3.05%, 7.31% and 4.73% in terms of AUC, and improve the baseline by 4.12% points on average (86.68% vs. 82.56%). Our framework achieves better performance, probably due to the elaborate Action Units Relation Transformer, and the generation of richer forgery samples in the image space and feature space.

4.4. Analysis

This section analyzes the effectiveness of each branch in ART, each process in TAP, the hyper-parameter in LS. We provide more analyses in the supplementary material.

Effect of Each Branch in ART. We compare our method with the following variants under the proposed TAP. (1) Backbone: the features after our backbone are extracted and fed into the global average pooling and full-connected layer. (2) w/o SPB: ART without the AU-specific branch. (3) w/o AGB: ART without the AU-agnostic branch.

The experimental comparison is shown in Table 4. Both the AU-specific Branch and the AU-agnostic Branch lead to the performance improvements (2.89% and 2.76%, on average) over the backbone. This shows the effectiveness of exploiting AU relation cues in face forgery detection. We also observe that the model performs marginally better with AGB (84.07%) than with SPB (83.94%). This may be attributed to the usage of Transformer encoder, which models correlations between image patches and obtains comprehensive forensic clues. However, AGB ignores the local information on the semantics of the face, which is beneficial for deepfake detection. When we combine AGB and SPB, the performance of our model is raised from 81.18% to 86.79%. This presents the effectiveness of simultaneously considering the relation between different Action Units at the AU-agnostic patches and the AU-specific regions.

Effect of Each Process in TAP. We compare our methods with the following variants. (1) w/o $\mathbf{M}_p$: the Partial Face Mask $\mathbf{M}_p$ is replaced with the convex hull of facial landmarks (*i.e.,* facial mask $\mathbf{M}_f$). (2) w/o IT: we remove Image-level Tampering from TAP. (3) w/o FM: the Feature-level Mixing is removed from TAP. (4) w/o LS: we disable Local Tampering Supervision in the pipeline of TAP.

The experimental comparison is shown in Table 4. Replacing the $\mathbf{M}_p$ with the facial mask $\mathbf{M}_f$ results in a drop of 2.16% on average. This suggests that randomly removing AU regions from $\mathbf{M}_f$ is necessary for the whole process. One reason is that the Partial Face Mask $\mathbf{M}_p$ provides dynamically changing AU regions for subsequent tamper-

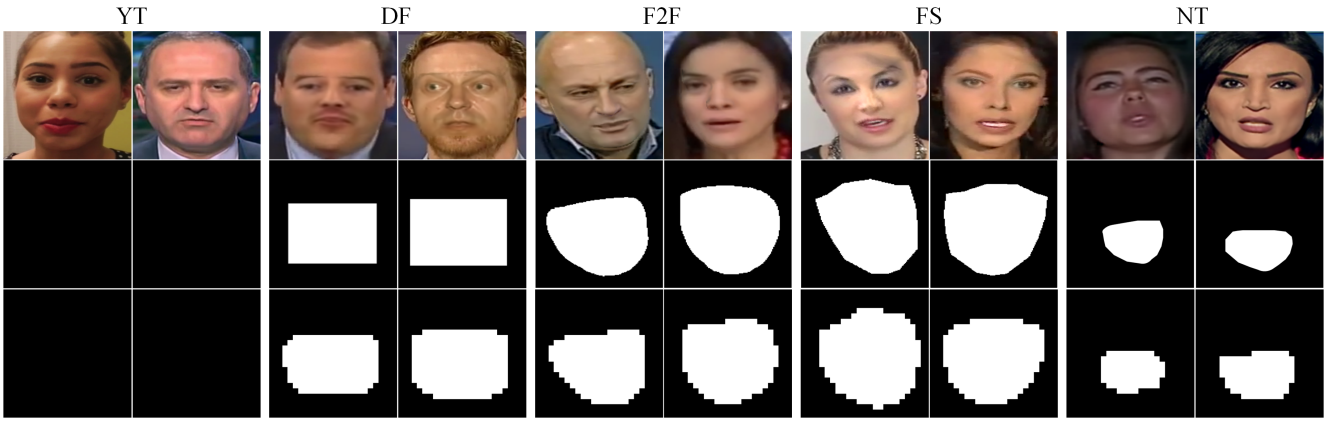

Figure 5. **Visual results on various manipulation methods.** The images in the first, second, and third rows are input, ground truth, and prediction, respectively. To detect the tampered area more accurately, we use images with the size of 384×384 as input and obtain masks with the size of 24×24 to indicate the tampered area. It can be clearly seen that our framework well captures the manipulation regions.

Hyper-Parameter	Testing Set AUC (%)			Avg
	CDF	DFDCP	FFIW	
$\lambda_t = 0$	90.80	82.17	78.06	83.68
$\lambda_t = 10$	91.98	84.87	81.03	85.96
$\lambda_t = 30$	92.77	**86.16**	81.45	86.79
$\lambda_t = 50$	**93.01**	85.65	82.59	**87.08**
$\lambda_t = 100$	92.95	85.39	**82.73**	87.02

Table 5. **Effect of Local Tampering Supervision.** The performance is considerably enhanced by the use of proper λ_t.

ing and prediction, which facilitates the diversity of forgery samples and increases the difficulty of the auxiliary task. We also observe that Image-level Tampering and Feature-level Mixing reproduce important artifacts because of the significant performance drop without them (from 86.79% to 80.48% and 81.08%, respectively). Additionally, without the Local Tampering Supervision, our approach suffers a drop of 3.11%. This demonstrates that guiding the model to learn additional location-specific forensic clues is actually effective for face forensic detection.

