Delocate: Detection and Localization for Deepfake Videos with Randomly-Located Tampered Traces

Juan Hu[1], **Xin Liao**[1]*, **Difei Gao**[2], **Satoshi Tsutsui**[3], **Qian Wang**[4], **Zheng Qin**[1], **Mike Zheng Shou**[2]

[1]College of Computer Science and Electronic Engineering, Hunan University, China
[2]Show Lab, National University of Singapore, Singapore
[3]Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, Singapore
[4]School of Cyber Science and Engineering, Wuhan University, China
{hujuan, xinliao, zqin}@hnu.edu.cn, difei.gao@vipl.ict.ac.cn, satoshi.tsutsui@ntu.edu.sg,
qianwang@whu.edu.cn, mike.zheng.shou@gmail.com

Abstract

Deepfake videos are becoming increasingly realistic, showing few tampering traces on facial areas that vary between frames. Consequently, existing Deepfake detection methods struggle to detect unknown domain Deepfake videos while accurately locating the tampered region. To address this limitation, we propose Delocate, a novel Deepfake detection model that can both recognize and localize unknown domain Deepfake videos. Our method consists of two stages named recovering and localization. In the recovering stage, the model randomly masks regions of interest (ROIs) and reconstructs real faces without tampering traces, leading to a relatively good recovery effect for real faces and a poor recovery effect for fake faces. In the localization stage, the output of the recovery phase and the forgery ground truth mask serve as supervision to guide the forgery localization process. This process strategically emphasizes the recovery phase of fake faces with poor recovery, facilitating the localization of tampered regions. Our extensive experiments on four widely used benchmark datasets demonstrate that Delocate not only excels in localizing tampered areas but also enhances cross-domain detection performance.

1 Introduction

Deepfakes, AI-generated videos of people, pose serious threats to society [Chesney and Citron, 2019; Wang *et al.*, 2022], emphasizing the need for *reliable* detection methods. By *reliable*, we believe the following three characteristics are necessary: First, the method should be robust to unseen forgery patterns (Fig. 3(a), [Rossler *et al.*, 2019; Zi *et al.*, 2020; Dolhansky *et al.*, 2020; Li *et al.*, 2020b]) with randomly located forgery traces (Fig. 3(b)), calling for cross-domain or cross-dataset evaluation. Second, a truly *reliable* method should convincingly explain the underlying reasons behind the model's decision by pointing to the manipulated part of a face. Unfortunately, we are not aware of works that satisfy all these criteria, so this paper develops a method that meets them.

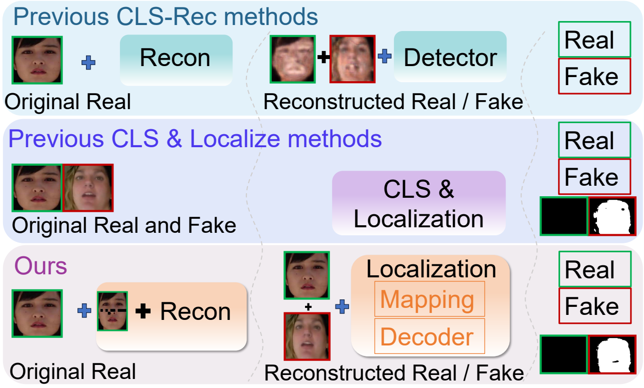


Figure 1: Diferences between ours and previous methods. Previous CLS-Rec methods mainly emphasize classification while overlooking localization aspects. Previous CLS & Localize methods leverage real and fake labels for feature extraction, without initially modeling real samples to extract robust features. Our method integrates both classification and localization, with a dedicated focus on real samples, enabling us to extract features for enhanced performance.

Recently, reconstruction-prediction-based methods have achieved relatively high detection performance. These methods typically involve the encoder and decoder that encode the input data into a low-dimensional representation and subsequently decode the original inputs from that representation. For example, [Khalid and Woo, 2020] uses reconstruction scores to classify real and fake videos. Moreover, reconstructing and predicting future frame representations [Hu *et al.*, 2022], forgery configurations [Chen *et al.*, 2022a], pseudo training samples [Chen *et al.*, 2022b], artifact representations [Dong *et al.*, 2022], the whole faces [Cao *et al.*, 2022; Shi *et al.*, 2023], the masked relation [Yang *et al.*, 2023], and mask regions [Chen *et al.*, 2023] can boost the detection performance. As illustrated in the first row of Fig. 1, since these methods typically train the reconstruction model

*Corresponding author.

solely with real images without specifically targeting any fake patterns, they have relatively good cross-domain performance. However, these methods overlook the importance of locating the forgery areas. Meanwhile, Deepfake localization methods[He *et al.*, 2021; Guo *et al.*, 2023; Kong *et al.*, 2022; Lai *et al.*, 2023; Huang *et al.*, 2022; Zhao *et al.*, 2023; Ţânţaru *et al.*, 2024] can locate forgery areas, but they learn the representation directly using real and fake videos, leading to a performance drop in detecting unseen types of fakes.

In this paper, we design a method that (1) can robust to unseen forgery patterns with randomly located forgery traces and (2) can locate the manipulated parts of faces. We name our method as Delocate, which, in essence, works as follows. The first stage, *Recovering for Consistency Learning* (Fig 2-top), pretrains a masked autoencoder using real faces only. Training on real faces ensures that the method does not overfit to any Deepfake patterns, enabling better generalization to unseen generation techniques, addressing point (1). Moreover, to detect tampered traces that appear randomly on a face, we design a unique masking strategy guided by facial parts. Subsequently, the masked autoencoder predicts the masked regions of interest (ROIs) based on the unmasked facial parts and interframes. This strengthens the understanding of relationships between facial parts and their temporal consistency, addressing point (1). Furthermore, the second stage, *Localization for Discrepancy Learning* (Fig 2-bottom), combines meta-learning with localization supervision to explicitly enhance cross-dataset generalization performance while simultaneously localizing the tampered regions of fake faces, addressing points (1) and (2).

Contributions. (1) We propose Delocate to learn representations guided by facial parts, enabling the detection of Deepfake videos in unknown domains.

(2) Unlike most detection methods that simply predict real or fake, Delocate can precisely localize tampered regions on faces. Learning to localize actually enhances the model's ability to detect fake videos.

(3) Extensive experiments on benchmark datasets, including FaceForensics++ (FF++) [Rossler *et al.*, 2019], Celeb-DF (CDF) [Li *et al.*, 2020b], DeeperForensics-1.0 (DFo) [Jiang *et al.*, 2020], DFDC [Dolhansky *et al.*, 2020] show that Delocate achieves effective performance under various metrics.

