Real Appearance Modeling for More General Deepfake Detection

Jiahe Tian[1,2], Cai Yu[1,2], Xi Wang[3], Peng Chen[4], Zihao Xiao[4], Jiao Dai[2], Jizhong Han[2], and Yesheng Chai[2]

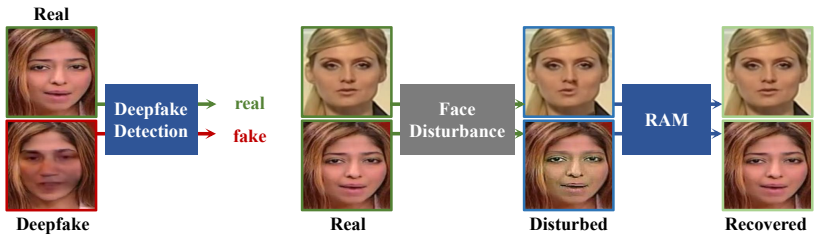
[1] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China
[2] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100094, China
[3] Institute of Microelectronics, Chinese Academy of Sciences, Beijing 100094, China
[4] RealAI Inc., Beijing 100094, China

Abstract. Recent studies in deepfake detection have shown promising results when detecting deepfakes of the same type as those present in training. However, their ability to generalize to unseen deepfakes remains limited. This work improves the generalizable deepfake detection from a simple principle: an ideal detector classifies any face that contains anomalies not found in real faces as fake. Namely, detectors should learn consistent real appearances rather than fake patterns in the training set that may not apply to unseen deepfakes. Guided by this principle, we propose a learning task named Real Appearance Modeling (RAM) that guides the model to learn real appearances by recovering original faces from slightly disturbed faces. We further propose Face Disturbance to produce disturbed faces while preserving original information that enables recovery, which aids the model in learning the fine-grained appearance of real faces. Extensive experiments demonstrate the effectiveness of modeling real appearances to spot richer deepfakes. Our method surpasses existing state-of-the-art methods by a large margin on multiple popular deepfake datasets.

Keywords: Deepfake Detection · Face Forgery Detection · Multimedia Forensic

1 Introduction

Current deep generative models allow for the generation of realistic fake faces, *i.e.*, deepfakes, making it easy to manipulate face media. Therefore, these techniques can be used for committing fraud, bypassing identity authentication, or spreading false information. The potential misuses raise concerns regarding privacy and reputation, and thus deepfake detectors are needed to counter the risks posed by deepfakes. Moreover, the proliferation of deepfake algorithms presents a challenge to the generalizability of deepfake detection, making it crucial to develop detectors that generalize across various deepfake types.



(a) Deepfake Detection. **(b)** The RAM task.

Fig. 1: Illustration of (a) the Deepfake Detection task and (b) the proposed RAM task. Compared to Deepfake Detection, the RAM task guides models to restore disturbed faces back to their original appearance.

Previous works usually tackled deepfake detection by guiding detectors to recognize specific deepfake patterns, such as the boundaries in swapped faces [1,2], inconsistency [3,4], frequency anomalies [5,6], and movement anomalies [7,8]. They achieved good performance in detecting deepfakes of the same types as those present in training. However, their performance often significantly deteriorates when detecting deepfakes generated by novel algorithms, which mirrors real-world scenarios where various deepfakes emerge. This is due to different deepfake algorithms producing different fake patterns, and learned patterns are not universally applicable to all deepfakes. Namely, more than knowledge of specific deepfake patterns is required to generalize to unseen types of deepfake.

Though different deepfakes exhibit different patterns, real faces maintain consistent appearance. Thus, it is better to learn the appearance of real faces to verify their authenticity instead of learning deepfake patterns to spot fakes. Namely, an ideal classifier would classify faces as fake if they exhibit patterns not found in real faces, *i.e.*, anomalies. As the detector learns explicit real face appearance, it should be capable of predicting real faces at pixel-level. This guides us to design a task that densely predicts real faces to learn real appearance. To fully define this task, dense prediction targets and input information that assists the prediction are needed. Specifically, we propose the Real Appearance Modeling (RAM) task that recovers corresponding original faces from disturbed faces, as illustrated in Figure 5, where original faces are prediction targets and the inputted disturbed faces contain information needed for prediction. For the RAM task, Deepfakes are not applicable to serve as training data since their original identities or poses are altered, and the necessary information that enables prediction is lost. We thus propose Face Disturbance to disturb real faces to *pseudo-fakes* as training data for RAM. In the disturbed faces produced by Face Disturbance, the original information is mostly retained. Then, the detector learns real face appearance by predicting original faces.

To disturb real faces, the proposed Face Disturbance includes two parts, Texture and Structure Disturbance, to introduce anomalies in textures and structures, respectively. This is based on the premise that an image can be regarded as composed of structures and textures [9,10]. Texture Disturbance disturbs partial

textures to produce anomalies in the form of texture inconsistency. The reason for disturbing partially rather than globally is that distribution of real textures varies, and globally modifying facial textures may fail to produce anomalies. Meanwhile, having inconsistent textures between different facial parts is indeed anomalous. Hence, Texture Disturbance is designed to produce inconsistencies. Structure Disturbance disturbs structure by covering key facial areas, including facial features and boundaries, which are essential components of facial structure. The specific covering method involves copying and tiling the surrounding texture to the disturbed areas to ensure the original texture is intact. Jointly using Texture and Structure Disturbance to produce disturbed faces for RAM helps the detector to learn real structures and textures to spot anomalies in unseen deepfakes.