Effect of Local Tampering Supervision. In this subsection, the models are trained on real data of FF++ with increasing λ_t. The results are shown in Table 5. We observe that the models with Local Tampering Supervision ($\lambda_t > 0$) outperform the model with the binary classification loss alone ($\lambda_t = 0$). Especially, the model trained with parameter $\lambda_t = 50$ achieves a better AUC score (3.40%) than trained with parameter $\lambda_t = 0$, on average. The results validate that it is beneficial to use proper λ_t during training, which also suggests that the Local Tampering Supervision plays an important role in the success.

4.5. Qualitative Results

Our framework enhances the representation learning for deepfake detection while also generating interpretable visualizations clues about the modified region. To represent the tampered regions, the 1D prediction results based on patch tokens are converted into 2D masks. Figure 5 visualizes some prediction results (the third row) along with the corresponding input images (the first row) and ground truth (the second row). When feeding an authentic image, in most cases, the visualization is a pure blank image, indicating that the input image has not been manipulated. When testing a fake image, the predicted mask can adequately match the ground truth. The results demonstrate that our framework can capture the corresponding manipulation region rather than simply segmenting the full-face part.

5. Conclusion

In this paper, we propose the Action Units Relation Learning framework, which consists of the Action Units Relation Transformer (ART) and the Tampered AU Prediction (TAP). In ART, we model the relation between different Action Units at the AU-agnostic patches and the AU-specific regions. The two levels of relation learning complement each other and work together to uncover forgery clues. We also formulate an auxiliary task, *i.e.,* the Tampered AU Prediction, to implement Image-level Tampering and Feature-level Mixing in the AU-related regions and enhance the capacity of the model to sense local forgery defects with the Local Tampering Supervision. Experimental results showed that our framework is competitive against state-of-the-art methods on popular datasets, providing a strong baseline for future research.

Acknowledgements. This work was supported by the National Key Research and Development Program of China (Grant No. 2020AAA0106800), the National Natural Science Foundation of China (No. 62192785, No. 61972071, No. U1936204, No. 62122086, No. 62036011, No. 62192782, No. 61721004, No. U2033210, No. 62172413), the Beijing Natural Science Foundation No. M22005 and L223003, the Major Projects of Guangdong Education Department for Foundation Research and Applied Research (Grant: 2017KZDXM081, 2018KZDXM066), Guangdong Provincial University Innovation Team Project (Project No. 2020KCXTD045). The work of Bing Li was also supported by the Youth Innovation Promotion Association, CAS.

References

[1] Contributing data to deepfake detection research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. Accessed 2022-11-10. 2, 6

[2] Deepfakes. https://github.com/deepfakes/faceswap. Accessed 2022-11-10. 2, 6

[3] Faceswap. https://github.com/MarekKowalski/FaceSwap/. Accessed 2022-11-10. 2, 6

[4] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018. 1

[5] Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE Transactions on Image Processing*, 28(7):3286–3300, 2019. 2

[6] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18710–18719, 2022. 1

[7] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1081–1088, 2021. 6, 7

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 6

[9] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 1

[10] 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. 1, 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[12] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. 5

[13] 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. 2, 6

[14] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019. 2, 6

[15] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9468–9478, 2022. 2, 7

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[17] Mengnan Du, Shiva Pentyala, Yuening Li, and Xia Hu. Towards generalizable deepfake detection with locality-aware autoencoder. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 325–334, 2020. 1

[18] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 2

[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. 6

[20] 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. 2

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

[22] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. 7

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

[24] 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. 6, 7

[25] Juan Hu, Xin Liao, Jinwen Liang, Wenbo Zhou, and Zheng Qin. Finfer: Frame inference-based deepfake detection for high-visual-quality videos. *IEEE Trans. Pattern Anal. Mach. Intell*, pages 1–9, 2022. 7

[26] Tackhyun Jung, Sangwon Kim, and Keecheon Kim. Deepvision: Deepfakes detection using human eye blinking pattern. *IEEE Access*, 8:83144–83154, 2020. 2

[27] Ali Khodabakhsh, Raghavendra Ramachandra, Kiran Raja, Pankaj Wasnik, and Christoph Busch. Fake face detection methods: Can they be generalized? In *2018 international conference of the biometrics special interest group (BIOSIG)*, pages 1–6. IEEE, 2018. 1

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

[29] Prabhat Kumar, Mayank Vatsa, and Richa Singh. Detecting face2face facial reenactment in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2589–2597, 2020. 1

[30] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6458–6467, 2021. 2

[31] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020. 1, 2, 6, 7

[32] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2583–2596, 2018. 6

[33] Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. Sharp multiple instance learning for deepfake video detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1864–1872, 2020. 6

[34] 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. 2

[35] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018. 2

[36] 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. 2, 6

[37] 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. 2

[38] 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. 2, 6, 7

[39] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2019. 1

[40] 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. 2

[41] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 6

[42] 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. 1, 6

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

[44] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European conference on computer vision (ECCV)*, pages 705–720, 2018. 6

[45] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022. 1, 2, 4, 6, 7

[46] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2316–2324, 2022. 6, 7

[47] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2, 6

[48] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2, 6

[49] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14923–14932, 2021. 1, 2

[50] 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. 1, 2

[51] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018. 6

[52] Xinsheng Xuan, Bo Peng, Wei Wang, and Jing Dong. On the generalization of gan image forensics. In *Chinese conference on biometric recognition*, pages 134–141. Springer, 2019. 1

[53] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019. 2

[54] 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. 2

[55] 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. 1, 2, 6, 7

[56] 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. 2, 6, 7

[57] Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5778–5788, 2021. 2, 6, 7