2 Related Work

Deepfake detection. Detection methods that focus on classifying real and fake videos can be broadly divided into two types: classification based on generalized methods (CLS-Gen) and classification based on reconstruction-prediction methods (CLS-Rec). The generalized methods contain methods based on implicit clues, explicit clues, and both implicit and explicit clues. Methods that explore implicit clues [Rossler *et al.*, 2019; Li *et al.*, 2020a; Sun *et al.*, 2021; Sabir *et al.*, 2019; Zhao *et al.*, 2021a; Sun *et al.*, 2022; Chen *et al.*, 2021; Zhao *et al.*, 2021b; Dong *et al.*, 2023] use supervised learning to distinguish genuine and fake videos without explicitly incorporating clues to detect Deepfake videos, making it challenging to understand the underlying detection clues. Methods that employ explicit clues [Li *et al.*, 2018; Yang *et al.*, 2019; Mittal *et al.*, 2020; Luo *et al.*, 2021; Nadimpalli and Rattani, 2022; Haliassos *et al.*, 2021; Hu *et al.*, 2022; Chai *et al.*, 2020; Zheng *et al.*, 2021; Guan *et al.*, 2022; Hu *et al.*, 2022; Shiohara and Yamasaki, 2022; Wang and Chow, 2023] have achieved more promising performance. Furthermore, Huang et al. [Huang *et al.*, 2023] explore explicit and implicit embeddings for Deepfake detection. However, given the rapid advancement of Deepfake technology, various falsification traces can be left behind, rendering detection methods that rely on explicit features vulnerable to attack.

The reconstruction-prediction-based methods [Khalid and Woo, 2020; Hu *et al.*, 2022; Chen *et al.*, 2022a; Chen *et al.*, 2022b; Dong *et al.*, 2022; Cao *et al.*, 2022; Yang *et al.*, 2023; Chen *et al.*, 2023; Shi *et al.*, 2023] are explained in Sec. 1. Though these methods achieve promising detection performance, they do not focus on forgery localization.

Deepfake localization. Kindly note that while there is a vast array of research papers on image localization, our discussion here is specifically focused on papers related to Deepfake classification and localization (CLS & Localize). There are few works that focus on Deepfake localization. Kong et al. [Kong *et al.*, 2022] use the noise map and semantic map to predict the forgery regions. Lai et al. [Lai *et al.*, 2023] use the mask decoder to locate forgery areas and classify videos. Zhao et al. [Zhao *et al.*, 2023] proposed RGB-Noise correlation to obtain the predicted manipulation regions. A recent paper [Shuai *et al.*, 2023] proposes a two-stream network for improving detection performance. These methods push the Deepfake forensic one step further to forgery localization, but they struggle to classify cross-domain Deepfake videos.

3 Method

This section presents the details of Delocate for Deepfake video detection. Specifically, the proposed method is composed of two stages: (1) Recovering for Consistency Learning, and (2) Localization for Discrepancy Learning stage, as shown in Fig. 2. We demonstrate the logic design in Algorithm 1.

Notations. Let A^{or}, A^{of}, A^{rr}, A^{rf}, A^{mr}, A^{mf}, A^{olr}, A^{olf}, A^{plr}, A^{plf} be original real faces, original fake faces, recovered real faces, recovered fake faces, masked real faces, masked fake faces, original real face localization, original fake face localization, predicted real face localization, and predicted fake face localization.

3.1 Recovering for Consistency Learning

In this stage, we perform self-supervised learning of real faces to learn generic facial part consistency features. As a result, the unspecific inconsistencies of fake faces with randomly-located tampered traces are exposed. Furthermore, we finetune the model with real faces and fake faces.

Masking strategy tailored to learn the consistent face representation. We design a facial part masking strategy to ensure that the model can learn the consistencies of all facial parts. The designed facial part masking strategy is different from the frame masking strategy of VideoMAE [Tong *et al.*, 2022].

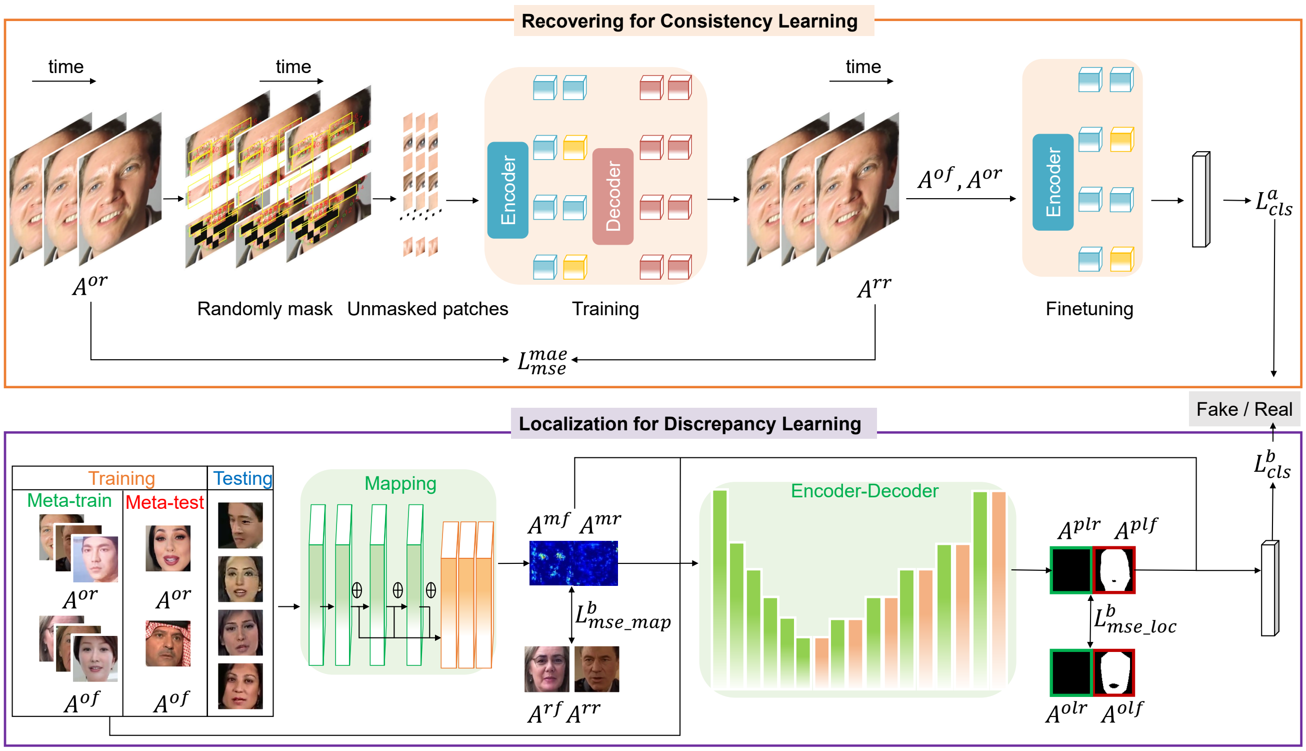
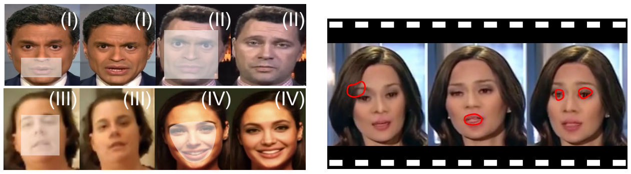


Figure 2: Pipeline of the proposed Delocate. In the Recovering stage, Delocate learns unspecific features by developing the designed masking strategy and recovery process. In the Localization stage, Delocate leverages devised mapping module and encoder-decoder module to maximize the discrepancy between real videos and Deepfake videos and locate the forgery areas.