The proposed RAM uses disturbed faces produced by Face Disturbance as inputs and the original faces as prediction targets, making this task unsupervised. Therefore, RAM is adopted as an auxiliary task to enhance deepfake detectors. In training, real and disturbed faces are used as inputs, and the detector is expected to both discriminate disturbed faces and recover their real appearances. Specifically, RAM involves guiding the detector to recover structure and texture anomalies. For structure anomalies, the detector recovers the structure of the disturbed areas with the undisturbed areas as the reference. For texture anomalies, the detector recovers the inner texture based on the texture of the outer face. By leveraging learned real appearances, our model is equipped to detect anomalies within unseen deepfakes. As illustrated in Figure 5, our model diverges from the baseline model by grouping various deepfakes to general anomalies rather than identifying specific deepfake patterns. This indicates RAM narrows the gap across deepfakes and thus improves the generalizability.

Our contributions can be summarized as follows:

- We introduce Real Appearance Modeling to guide detectors to learn the appearance of real faces by restoring disturbed faces to their original appearances, which helps to improve the generalization of deepfake detectors.
- We propose Face Disturbance to disturb real faces in texture and structure views as pseudo-fakes, which aids the detector in learning real texture and structure appearance when predicting original faces.
- We conduct extensive experiments using various evaluation protocols to demonstrate the generalizability of our framework, which surpasses previous methods on multiple popular deepfake datasets.

2 Related Work

2.1 Deepfake Detection

Early works leveraged biological artifacts, including distorted pupils [11], heartbeat frequency [12], and anomalous head pose [13] for deepfake detection. More recent works treated deepfake detection as a binary classification task and focused on detecting specific deepfake patterns. For instance, [14] investigated

mesoscopic features in shallow networks that contain rich features, while [15] employed an attention mechanism to combine RGB and texture features. [6, 16] leveraged high-frequency in the discrete cosine domain as complementary modalities. Another direction involves employing proxy tasks for pretraining to learn high-level representations capable of identifying deepfake videos. Examples of such proxy tasks include lip-reading [8] and audio-visual contrastive learning [17, 18]. Recently, the unsupervised learning object MAE [19] is introduced in deepfake detection [20, 21]. These works usually recover masked facial features through dense prediction to model real faces. As they use masked images as input where information is lost, the set recovery target, *i.e.*, original face, is merely one of the multiple feasible solutions. In contrast, RAM for each disturbed face has a definitive prediction target and thus is not an ill-posed problem with multiple solutions. Compared to MAE, RAM is more fine-grained in supervising detectors to model real face appearance.

2.2 Pseudo-fakes for Deepfake Detection

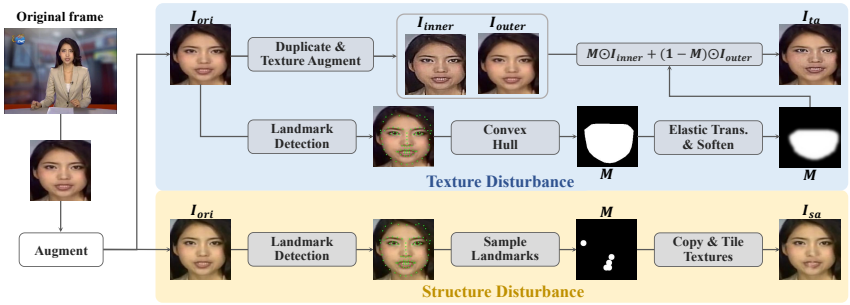
Instead of training detectors using deepfakes in deepfake datasets, several studies [1, 2, 4, 22–24] suggested constructing pseudo-fakes for training. Since the introduced pattern in pseudo-fakes is controllable, detectors learn prior patterns accurately. Specifically, they introduced texture inconsistencies to simulate swapped faces and guide detectors to learn this prior pattern. For instance, [1] introduced blur to inner faces to replicate clarity inconsistencies. [2,4,22] blended two faces with similar facial landmarks to generate texture inconsistencies. Moreover, [23, 24] utilized augmentations to produce inconsistencies within a single face, achieving good generalization. Although they generalize well to face swap deepfakes, their performance on deepfakes generated by reenactment algorithms [25, 26] or Diffusion model [27] drops where texture inconsistency is not the major artifacts.

3 Method

Different from most deepfake detection methods, which are typically trained to spot specific deepfake patterns, our approach emphasizes learning real face appearances through the RAM task. The Face Disturbance strategy is adopted to produce disturbed faces as training data for RAM, as illustrated in Figure 2. The subsequent subsections delve into the details of the proposed Face Disturbance strategy, the definition of the RAM task, and the design of the proposed Recovery Autoencoder (RAE) model in Figure 3, which simultaneously conducts deepfake detection and RAM in a multi-task learning manner.