First, as shown in Fig. 3(b), the tampered traces may only be sporadically present in one part and not related to other facial parts. Hence, we devise the masking strategy by considering Deepfake's domain knowledge. Specifically, we split the faces into different facial parts, i.e., eyes, cheek & nose, and lips, enabling the model to focus on both local and global consistencies among all facial parts. We choose region-specific masking strategy instead of a haphazard approach because random masks can fail to maintain the crucial global consistency among various facial regions. Neglecting such global facial part consistency could impede the model's ability to learn accurate facial part consistency features, making it challenging to distinguish real from fake videos based on reconstructed faces.

Second, the original masking strategy of VideoMAE [Tong *et al.*, 2022], with a high masking ratio, would make it too challenging to restore the original appearance without any artifacts or distortions. If reconstruction artifacts occur, real faces will contain them, and fake faces will display both reconstructed artifacts and tampering artifacts. This makes it difficult to distinguish real videos from fake videos since both have artifacts. Therefore, we propose a masking strategy that focuses on ROIs and utilizes a relatively low masking ratio to enable the model to reconstruct the original faces more accurately.

The ROIs extraction is partially inspired by Facial Action Coding System (FACS) [Friesen and Ekman, 1978], which considers the action units of FACS as fundamental elements. Drawing from psychology studies [Friesen and Ekman, 1978; Wang *et al.*, 2014; Wang *et al.*, 2015; Li and Deng, 2020; Russell and Fernandez-Dols, 1997], it is well-known that real faces exhibit inherent consistency in these elements. Consequently, when we mask ROIs, it becomes more challenging to reconstruct these regions for fake faces compared to real faces.



(a) Different forgery patterns (b) Random forgery traces

Figure 3: The significance of the randomly-located traces. Different forgery patterns employ different shapes to alter the face area, rendering random tampered traces across different frames, which cannot be predicted based on the current frame, resulting in strong unpredictability. (I) Face2Face in FF++. (II) FSGAN in DFDC (III) DeepFakes in FF++. (IV) Deepfake in Celeb-DF.

We reference the action units of eyebrows, lower eyelid, nose root, cheeks, mouth corner, side of the chin, and chin to calculate bounding box coordinates according to facial key-

Algorithm 1: The algorithm process of Delocate.

```
Input:
The original real faces $A^{or}$, the original fake faces $A^{of}$. The original localization $A^{ol}$ =$\{A^{olr}$,$A^{olf}\}$. The batch size $bs = 8$. The number of iterations $num\_iter$. The learning rate $\alpha$.
Output:
Trained model in recovery $\theta_{Rec}^{Ted}$, finetuning $\theta_{Fin}^{Ted}$, and localization $\theta_{Loc}^{Ted}$ process.

while (Recovery process) $\theta_{Rec}$ have not converged do
    for $i = 1 \rightarrow num\_iter^{Rec}$ do
        $A^{rr} = \theta_{Rec}(A^{or})$
        $\theta_{Rec}^{grad} \leftarrow \nabla_{\theta_{Rec}}(\frac{1}{b}\sum_{i=1}^{i=bs} L_{mse}^{mae}(A^{or}, A^{rr}))$
        $\theta_{Rec}^{Ted} \leftarrow \theta_{Rec} - \alpha_{Rec} \cdot AdamW(\theta_{Rec}, \theta_{Rec}^{grad})$
    end for
end while
while (Finetunig process) $\theta_{Fin}$ have not converged do
    for $i = 1 \rightarrow num\_iter^{Fin}$ do
        $p_i^{A^o\_Fin} = \theta_{Fin}(A^{or}, A^{of})$
        $\theta_{Fin}^{grad} \leftarrow \nabla_{\theta_{Fin}}(\frac{1}{bs}\sum_{i=1}^{i=bs} L_{cls}^{a}(p_i^{A^o\_Fin}, y_i^{A^o\_Fin}))$
        $\theta_{Fin}^{Ted} \leftarrow \theta_{Fin} - \alpha_{Fin} \cdot AdamW(\theta_{Fin}, \theta_{Fin}^{grad})$
    end for
end while
while (Localization process) $\theta_{Loc}$ have not converged do
    for $i = 1 \rightarrow num\_iter^{Loc}$ do
        $A^{r} = \{A^{rr}, A^{rf}\} = \theta_{Rec}^{Ted}(A^{or}, A^{of})$
        $A^{m} = \{A^{mr}, A^{mf}\} = \theta_{Loc}^{map}(A^{or}, A^{of})$
        $p_i^{A^o\_Loc}, A^{pl} = \theta_{Loc}^{Cls}(A^{or}, A^{of}, A^{ol})$
        $\theta_{Loc}^{map^{grad}} \leftarrow \nabla_{\theta_{Loc}}(\frac{1}{bs}\sum_{i=1}^{i=bs} L_{mse\_map}^{b}(A^{m}, A^{r}))$
        $\theta_{Loc}^{Cls^{grad}} \leftarrow$
            $\nabla_{\theta_{Loc}}(\frac{1}{bs}\sum_{i=1}^{i=bs} L_{cls}^{b}(p_i^{A^o\_Loc}, y_i^{A^o\_Loc}))$
        $\theta_{Loc}^{grad} \leftarrow \nabla_{\theta_{Loc}}(\frac{1}{bs}\sum_{i=1}^{i=bs} L_{mse\_loc}^{b}(A^{ol}, A^{pl}))$
        $\theta_{Loc}^{Meta^{grad}} \leftarrow \nabla_{\theta_{Loc}}(\frac{1}{bs}\sum_{i=1}^{i=bs} L^{b})$
        $\theta_{Loc}^{Ted} \leftarrow \theta_{Loc} - \alpha_{Loc} \cdot$
            $SGD\ (\theta_{Loc}, \theta_{Loc}^{map^{grad}}, \theta_{Loc}^{Cls^{grad}}, \theta_{Loc}^{grad}, \theta_{Loc}^{Meta^{grad}})$
    end for
end while
```

points. We discuss more about the masking strategy in the ablation study.

Network architecture. Our masked autoencoder is based on an asymmetric encoder-decoder architecture [He *et al.*, 2022]. To consider temporal correlation, the vanilla Vision Transformers (ViT) and joint space-time attention [Tong *et al.*, 2022] are adopted for recovering.

Recover masked faces. The masked patches of faces are dropped in the processing of the encoder, leaving the unmasked areas. In this way, the decoder predicts the missing facial part based on the unmasked areas. The reconstruction quality of masked patches is calculated with the MSE loss function L_{mse}^{mae}.

$$L_{mse}^{mae} = \frac{1}{n}\sum_{i=1}^{n}(A_i^{or} - A_i^{rr})^2. \tag{1}$$

If the model learns consistencies among facial parts, the loss between the reconstructed patches and the input patches should decrease. Our facial part masking strategy makes each part selected randomly, which enforces the model to learn the representation unspecific to any facial part. Furthermore, since this phase solely utilizes authentic videos and excludes any Deepfake content, it helps prevent the model from overfitting to particular Deepfake tampering patterns. In this way, the pretrained recovery model is obtained. Let $A^{rr} = RE^{r} \times A^{or}$, subject to $0 < RE^{r} < 1$, where RE^{r} represents the recovery quality of A^{or}. A higher score of RE^{r} indicates better reconstruction quality.

Finetuning the recovery model. We discard the decoder and apply the encoder to uncorrupted A^{or} and A^{of} for finetuning. The finetuning process uses a cross-entropy loss for detection.

$$L_{cls}^{a} = \frac{-1}{N}\sum_{i=1}^{N}\left[y_i^{A^o}\log p_i^{A^o} + (1 - y_i^{A^o})\log(1 - p_i^{A^o})\right], \tag{2}$$

where $p_i^{A^o}$ is the predicted label of original faces, $y_i^{A^o}$ is the groud truth label of original faces. Since the recovery model learns the facial part consistency of real videos, the well-trained encoder can extract the consistency features of real videos. For fake videos, as shown in Fig. 3, they are generated by different forgery patterns and tampered with different areas, and the tampered traces can show up in random regions. Consequently, the tampered traces can not be predicted. If the masked areas contain tampered traces, the recovery process would be affected. If there are no tampered traces in the masked area, the tampered traces in unmasked areas can not be recovered. That is, regardless of whether the tampering traces are covered, the video with randomly-located tampering traces will influence the recovery process, which makes the features extracted from the encoder different from those of the original videos.

3.2 Localization for Discrepancy Learning

In this stage, we leverage the well-trained recovery model from the first stage and map the recovery result to enlarge the discrepancy between real and fake videos.

Input data. We load the trained recovery model to obtain A^{or}, and input A^{of}, A^{or}, A^{rf}, and A^{rr} into the Localization stage. Let $A^{rf} = RE^{f} \times A^{of}$, where RE^{f} represents the recovery quality of A^{of}.

Data split strategy. To avoid over-fitting to specific Deepfake patterns, we use meta-learning [Jia *et al.*, 2021] and randomly divide the training data into Meta-train set and Meta-test set, where fake faces in Meta-train and Meta-test have different manipulated patterns.

Network architecture. We utilize the first convolutional layer of ResNet-18 [He *et al.*, 2016]. Instead of directly utilizing ResNet-18, we employ the first three residual blocks of ResNet-18 and concat the outputs of these residual blocks. Second, the concatenated outputs are fed into three convolutional layers for face mapping. The dimensions of the mapped faces are $56 \times 56 \times 3$. In this way, A^{mr} and A^{mf} can be represented as, $A^{mr} = MA^{r} \times A^{rr}$, subject to $0 < MA^{r} < 1$, $A^{mf} = MA^{r} \times A^{rf}$, subject to $0 < MA^{f} < 1$, where MA^{r} and MA^{f} represent the mapping quality of A^{rr} and A^{rf}. Third, the extracted mapping features are leveraged for classification purposes. Lastly, these features, alongside the original faces and localization labels, are fed into an encoder-decoder framework.

The encoder incorporates the SENet architecture [Hu *et al.*, 2018], while the decoder adopts the UNet framework [Ronneberger *et al.*, 2015]. To enhance the network's focus on pivotal regions, the SCSE Module [Wu *et al.*, 2022] is integrated into the decoder. The classification outcomes derived from the mapping features are governed by the constraint $L^{b}_{mse\_map}$, establishing a link with the encoder-decoder's localization component. This localization module is similarly regulated by $L^{b}_{mse\_loc}$. Both the mapping and localization results collectively contribute to the overall classification constraint L^{b}_{cls}. Instead of focusing on one task of classification and localization, our classification and localization results are mutually constrained and mutually promoted to facilitate us to complete multi-tasks of classification and localization.

Detection loss. To amplify the differences between A^{or} and A^{of}, we should satisfy:

$$A^{mr} - A^{mf} \gg A^{or} - A^{of}. \quad (3)$$

Combine the analyses of the A^{mr}, A^{mf}, A^{or}, and A^{of}, Eq. (3) can be represented as:

$$A^{rr}(MA^{r} - \frac{1}{RE^{r}}) \gg A^{rf}(MA^{f} - \frac{1}{RE^{f}}). \quad (4)$$

Since the recovery model is trained on real data A^{or} and the randomly-located traces of A^{of} could influence the recovery process, we have $A^{rr} > A^{rf}$. Moreover, the recovery quality of A^{of} can be smaller than that of A^{or}. That is, $0 < RE^{f} < RE^{r} < 1$. To satisfy Eq. (4), it is necessary to ensure that $MA^{r} \gg MA^{f}$. Therefore, we minimize the MSE loss $L^{b}_{mse\_map}$ between the mapped faces and the reconstructed faces. Consequently, $L^{b}_{mse\_map}$ allows the A^{mr} to be constrained by the consistency of A^{rr}, while the A^{rf} are constrained by the inconsistency. In this way, the model is able to recover A^{rr} but fails to recover A^{rf}, ensuring that $MA^{r} \gg MA^{f}$. These discrepancies enable the model to detect the failed reconstructed faces and better locate the tampered areas without misjudging the real faces.

For each pixel value in predicted localization masks A^{plr}_{pix}, we normalize it and process it as follows.

$$A^{plr}_{pix} = \begin{cases} 1, & \text{if } A^{plr}_{pix} \geq 0.5 \\ 0, & \text{if } A^{plr}_{pix} < 0.5 \end{cases}. \quad (5)$$

The primary objective of localization is to minimize the MSE loss $L^{b}_{mse\_loc}$ between the ground truth localization mask and predicted localization mask.

Moreover, we also minimize the binary cross-entropy L^{b}_{cls} between the video labels and the combined outputs of the mapping and localization features.

For each epoch, a sample batch is formed with the same number of fake videos and real videos to construct the binary detection task. To simulate unknown domain detection during training, the Meta-train phase performs training by sampling many detection tasks, and is validated by sampling many similar detection tasks from the Meta-test. Thereafter, the parameters of Meta-train phase can be updated. The goal of Meta-test phase is to enforce a classifier that performs well on Meta-train and can quickly generalize to the unseen domains of Meta-test, so as to improve the cross-domain detection performance.