3.1 Face Disturbance

We aim to disturb real faces to pseudo-fake faces as training data. Previous works devoted to producing texture inconsistency, which is a typical pattern of



(a) Face Disturbance

(b) Pseudo-fake Examples

Fig. 2: (a) Face Disturbance with TD and SD. (b) Pseudo-fake examples, where for each character, from left to right, are real faces, faces with texture anomalies I_{ta}, and faces with structure anomalies I_{sa}. We highlight their disturbed areas.

face swap deepfakes. Therefore, these detectors exhibited biased generalization to face swap deepfakes and fell short when detecting reenactment deepfakes in [28–30]. Similarly, employing these data for RAM enables the detector to learn real texture appearance only, making the detector struggle to spot reenactment deepfakes with intact textures. Thus, it is necessary to introduce richer anomalies to help detectors handle a wider spectrum of deepfakes. As an image can be regarded as composed of structures and textures [9, 10], we categorize anomalies in fake faces into two categories: structure and texture anomalies. Therefore, we implement Texture Disturbance (TD) and Structure Disturbance (SD) in Face Disturbance to produce these anomalies. By adopting this divide-and-conquer strategy, we comprehensively disturb faces and thus make the detector learn richer real appearances from both the texture and structure views.

Texture Disturbance The aim of TD is to produce texture anomalies. Given the inherent diversity of real facial textures, which vary in color, brightness, and clarity, a reference facial area is needed to discern whether the textures in other facial areas are abnormal. Namely, texture anomalies lie in the texture inconsistency between facial parts. There are two existing approaches to produce such inconsistency: Face Blending [2, 4] using two faces and Partial Augmentation [23, 24] using one face. Face blending involves selecting two faces with similar poses and then merging the inner region of one face into the other face. Partial Augmentation involves disturbing either the inner or outer facial texture through image augmentation. An important advantage of Partial Augmentation is preserving the original identity, which prevents detectors from being overfitted to the training set identities and harming generalization [31]. Thus, the proposed TD adopts a disturbing scheme similar to Partial Augmentation.

Figure 2 illustrates the process of TD. The input image I_{ori} is duplicated to I_{inner} and I_{outer}, and one of them is augmented in color, brightness, and clarity. I_{inner} and I_{outer} are then fused using a mask M:

$$I_{ta} = M \odot I_{inner} + (1 - M) \odot I_{outer}, \tag{1}$$

where $\odot$ is Hadamard production, I_{inner} is the image that provides the inner face, I_{outer} provides the background, and I_{ta} is the fused face with texture anomaly. M is the mask for the inner face, and its pixel values are between 0 and 1. Specifically, we first apply a landmark detector to the input image to predict its landmarks. Next, we calculate the convex hull of the face area in I to initialize M. To create inconsistent textures between a richer diversity of different areas, we introduce random deformations to M using a 2D elastic transformation. Before using Equation 1 to blend I_{inner} and I_{outer}, we soften M from a binary mask into a soft mask with Gaussian Blur to produce I_{ta} with natural boundaries. This allows the detector to recognize anomalies in long-span texture inconsistencies rather than merely adapt to local boundary artifacts around M.

Structure Disturbance The aim of SD is to introduce anomalies in the structure view to complement texture anomalies. To disturb facial structures, we disturb the elements that constitute the facial structure. Clearly, these elements are facial boundaries and facial features such as the eyes and mouth. We note that facial landmarks are also positioned at these elements. Therefore, it is intuitive to disturb areas around landmarks to create disturbances in facial structures. Besides, to distinguish with TD, it is also necessary for SD not to disturb the original texture. To disturb facial structures while keeping textures intact, we copy and tile the surrounding textures into areas around a few facial landmarks. The crux lies in the fact that the copy-and-tile operation preserves original textures and thus is the ideal method for disturbing structures. This aids the detector in accurately learning real facial structure in RAM, enabling them to distinguish faces with structural anomalies during detection.

Figure 2 illustrates the process of SD. Given an input face image I_{ori}, we first detect its facial landmarks. We randomly sample several landmark points and draw circles around them as the mask M for disturbed areas. We avoid sampling landmarks of symmetric positions in the face, which enables the detector to rely on the structure information from the symmetric region to perform RAM. We then employ the Fast Marching Method [32], a non-learning image inpainting algorithm, which produces less blur and maintains clearer textures than other textures replicating algorithms, to copy and tile the surrounding textures to the disturbed area. After the surrounding textures are tiled to the disturbed areas, the original structures of the disturbed areas are eliminated and become anomalous. Incorporating SD in Face Disturbance emrichs anomalies in constructed pseudo-fakes, which helps the detector generalize to a wider range of deepfakes.