The final loss function of the Localization stage is:

$$\begin{aligned} L^{b} = {} & (L^{b}_{cls} + L^{b}_{mse\_map} + L^{b}_{mse\_loc})_{Meta^{train}} + \\ & (L^{b}_{cls} + L^{b}_{mse\_map} + L^{b}_{mse\_loc})_{Meta^{train}})_{Meta^{test}}. \end{aligned} \quad (6)$$

which combines the Meta-test loss of L^{b}_{cls}, $L^{b}_{mse\_map}$, and $L^{b}_{mse\_loc}$ and Meta-train loss of L^{b}_{cls}, $L^{b}_{mse\_map}$, and $L^{b}_{mse\_loc}$ to achieve joint optimization.

Detection results. We average the output of Recovering stage and Localization stage to get the final detection score.

4 Experiment

4.1 Experimental Setup

Datasets. Four public Deepfake video datasets, i.e., FF++ [Rossler *et al.*, 2019], CDF [Li *et al.*, 2020b], DFo [Jiang *et al.*, 2020], DFDC [Dolhansky *et al.*, 2020] are utilized to evaluate the proposed method and existing methods. FF++ is made up of 4 types manipulated algorithms: DeepFakes (DF) [DeepFakes, 2018], Face2Face (F2F) [Thies *et al.*, 2018], FaceSwap (FS) [FaceSwap, 2018], NeuralTextures (NT) [Thies *et al.*, 2019]. Moreover, 4000 videos are synthesized based on the 4 algorithms. These videos are widely used in various Deepfake detection scenarios. Celeb-DF contains 5639 videos that are generated by an improved DeepFakes algorithm [Li *et al.*, 2020b]. The tampered traces in some inchoate datasets are relieved in Celeb-DF. DeeperForensics-1.0 dataset is published for real-world Deepfake detection. DFDC is a large-scale Deepfake detection dataset published by Facebook.

Implementation details. In the Recovering stage, the masking ratio, batch size, patch size, and input size are set as 0.75, 8, 16, 224, respectively. The AdamW [Loshchilov and Hutter, 2017] optimizer with an initial learning rate 1.5×10^{-4}, momentum of 0.9 and a weight decay 0.05 is utilized to train the recovery model. The finetuning of the Recovering stage utilizes the AdamW optimizer with an initial learning rate 1×10^{-3} to detect videos. The SGD optimizer is used for optimizing the Localization stage with the initial learning rate 0.1, momentum of 0.9, and weight decay of 5×10^{-4}. We use `FFmpege` [Lei *et al.*, 2013] to extract 30 frames from each video. The `dlib` [Sharma *et al.*, 2016] is utilized to extract faces and detect 68 facial landmarks. We follow Kong et al. [Kong *et al.*, 2022] to extract the ground truth of forgery localization.

Comparison methods. We compare Delocate with the CLS-Gen methods that are representative of implicit methods, explicit methods, and explicit and implicit combined methods, i.e., MultiAtt [Zhao *et al.*, 2021a], LipForensics [Haliassos *et al.*, 2021], Huang et al. [Huang *et al.*, 2023]. We also compare Delocate with CLS-Rec methods, i.e., OST [Chen *et al.*, 2022b], RECCE [Cao *et al.*, 2022], MRL [Yang *et al.*, 2023], and DisGRL [Shi *et al.*, 2023]. Furthermore,

Method	CDF		DFo		DFDC	
	AUC ↑	EER↓	AUC ↑	EER↓	AUC ↑	EER↓
MultiAtt	76.7	32.8	72.4	34.7	67.3	38.3
LipForensics	82.4	24.2	97.6	10.6	73.5	36.5
Huang et al.	83.8	24.9	90.8	15.3	81.2	26.8
OST	74.8	31.2	95.1	9.7	83.3	25.0
RECCE	73.7	30.3	89.3	16.9	74.0	31.1
MRL	86.7	18.3	91.1	15.6	74.5	30.1
DisGRL	76.7	28.3	88.4	18.5	74.8	30.0
Kong et al.	70.7	35.5	82.6	24.7	63.3	40.8
Zhao et al.	74.8	30.0	80.9	25.8	79.0	26.1
Chao et al.	86.2	18.1	99.0	7.6	82.5	25.1
Delocate	**91.3**	**14.1**	**99.1**	**6.6**	**84.0**	**24.7**

Table 1: Comparisons of detection performance (AUC (%) and EER (%)) between Delocate and other methods on CDF, DFo, and DFDC datasets when trained on 4 types of videos of FF++.

we compare the CLS & Localize methods, i.e., Kong et al. [Kong *et al.*, 2022], Zhao et al. [Zhao *et al.*, 2023], Chao et al. [Shuai *et al.*, 2023] .

4.2 Generalization to Unknown Domains

We enforce Delocate to learn unspecific features for Deepfake video detection with randomly-located tampered traces. The unknown domain detection is precisely the scenario where tampered traces are often randomly-located. To test the performance of Delocate, we simulate unknown domain Deepfake detection in multiple scenarios.

Comparisons of classification. First, we conduct experiments by training the model on FF++ with all 4 types of videos, but testing on other datasets, i.e., CDF, DFo, DFDC, and we use Area Under Curve (AUC) and Equal Error Rate (EER) to evaluate the classification performance. The enormous differences between the training domain and the testing domain make it challenging to improve unknown domain detection performance. Nonetheless, the results in Table 1 show that Delocate manages to improve the classification performance and achieve comparable localization performance at the same time. For example, Delocate improves the AUC on CDF from 86.2% (the localization method: Chao et al. [Shuai *et al.*, 2023]) to 91.3%.

Second, to avoid performing experiments on a particular training mode, we change the training mode and conduct other unknown domain detection experiments. Specifically, we implement experiments by selecting one type of FF++ for training, but testing on other datasets, i.e., CDF, DFo, DFDC. Since there is only one type of video for training in experiments, we randomly split the training data into Meta-train and Meta-test with $7 : 3$. Results in Table 2 illustrate that Delocate outperforms previous methods in many scenarios. Compared with classification methods, OST [Chen *et al.*, 2022b] performs better than Delocate in 3 scenarios. Despite these results, it is worth noting that Delocate achieves better classification performance, especially with a 2.4% improvement over OST [Chen *et al.*, 2022b] when training on FS and testing on CDF. We also observe that Delocate performs better AUC performance than that of localization methods [Kong *et al.*, 2022; Zhao *et al.*, 2023; Shuai *et al.*, 2023]. For instance, when training on DF and testing on CDF, Delocate achieves a 5.8% AUC improvement over Chao et al. [Shuai *et al.*, 2023].

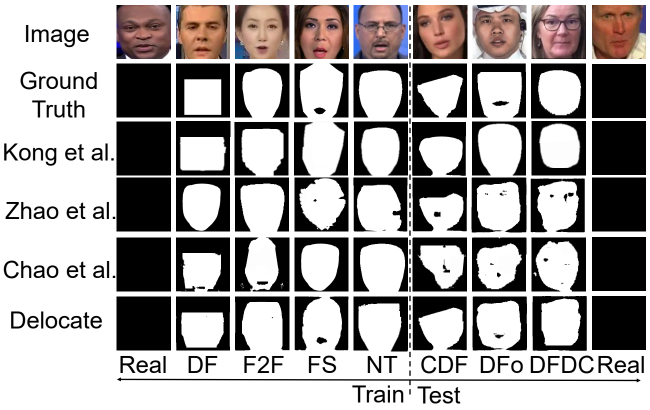


Figure 4: Comparisons of predicted forgery regions on CDF, DFo, and DFDC datasets when trained on 4 types of videos of FF++.

Comparisons of localization. We use Intersection over Union (IoU) and Pixel-wise Binary Classification Accuracy (PBCA) [Kong *et al.*, 2022] to evaluate the localization performance. We train the model on FF++ and test it in other datasets. Table 3 shows that Chao et al. [Shuai *et al.*, 2023] the best IoU results in testing DFDC. Delocate performs best results in other scenarios.

We also conduct forgery localization analyses for the CLS & Localize methods and show the results in Fig. 4. It shows that the localization area identified by Kong et al. [Kong *et al.*, 2022], Zhao et al. [Zhao *et al.*, 2023] and Chao et al. [Shuai *et al.*, 2023] exhibits sporadic mismatches across various regions when compared to the ground truth. For the CDF, DFo, and DFDC datasets, Delocate aligns more closely with the ground truth region compared to the area localized by Zhao et al. [Zhao *et al.*, 2023] and Chao et al. [Shuai *et al.*, 2023]. It may be because Delocate focuses on unspecific features during the reconstruction stage, thereby revealing inconsistencies in the synthetic faces. In the localization stage, it maps the outcomes of the reconstruction, where the classification and localization results mutually influence and enhance each other. This process leads to the extraction of more generalized features, consequently improving the cross-domain performance.

4.3 Intra-dataset Detection Performance

To provide a comprehensive assessment of the proposed Delocate, we compare Delocate with the state-of-the-art methods in the scenario of intra-dataset detection. Specifically, we conduct experiments on 4 subsets of FF++ (C23). The training data and testing data of intra-dataset experiments are from the same subset of FF++. Table 4 shows that most methods perform well in intra-dataset detection. Chao et al. [Shuai *et al.*, 2023] achieves the highest intra-dataset detection score while Delocate has a slight decrease of 0.2% in average accuracy compared. This drop may be due to the fact that the model improves the unknown domain performance while sac-

Method	DF			F2F			FS			NT		
	CDF	DFo	DFDC	CDF	DFo	DFDC	CDF	DFo	DFDC	CDF	DFo	DFDC
MultiAtt	68.7	80.6	70.1	69.6	81.9	68.6	70.4	82.5	70.1	70.2	82.9	66.9
LipForensics	69.3	90.1	70.8	69.1	72.4	71.4	72.3	71.9	71.8	70.9	73.2	69.8
Huang et al.	72.9	90.9	72.8	74.2	91.2	75.8	72.7	89.9	71.9	74.8	91.3	73.5
OST	76.6	93.8	75.7	79.9	**94.7**	79.8	79.2	90.9	80.2	75.3	**92.9**	75.2
RECCE	69.7	78.0	68.0	70.5	75.7	71.1	69.7	73.3	71.1	70.1	74.5	70.2
MRL	72.9	79.3	72.2	70.6	79.5	71.2	73.1	84.2	70.5	71.4	82.1	72.4
DisGRL	71.5	79.2	70.2	70.3	78.9	72.0	73.3	82.9	71.0	72.8	83.7	72.3
Kong et al.	69.3	80.8	62.6	68.4	79.4	62.1	69.2	79.2	62.9	70.1	79.2	62.3
Zhao et al.	71.2	79.8	76.2	70.4	79.6	76.1	73.0	79.0	75.9	71.8	79.2	74.6
Chao et al.	72.4	89.1	75.0	79.7	90.6	76.2	80.4	90.5	80.1	75.4	91.6	72.3
Delocate	**78.2**	**94.5**	**76.3**	**80.9**	93.3	**79.9**	**81.6**	**91.5**	**80.8**	**76.8**	90.9	**75.9**

Table 2: Comparisons of the detection performance (AUC (%)) between Delocate and other methods on CDF, DFo, and DFDC datasets when trained on one type of videos of FF++.

Method	CDF		DFo		DFDC	
	IoU ↑	PBCA↑	IoU ↑	PBCA↑	IoU ↑	PBCA↑
Kong et al.	0.709	0.721	0.843	0.826	0.616	0.624
Zhao et al.	0.789	0.767	0.904	0.905	0.708	0.706
Chao et al.	0.798	0.784	0.921	0.919	**0.741**	0.726
Delocate	**0.801**	**0.802**	**0.937**	**0.926**	0.738	**0.727**

Table 3: Comparisons of localization performance (IoU and PBCA) between Delocate and localization methods on CDF, DFo, and DFDC datasets when trained on 4 types of videos of FF++.

Method	DF	FS	F2F	NT
MultiAtt	99.6	**100**	99.3	98.3
LipForensics	99.8	**100**	99.3	**99.7**
Huang et al.	99.6	99.8	99.5	98.4
OST	99.0	98.8	99.1	95.9
RECCE	99.7	99.9	99.2	98.4
MRL	99.2	98.1	97.3	98.6
DisGR	99.0	99.1	98.3	99.6
Kong et al.	99.7	99.6	99.4	98.9
Zhao et al.	99.8	99.4	99.0	97.9
Chao et al.	**100**	**100**	**99.9**	99.4
Delocate	99.8	99.7	99.7	99.3

Table 4: Comparisons of the Intra-dataset evaluation (AUC (%)) between Delocate and other methods.

Mask ratio	CDF	DFo	DFDC
55%	89.0	94.2	80.8
65%	90.6	92.6	81.8
75%	**91.3**	**99.1**	**84.0**
85%	90.3	92.8	81.9
95%	89.9	92.7	81.0

Table 5: Ablation study - The detection performance (AUC (%)) of different masking ratios on testing datasets after training on FF++.

rificing a little bit of intra-domain performance to fit the unseen domain.

Masking strategy	CDF	DFo	DFDC
MAE masking	86.4	95.8	79.1
VideoMAE masking	86.5	95.6	79.5
Eye	91.1	98.7	80.1
cheek & nose	90.2	88.2	81.3
Lip	90.8	88.2	81.7
w/o ROIs	90.9	88.9	83.5
Proposed strategy	**91.3**	**99.1**	**84.0**

Table 6: Ablation study - The detection performance (AUC (%)) of different mask strategies on the testing datasets after training on FF++.

4.4 Ablation Study

We conduct ablation study experiments by training on FF++ but testing on CDF, DFo, and DFDC datasets.