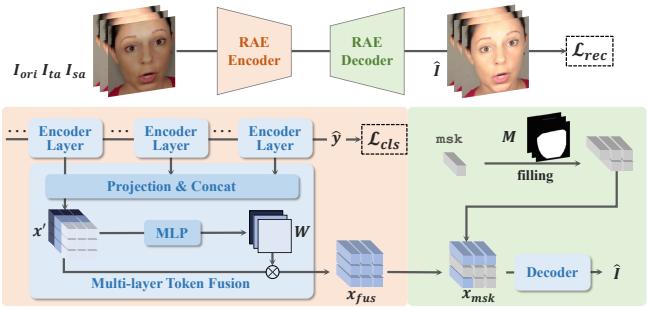


Fig. 3: The proposed RAE is optimized using two learning objects, *i.e.*, $\mathcal{L}_{rec}$ for the RAM and $\mathcal{L}_{cls}$ for the deepfake detection.

3.2 Real Appearance Modeling

We believe that guiding models to learn the appearances of real faces is more beneficial for the generalization across deepfake algorithms than learning specific deepfake patterns. Thus, we propose the novel Real Appearance Modeling (RAM) task to guide detectors to learn real face appearances. RAM requires the model to restore anomalous facial areas produced by Face Disturbance to their original appearances. In comparison to existing methods, our model is guided to learn patterns of real faces instead of deepfake patterns in training fake faces. Then, it classifies faces that display anomalies not found in real faces as fake. Therefore, RAM helps to improve generalization across deepfake algorithms.

Task Definition Given an input face I_{ori}, we obtain pseudo-fake faces I_{ta} with texture anomalies and I_{sa} with structure anomalies after the Face Disturbance strategy. These images are fed into the detector for deepfake detection and RAM, *i.e.*, classification and densely real appearance recovery. In the training procedure, the deepfake detection object and RAM object are optimized in a multi-task learning manner. In deepfake detection, the labels for constructed pseudo-fakes are 1, while for real faces are 0. In RAM, the recovery target I_{target} for I_{sa} is I, while the recovery target I_{target} for I_{ta} is I_{outer}. This is because I_{ta} is derived from I_{outer} and I_{inner}, so there are two potential recovery targets, and we fix the target as I_{outer}. This allows the model to restore the inner texture from I_{inner} using the background texture from I_{outer} as the reference. Thus, the model is guided to align the inner texture with the outer texture. The feasibility of recovering pseudo-fakes to original faces lies in the fact that the pseudo-fakes contains all the information of the corresponding original faces, with only variations between the inner and outer face regions or partially disrupted structure.

3.3 Recovery Autoencoder

In order to perform both deepfake detection and RAM, the detector in our framework requires a classification head and a dense prediction head. Thus, we

incorporate a dense prediction branch for the RAM object into a classifier. This configuration makes our model resemble an autoencoder, thus named Recovery Autoencoder (RAE). The RAE is illustrated in Figure 3, which is constructed based on a vanilla Vision Transformer Base (ViT Base) as the encoder. We adopt eight transformer layers and a linear projection layer as the decoder, and a Multi-layer Token Fusion (MTF) block to extract features from both the shallow and deep encoder layers. We refer readers to [33] for detailed information of the ViT encoder. Following this, we detail the entire pipeline of RAE.

After the image $I \in R^{H\times W\times 3}$ is fed to the ViT encoder, it is extracted to a sequence of tokens $x_l \in R^{(N+1)\times D}$ in each encoder layer. Here, $l \in 1, 2\cdots, L$ is the encoder layer index, $L = 12$ is the depth of the encoder, N is the number of image patches, and D represents the embedding dimension. We note $x_L[0]$ is the `cls` token [33] for classification. The classification process involves passing the `cls` token to a linear layer and a softmax operator to predict the class probabilities $\hat{y} \in \mathbb{R}^2$, and 2 is the number of classes. Then, the remaining tokens in x_l are reshaped to $\mathbb{R}^{h\times w\times D}$ where $h = w = \sqrt{N}$, and then inputted to MTF for cross-layer token fusion and the decoder for RAM.

As there are 12 transformer layers in the ViT encoder, inputting all layers in x_l to MTF is computationally intensive. Thus, we select $S = 3$ layers at equal intervals and tokens in these layers as the inputs to MTF to provide shallow, mid, and deep features. Formally, the selected tokens of shallow, mid, and deep layers are noted as x_s, x_m, and x_d. In MTF, we first project x_s, x_m, and x_d into the same feature space and concatenate them to $x' \in R^{S\times h\times w\times D}$. We employ a linear layer as the projection layer, which is widely used for aligning different spaces. After projection, we further extract and fuse both structural and textural information from different layers. As tokens in different ViT layers have identical dimensions, we adopt a cross-layer attention mechanism for fusion. Specifically, MTF predicts weights $W \in R^{S\times h\times w}$ for each token with an MLP. W is normalized using softmax along the first dimension S, ensuring that the weights for tokens at the same spatial position across S layers sum to 1. W and fused tokens $x_{fus} \in R^{h\times w\times D}$ are derived as:

$$x' = concat(x_s, x_m, x_d), \tag{2}$$

$$W = softmax(\mathrm{MLP}(x')), \tag{3}$$

$$x_{fus} = \sum_{i=0}^{S} W[i] \times x'[i]. \tag{4}$$

After extracting x_{fus}, we aim to recover the original face with x_{fus}. While it might be intuitive to directly feed x_{fus} into the decoder to recover real faces, this leads to easy mining that outputs the inputted pseudo-fakes directly. This is because the inputted pseudo-fakes and the target real faces are identical in the undisturbed areas $1 - M$ and are similar in disturbed areas M. To avoid such easy mining, we introduce a learnable mask `msk` token to replace tokens of the disturbed areas. Specifically, given a binary mask M for disturbed areas, it is divided into mask patches $m \in R^{h\times w\times p\times p}$, where p is the side length of a patch. If a mask patch is disturbed, the corresponding patch is replaced with `msk`

token, and the mask for masked patches is M_{patch}. After masking, the masked tokens $x_{msk} \in R^{h \times w \times D}$ are added with positional encodings and then fed into the decoder to $\hat{x} \in R^{N \times p \times p \times 3}$. Lastly, through a reshape operation, we obtain the recovered face image $\hat{I} \in R^{H \times W \times 3}$ where $H = W = \sqrt{N} \times p$. We emphasize that the masking is applied after encoding. Thus, the classification head and the predicted $\hat{y}$ remain unaffected. Moreover, information in masked tokens is fused to unmasked tokens in encoding, thus necessary information that enables the recovery is still conveyed to the decoder.

In predicting the real appearance of disturbed areas, RAE extracts texture and structure features from undisturbed areas and semantic information from disturbed areas for RAM. Consequently, the decoder can utilize the combination of these features to model the appearance of disturbed areas through pixel-level prediction. This enhances the real appearance learned in the encoder, thereby improving the generalization of the deepfake detection.

3.4 Loss Function

The proposed RAE is optimized to both discriminate disturbed faces and model real appearance. We use cross-entropy loss $\mathcal{L}_{cls}$ for classification:

$$\mathcal{L}_{cls} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}), \tag{5}$$

where $\hat{y}$ is the predicted class probabilities and y is the class label. Besides, we use the following L1-loss $\mathcal{L}_{rec}$ for RAM:

$$\mathcal{L}_{rec} = \frac{1}{\|M_{patch}\|_1} \left\| M_{patch} \odot (\hat{I} - I_{target}) \right\|_1, \tag{6}$$

where $\|\cdot\|_1$ is the L1-norm, M_{patch} is the mask for disturbed tokens and thus $\mathcal{L}_{rec}$ is performed on the disturbed tokens. The total loss $\mathcal{L}$ for optimizing RAE is formulated as the weighted sum of $\mathcal{L}_{cls}$ and $\mathcal{L}_{rec}$:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha \mathcal{L}_{rec}. \tag{7}$$

We set α as 0.1 to make $\mathcal{L}_{cls}$ and $\mathcal{L}_{rec}$ converge to the same order of magnitude.

4 Experiments

4.1 Implementation Details

Dataset. We conduct evaluations on six widely used deepfake datasets, *i.e.*, FaceForensics++ (FF++) [34], Celeb-DF v2 (CDF) [35], DFD [34], DFDC [36], DFDC preview (DFDCp) [36], and WildDeepfake (WDF) [37]. **FF++** is a deepfake dataset containing 4,000 forged videos and 1,000 real videos. All videos are provided in three compression levels: raw, high-quality (HQ), and low-quality (LQ). **CDF** contains 590 real and 5,639 fake videos corresponding to 59 celebrities using an improved face swap algorithm. **DFD** contains over 3,000 face swap

videos. **DFDC** is a large-scale dataset that contains more than 120,000 deepfake clips generated using eight deepfake techniques. **DFDCp** is the preview dataset for the Deepfake Detection Challenge. Its fake videos are generated using two face swap algorithms. **WDF** consists of 10,000 real videos and 10,000 online collected fake videos. For DFDC, we use videos with one identity following [8,17,38] in case of labeling noise. For other datasets, we use the official dataset splits. We use the real videos of FF++(HQ) to construct pseudo-fakes for training, while other datasets are for testing.

Preprocess. We extract 10 frames at equal temporal intervals from each video in training and 32 frames from each video in testing. Then, the faces are cropped using Retinaface [39], maintaining a 15% margin around each face. Subsequently, we utilize Dlib [40] to extract 81 facial landmarks for every face and save them to a key-value database to boost the Face Disturbance process.

Face Disturbance Details. As for the texture augmentations involved in generating I_{ta}, we employ RGBShift, HueSaturationValue, RandomBrightnessContrast, Downscale, and Sharpen to augment textures of either I_{inner} or I_{outer} to produce inconsistencies. As for the disturbed areas in generating I_{sa}, we draw circles around 2 to 6 random landmarks with a radius sampled between $1/14$ and $1/16$ of the side length. The number and size of the sampled circles for SD ensure have been found empirically to aid in the effectiveness of RAM tasks as shown in supplementary. Then, all images are resized to 224×224 as inputs.

Training Details. We use a ViT Base pretrained on ImageNet [41] to initialize our RAE. RAE is trained using SAM [42] optimizer with a learning rate of 8e-4 and a batch size of 192 for 100 epochs, and we use cosine decay as the learning rate scheduler. Each batch consists of 96 real faces, 48 disturbed faces with texture anomalies, and 48 disturbed faces with structure anomalies. In the PyTorch framework, training with automatic mixed precision requires approximately 60GB of GPU memory.