Influence of the masking ratio. We trained models on the FF++ dataset with different masking ratios. Note that instead of defining the masking ratio as the ratio of masked area to the entire face, we define the masking ratio as the ratio of masked area to the corresponding ROIs facial parts. We choose not to use the original definition of mask ratio, which measures the ratio of the mask area to the entire face. Instead, we focus on specific regions of interest (ROIs) and divide the face into three parts. Then, we randomly mask only one part at a time. Our attention is directed towards the specific masked ROIs during the masking procedure, rather than considering the entire face as a whole.

In Table 5, we observe that Delocate scales well with the masking ratio of 75%. The performance gets a slight drop in the masking ratio of 55% and 65% indicating that low masking ratios may hinder learning robust features. When the mask rate is 85% and 95%, the detection performance is also degraded. That may be because that high masking ratio can raise the difficulty of reconstructing faces. If both real faces and fake faces are not reconstructed well, the distinction between them can be reduced. Therefore, we set the masking ratio as 75% in the experiments.

Influence of the masking strategy. We modify the masking strategies of MAE [He *et al.*, 2022] to improve the general-

	CDF	DFo	DFDC
MAE	76.4	88.2	71.1
VideoMAE	77.4	89.3	71.8
w/o Recovering stage	89.0	96.5	80.9
w/o Localization stage	85.8	95.4	80.0
w/o Meta-learning	89.6	96.1	81.4
w/o Mapping	88.2	95.3	80.9
w/o Encoder-Decoder	89.9	96.7	82.8
MAE + Localization stage	82.8	92.9	75.1
VideoMAE + Localization stage	83.7	93.2	75.8
RECCE + Localization stage	81.8	92.4	76.2
Delocate	**91.3**	**99.1**	**84.0**

Table 7: Ablation study - Effects of MAE, VideoMAE, Recovering stage, Localization stage, Meta-learning, Mapping and Encoder-Decoder.

ization. To evaluate the effectiveness of the improved masking strategy, we compare the proposed masking strategy with masking strategies of MAE and VideoMAE. Furthermore, since the modified strategy randomly selects parts to mask, evaluating the effects of different masked parts is important. To analyze the effectiveness of the ROIs, we compare the proposed strategy with the masking strategy that does not focus on ROIs. We trained models on the FF++ dataset with different masking strategies.

The results of 1^{st}, 2^{nd}, and 7^{th} lines in Table 6 demonstrate that modifying the masking strategies of MAE [He *et al.*, 2022] and VideoMAE [Tong *et al.*, 2022] can improve the detection performance. The results in the 3^{rd}, 4^{th} and 5^{th} lines, which represent methods that mask eye areas, cheek and nose areas, and lip areas, respectively, show a performance degradation compared to the proposed strategy. That is, random masking a part of all facial parts is more conducive to extracting robust features than masking a certain part only. Moreover, the results of the 6^{th} line and 7^{th} lines show that the proposed masking strategy that focuses on ROIs achieves better performance than the masking strategy without ROIs. The reason is that the model can better capture the differences between real and fake videos by masking patches in these ROIs, as fake videos typically lack consistency. Therefore, the proposed masking strategy is effective in detecting Deepfake videos.

Influence of MAE and VideoMAE. We compare the detection performance of the Delocate with the original MAE and VideoMAE methods for Deepfake detection. The results are shown in the 1^{st} and 2^{nd} line of Table 7. The detection performance of the original MAE and VideoMAE is lower than that of Delocate, demonstrating the effectiveness of the modifications in Delocate.

Influence of Recovering stage and Localization stage. To validate the performance of each stage, we compare the performance of a single stage with that of both stages combined. The results are shown in the 3^{rd} and 4^{th} lines of Table 7. We can see that removing either the Recovering stage or the Localization stage degraded the detection performance, as each stage plays a crucial role in Deepfake detection. Combining both stages improves the performance by magnifying the distinction between real and fake videos.

Influence of Meta-learning, Mapping, and Encoder-Decoder. We remove the meta-learning, mapping, and Encoder-Decoder module to carry out experiments, respectively, and the results are shown in the 5^{th}, 6^{th}, 7^{th} line of Table 7. Compared with results of 11^{th} line, the method without meta-learning, mapping, and Encoder-Decoder module achieves worse results than the proposed Delocate with these modules. The meta-learning approach simulates cross-domain detection in the training phase, improving detection performance. The mapping module can reveal the inconsistencies by developing the autoencoder of the Recovering stage, which facilitates the Encoder-Decoder module to locate the forgery regions. The Encoder-Decoder module achieves the forgery localization, providing a guidance for the classification results.

5 Conclusion

This paper focuses on the detection and localization of Deepfakes, particularly in identifying Deepfake videos with randomly-located tampered traces. By focusing equally on all facial parts rather than relying on specific facial parts, our two-stage model can learn unspecific facial consistencies and general representations. In the Recovering stage, the model is trained to recover faces from partially masked ROIs on the face, which facilitates the model in learning the facial part consistencies of real videos. In the Localization stage, the model enforces a mapping and an encoder-decoder strategy to expose the forgery areas in synthetic ones. Extensive experiments illustrate the generalizability of Delocate in detection and localization.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant Nos. U22A2030, U20A20174), National Key R& D Program of China (Grant No. 2022YFB3103500), Hunan Provincial Funds for Distinguished Young Scholars (Grant No. 2024JJ2025). Juan Hu thanks the support of the Ministry of Education, Singapore, under its MOE AcRF TIER 3 (Grant MOE-MOET32022-0001). Satoshi Tsutsui is supported by the NTU-PKU Joint Research Institute (a collaboration between the Nanyang Technological University and Peking University that is sponsored by a donation from the Ng Teng Fong Charitable Foundation). Mike Zheng Shou does not receive any funding for this work.

References

[Cao *et al.*, 2022] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, pages 4113–4122, 2022.

[Chai *et al.*, 2020] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *ECCV*, pages 103–120, 2020.

[Chen *et al.*, 2021] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *AAAI*, volume 35, pages 1081–1088, 2021.

[Chen *et al.*, 2022a] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *CVPR*, pages 18710–18719, 2022.

[Chen *et al.*, 2022b] Liang Chen, Yong Zhang, Yibing Song, Jue Wang, and Lingqiao Liu. Ost: Improving generalization of deepfake detection via one-shot test-time training. In *NeurIPS*, volume 35, pages 24597–24610, 2022.

[Chen *et al.*, 2023] Han Chen, Yuzhen Lin, Bin Li, and Shunquan Tan. Learning features of intra-consistency and inter-diversity: Keys toward generalizable deepfake detection. *IEEE TCSVT*, 33(3):1468–1480, 2023.

[Chesney and Citron, 2019] Bobby Chesney and Danielle Citron. Deepfakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019.

[DeepFakes, 2018] DeepFakes. Accessed october 10, 2018. https://github.com/deepfakes/faceswap, 2018.

[Dolhansky *et al.*, 2020] 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.

[Dong *et al.*, 2022] Shichao Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. Explaining deepfake detection by analysing image matching. In *ECCV*, pages 18–35, 2022.