4.2 Cross-Dataset Evaluation

In cross-dataset evaluation, detectors are trained using faces in FF++(HQ) and evaluated on other datasets to assess their generalization to unknown deepfake techniques. Here, we compare the proposed RAE with recent deepfake detection methods. The results are listed in Table 1, and we annotated the input type and training faces. RAE achieves top generalization performance and outperforms state-of-the-art detectors on CDF, DFDC, and DFDCp by 2.6%, 3.4%, and 3.2%, respectively. Compared to detectors trained using fake faces, such as TALL and IDDDM, the superior generalization of RAE demonstrates the effectiveness of using Face Disturbance to generate pseudo-fakes for training. Compared to detectors trained using pseudo-fakes with texture inconsistency, such as SBI and AltFreezing, RAE exhibits better generalization on face swap datasets such as CDF and DFDCp. As the design of TD is similar to previous works, the effectiveness of RAE against face swap deepfakes indicates RAM helps to spot richer anomalies through learning real appearance. Moreover, previous detectors trained with pseudo-fakes fail to exhibit better performance on

Table 1: Cross-dataset evaluation of deepfake detection methods on CDF, DFD, DFDC, DFDCP, and WDF. All models are trained using FF++(HQ). All compared results are video-level AUC and cited from the original papers or their subsequences.

Method	Input Type	Year	Training Set			Test Set AUC (%)				
			Real	Fake	Pseudo-fake	CDF	DFD	DFDC	DFDCp	WDF
PCL+I2G [4]	Frame	2021	✓		✓	90.0	**99.0**	67.5	74.4	-
ENB4+SBI [23]	Frame	2022	✓		✓	92.9	98.2	72.0	85.5	-
Mover [20]	Frame	2023	✓	✓		87.1	-	-	78.9	82.0
RFFR [21]	Frame	2023	✓	✓		89.0	-	67.8	-	-
IDDDM [31]	Frame	2023	✓	✓		91.2	-	71.5	-	-
SeeABLE [43]	Frame	2023	✓		✓	87.3	-	75.9	86.3	-
FTCN [38]	Clip	2021	✓	✓		86.9	94.4	71.0	74.0	-
RealForensics [17]	Clip	2022	✓	✓		86.9	-	75.9	-	-
AltFreezing [24]	Clip	2023	✓	✓	✓	89.5	98.5	-	-	-
TALL [44]	Clip	2023	✓	✓		90.8	-	76.8	-	-
RAE (Ours)	Frame	2024	✓		✓	**95.5**	**99.0**	**80.2**	**89.5**	**88.4**

Table 2: In-dataset evaluation on four subsets in FF++(HQ). †indicates reproduced on FF++(HQ).

Method	Test Set AUC(%)				
	DF	F2F	FS	NT	Avg
SLADD [22]	-	-	-	-	98.4
ENB4+SBI † [23]	99.2	99.1	99.0	96.7	98.5
SeeABLE [43]	99.2	98.8	99.1	96.9	98.5
RAE(Ours)	**99.6**	**99.1**	**99.2**	**97.6**	**98.9**

Table 3: Frame-level FP rate(%) on real face datasets. Lower FP is better. All models are trained on FF++(HQ).

Method	Test Set FP(%)↓	
	FFHQ	VoxCeleb
RECCE [45]	62.3	41.7
UIA-ViT [46]	40.6	18.9
ENB4+SBI [23]	57.7	59.2
RAE(ours)	**36.9**	**6.9**

DFDC as it contains reenacted faces that are difficult to be imitated by previous pseudo-fakes. Our method demonstrates a major improvement compared to previous methods as SD helps RAE spot structure anomalies in reenacted faces.

4.3 In-dataset Evaluation

In addition to cross-dataset evaluation, another frequently used evaluation protocol is in-dataset evaluation. We follow the protocol in [43] to conduct evaluations on the four types of deepfakes in FF++, *i.e.*, two types of face swap deepfakes DF and FS, and two types of face reenactment deepfakes F2F and NT, where the training data includes both deepfake and pseudo-fake samples. This evaluation protocol helps to assess the ability of our method in enhancing in-dataset deepfake detection. The results are listed in Table 2. The overall performance of RAE on all four deepfake types surpasses previous works. Specifically, RAE outperforms the compared methods in detecting reenacted faces generated by the two reenactment algorithms, *i.e.*, F2F and NT. RAE achieves a detection AUC of 97.6% on NT. This indicates that combining SD and the learning target RAM aids in recognizing distorted facial features in NT deepfakes, while previous works trained with texture inconsistency pseudo-fakes fail.

4.4 Evaluating RAE on Real Face Datasets

In practical deployment, a low False Positive rate (**FP**) is important for a deepfake detector, which indicates a low ratio of misclassifying real faces as fakes. For instance, a detector with a low FP rate in an authentication system helps to avoid having users repeatedly submit their portraits, thereby enhancing the user experience. Besides, at the same AUC, a lower FP helps reduce the adjustment of classification thresholds in practical applications. We use the mega-scale face dataset VoxCeleb [47], and the high-quality face dataset FFHQ [48] to test the FP of the proposed RAE on real faces. For comparison, we used three open-source detectors, RECCE [45], SBI [23], and UIA-ViT [46].

The FP scores are listed in Table 3. Compared to existing arts, RAE achieves the lowest FP on these two datasets. In stark contrast, the average FP scores of the compared RECCE and SBI are larger than 50%, which indicates they misclassified more than half real faces in these two datasets. This validates that RAM successfully makes the detector learn real face appearances and the trained RAE is less likely to classify a real face as fake.

4.5 Evaluation on Diffusion Deepfakes

In the cross-dataset evaluation, we validated the generalizability of our framework across five popular deepfake datasets. However, we noticed these datasets primarily consist of face swap and reenactment deepfakes. As a result, the effectiveness of detectors against other types of deepfakes, such as diffusion-generated faces, was not adequately evaluated. Given the fact that a proliferating number of deepfakes are diffusion-generated faces, we conduct evaluations on a diffusion face dataset DeepFakeFace(DFF) [27]. **DFF** is synthesized by Stable Diffusion Text2Img (T2I), Stable Diffusion Inpaint (Inpaint) [49], and InsightFace (IF) [50]. We conduct evaluation using the three subsets in DFF. For all compared methods, we use their official weights.

The evaluation results are listed in Table 4. RAE outperforms the compared methods in all three subsets in DFF. More significantly, RAE attains an AUC of 70.8% on average, which improves existing SOTA methods by 2.2%. In stark contrast, the AUC scores of RECCE and UIA-ViT on T2I are less than 50%, which indicates they perform worse than randomly guessing. The results on DFF verify the effectiveness of our proposed method in enhancing the detector's generalization to a wider range of deepfakes.

4.6 Ablation Study

We conduct ablations to verify the effectiveness of the RAM task, Face Disturbance strategy, and the MTF block. A ViT Base trained using I_{ta} and classification task is used to serve as the Baseline as it is similar to several previous detectors [1, 23]. We incrementally add the proposed RAM task, SD, and MTF block to the baseline to observe their effectiveness. The results are listed in Table

Table 4: Frame-level AUC on three subsets of diffusion deepfakes in DFF [27]. All models are trained on FF++(HQ).

Method	Test Set AUC(%)			
	T2I	Inpaint	IF	Avg
RECCE [45]	35.1	51.5	63.1	49.9
UIA-ViT [46]	45.2	55.8	74.0	58.3
ENB4+SBI [23]	62.8	71.4	71.5	68.6
RAE(ours)	**64.1**	**73.0**	**75.2**	**70.8**

Table 5: Ablation studies for the proposed Structure Disturbance, RAM task, and MTF module.

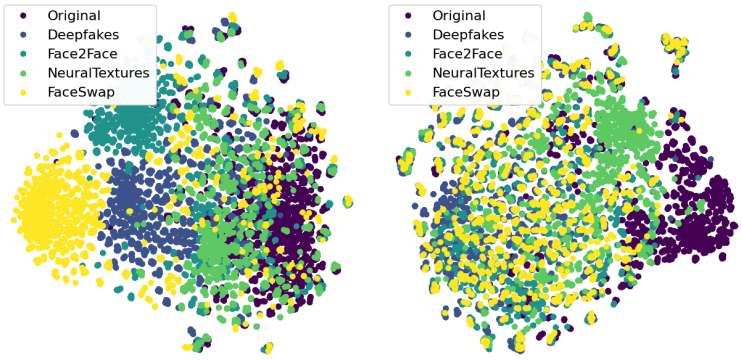
Method	Test set AUC(%)		
	CDF	DFDC	DFDCp
Baseline	92.5	75.5	85.9
+RAM	94.2	77.4	88.6
+I_{sa}	94.7	79.1	88.4
+MTF	**95.5**	**80.2**	**89.5**

Fig. 4: Visualizations of RAM. The left are recovery results for I_{sa} while the right are results for I_{ta}.

5, which show that each of them contributes to enhancing generalization. Compared to solely using a classification task, adding the RAM task significantly improves generalization by 2.2% on average. Furthermore, utilizing I_{sa} to model real facial structures mainly improves the generalization on the DFDC dataset, which includes reenactment deepfakes. Besides, the MTF block aids in improving generalization by leveraging multi-layer features for modeling real appearances.

4.7 Visualization

Face Recovery. We visualize the recovered disturbed faces to verify whether RAE learned real face appearances. The inputted disturbed faces, corresponding original faces, masks for disturbed facial areas, and recovered faces are presented in Figure 4. We observe that our model successfully recovers most of the perturbed areas within the anomalous faces. Though the recovered faces are not explicitly clear, the anomalous patterns are largely eliminated. For faces with structure anomalies, disturbed areas are finely recovered to their original structure appearance. For faces with texture anomalies, the texture pattern of the inner faces is aligned to normal appearance. This demonstrates the feasibility of modeling real appearance by recovering anomalous faces to their original appearance, which provides more fine-grained and interpretable supervision.



(a) Baseline model. **(b)** Our model.

Fig. 5: Distributions of learned representations for faces in FF++(HQ) [34] in the baseline model(a) and our model(b).

Feature Distribution. We apply t-SNE [51] to visualize the features in the last layer of the encoder to illustrate the effect of RAM. We use a ViT Base as the baseline detector and train it using real and fake faces in FF++(HQ). Then, we use the trained RAE and baseline detector to extract features from the test set of FF++(HQ). The results are illustrated in Figure 5. The baseline model learns specific features for each deepfake algorithm, thus separating deepfakes of different types in the feature space. In contrast, the proposed RAE aggregates real faces to a compact cluster in the feature space, which indicates it classifies deepfakes based on the learned general fake patterns. Moreover, RAE mixes different types of deepfakes in the feature space, which confirms that RAM helps to learn general deepfake representations. Compared to the baseline model, the anomalous features learned by RAE are more general across different deepfakes, thus demonstrating better generalization.

5 Conclusion

In this paper, we address the general deepfake detection through guiding the detector to learn real face appearance by recovering original faces from disturbed faces. Anomalies in both the texture and structure views are introduced to faces through the proposed Face Disturbance, and the detector is guided to recover real appearances from these disturbed faces through the proposed RAM task. We then propose the RAE model to conduct both the deepfake detection and the RAM task. Extensive experiments demonstrate the effectiveness of our method against existing methods in general deepfake detection. Future work involves scaling our method with large-scale human face datasets.

This project is supported by the National Key Research and Development Program of China (No.2022YFB2702500)

References

1. Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 46–52. Computer Vision Foundation / IEEE, 2019. 2, 4, 12
2. 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. 2, 4, 5
3. Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 2316–2324. AAAI Press, 2022. 2
4. 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. 2, 4, 5, 11
5. 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. 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, 4
7. 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 2018, Hong Kong, China, December 11-13, 2018*, pages 1–7. IEEE, 2018. 2
8. Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5039–5049. Computer Vision Foundation / IEEE, 2021. 2, 4, 10
9. Jean-François Aujol, Guy Gilboa, Tony F. Chan, and Stanley J. Osher. Structure-texture image decomposition - modeling, algorithms, and parameter selection. *Int. J. Comput. Vis.*, 67(1):111–136, 2006. 2, 5
10. Youngjung Kim, Bumsub Ham, Minh N Do, and Kwanghoon Sohn. Structure-texture image decomposition using deep variational priors. *IEEE Transactions on Image Processing*, 28(6):2692–2704, 2018. 2, 5
11. 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. 3
12. 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. 3
13. 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. 3

14. 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. 3
15. 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. 4
16. Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 735–743. AAAI Press, 2022. 4
17. 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. 4, 10, 11
18. Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Weiming Zhang, and Nenghai Yu. Self-supervised transformer for deepfake detection, 2022. 4
19. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 4
20. Juan Hu, Xin Liao, Difei Gao, Satoshi Tsutsui, Qian Wang, Zheng Qin, and Mike Zheng Shou. Mover: Mask and recovery based facial part consistency aware method for deepfake video detection, 2023. 4, 11
21. Liang Shi, Jie Zhang, and Shiguang Shan. Real face foundation representation learning for generalized deepfake detection, 2023. 4, 11
22. Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial examples: Towards good generalizations for deepfake detections. In *CVPR*, 2022. 4, 11
23. 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. 4, 5, 11, 12, 13
24. 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 (CVPR)*, pages 4129–4138, June 2023. 4, 5, 11
25. Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7135–7145, 2019. 4
26. Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021. 4

27. Haixu Song, Shiyu Huang, Yinpeng Dong, and Wei-Wei Tu. Robustness and generalizability of deepfake detection: A study with diffusion models. *arXiv preprint arXiv:2309.02218*, 2023. 4, 12, 13
28. Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. The deepfake detection challenge dataset. *CoRR*, abs/2006.07397, 2020. 5
29. Gereon Fox, Wentao Liu, Hyeongwoo Kim, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. Videoforensicshq: Detecting high-quality manipulated face videos. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 5
30. Jiajun Huang, Xueyu Wang, Bo Du, Pei Du, and Chang Xu. Deepfake mnist+: a deepfake facial animation dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1973–1982, 2021. 5
31. 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. 5, 11
32. Alexandru C. Telea. An image inpainting technique based on the fast marching method. *J. Graphics, GPU, & Game Tools*, 9(1):23–34, 2004. 6
33. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 8
34. 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. 9, 14
35. 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. 9
36. Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton-Ferrer. The deepfake detection challenge (DFDC) preview dataset. *CoRR*, abs/1910.08854, 2019. 9
37. Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2382–2390, 2020. 9
38. Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 15024–15034. IEEE, 2021. 10, 11
39. 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. 10
40. Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009. 10

41. 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 Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 10
42. Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 10
43. 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. 11
44. Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. *arXiv preprint arXiv:2307.07494*, 2023. 11
45. 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 (CVPR)*, pages 4113–4122, June 2022. 11, 12, 13
46. 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 (ECCV)*, 2022. 11, 12, 13
47. Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018. 12
48. Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228, 2021. 12
49. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 12
50. DeepInsight. insightface. https://github.com/deepinsight/insightface/, 2021. [Online;]. 12
51. Van Der Maaten Laurens and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2605):2579–2605, 2008. 14