[Dong *et al.*, 2023] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *CVPR*, pages 3994–4004, 2023.

[FaceSwap, 2018] FaceSwap. Accessed october 29, 2018. https://github.com/MarekKowalski/FaceSwap/, 2018.

[Friesen and Ekman, 1978] E Friesen and Paul Ekman. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, 3(2):5, 1978.

[Guan *et al.*, 2022] Jiazhi Guan, Hang Zhou, Zhibin Hong, Errui Ding, Jingdong Wang, Chengbin Quan, and Youjian Zhao. Delving into sequential patches for deepfake detection. In *NeurIPS*, volume 35, pages 4517–4530, 2022.

[Guo *et al.*, 2023] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *CVPR*, pages 3155–3165, 2023.

[Haliassos *et al.*, 2021] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *CVPR*, pages 5039–5049, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2021] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *CVPR*, pages 4360–4369, 2021.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[Hu *et al.*, 2022] Juan Hu, Xin Liao, Jinwen Liang, Wenbo Zhou, and Zheng Qin. FInfer: Frame inference-based deepfake detection for high-visual-quality videos. In *AAAI*, volume 36, pages 951–959, 2022.

[Huang *et al.*, 2022] Yihao Huang, Felix Juefei-Xu, Qing Guo, Yang Liu, and Geguang Pu. Fakelocator: Robust localization of gan-based face manipulations. *IEEE TIFS*, 17:2657–2672, 2022.

[Huang *et al.*, 2023] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *CVPR*, pages 4490–4499, 2023.

[Jia *et al.*, 2021] Yunpei Jia, Jie Zhang, and Shiguang Shan. Dual-branch meta-learning network with distribution alignment for face anti-spoofing. *IEEE TIFS*, 17:138–151, 2021.

[Jiang *et al.*, 2020] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, pages 2889–2898, 2020.

[Khalid and Woo, 2020] Hasam Khalid and Simon S Woo. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In *CVPRW*, pages 656–657, 2020.

[Kong *et al.*, 2022] Chenqi Kong, Baoliang Chen, Haoliang Li, Shiqi Wang, Anderson Rocha, and Sam Kwong. Detect and locate: Exposing face manipulation by semantic-and noise-level telltales. *IEEE TIFS*, 17:1741–1756, 2022.

[Lai *et al.*, 2023] Yingxin Lai, Zhiming Luo, and Zitong Yu. Detect any deepfakes: Segment anything meets face forgery detection and localization. In *CCBR*, pages 180–190, 2023.

[Lei *et al.*, 2013] Xiaohua Lei, Xiuhua Jiang, and Caihong Wang. Design and implementation of a real-time video stream analysis system based on ffmpeg. In *WCSE*, pages 212–216, 2013.

[Li and Deng, 2020] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE TAC*, 13(3):1195–1215, 2020.

[Li *et al.*, 2018] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In *IEEE WIFS*, pages 1–7, 2018.

[Li *et al.*, 2020a] 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 *ACM MM*, pages 1864–1872, 2020.

[Li *et al.*, 2020b] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *CVPR*, pages 3207–3216, 2020.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Luo *et al.*, 2021] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, pages 16317–16326, 2021.

[Mittal *et al.*, 2020] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *ACM MM*, pages 2823–2832, 2020.

[Nadimpalli and Rattani, 2022] Aakash Varma Nadimpalli and Ajita Rattani. On improving cross-dataset generalization of deepfake detectors. In *CVPR*, pages 91–99, 2022.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[Rossler *et al.*, 2019] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019.

[Russell and Fernandez-Dols, 1997] James A Russell and José Miguel Fernandez-Dols. The psychology of facial expression. *Cambridge university press*, 1997.

[Sabir *et al.*, 2019] 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.

[Sharma *et al.*, 2016] S Sharma, Karthikeyan Shanmugasundaram, and Sathees Kumar Ramasamy. Farec-cnn based efficient face recognition technique using dlib. In *IEEE ICACCCT*, pages 192–195, 2016.

[Shi *et al.*, 2023] Zenan Shi, Haipeng Chen, Long Chen, and Dong Zhang. Discrepancy-guided reconstruction learning for image forgery detection. In *IJCAI*, pages 1387–1395, 2023.

[Shiohara and Yamasaki, 2022] Kaede Shiohara and Toshihiko Yamasaki. Detecting Deepfakes with self-blended images. In *CVPR*, pages 18720–18729, 2022.

[Shuai *et al.*, 2023] Chao Shuai, Jieming Zhong, Shuang Wu, Feng Lin, Zhibo Wang, Zhongjie Ba, Zhenguang Liu, Lorenzo Cavallaro, and Kui Ren. Locate and verify: A two-stream network for improved deepfake detection. In *ACM MM*, pages 7131–7142, 2023.

[Sun *et al.*, 2021] Ke Sun, Hong Liu, Qixiang Ye, Yue Gao, Jianzhuang Liu, Ling Shao, and Rongrong Ji. Domain general face forgery detection by learning to weight. In *AAAI*, volume 35, pages 2638–2646, 2021.

[Sun *et al.*, 2022] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *AAAI*, volume 36, pages 2316–2324, 2022.

[Thies *et al.*, 2018] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, pages 2387–2395, 2018.

[Thies *et al.*, 2019] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *IEEE TOG*, 38(4):1–12, 2019.

[Tong *et al.*, 2022] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, volume 35, pages 10078–10093, 2022.

[Wang and Chow, 2023] Tianyi Wang and Kam Pui Chow. Noise based deepfake detection via multi-head relative-interaction. In *AAAI*, volume 37, pages 14548–14556, 2023.

[Wang *et al.*, 2014] Su-Jing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, and Xiaolan Fu. Micro-expression recognition using dynamic textures on tensor independent color space. In *ICPR*, pages 4678–4683, 2014.

[Wang *et al.*, 2015] Su-Jing Wang, Wen-Jing Yan, Xiaobai Li, Guoying Zhao, Chun-Guang Zhou, Xiaolan Fu, Minghao Yang, and Jianhua Tao. Micro-expression recognition using color spaces. *IEEE TIP*, 24(12):6034–6047, 2015.

[Wang *et al.*, 2022] Zhi Wang, Yiwen Guo, and Wangmeng Zuo. Deepfake forensics via an adversarial game. *IEEE TIP*, 31:3541–3552, 2022.

[Wu *et al.*, 2022] Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. Robust image forgery detection over online social network shared images. In *CVPR*, pages 13440–13449, 2022.

[Yang *et al.*, 2019] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP*, pages 8261–8265, 2019.

[Yang *et al.*, 2023] Ziming Yang, Jian Liang, Yuting Xu, Xiao-Yu Zhang, and Ran He. Masked relation learning for deepfake detection. *IEEE TIFS*, 18:1696–1708, 2023.

[Zhao *et al.*, 2021a] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *CVPR*, pages 2185–2194, 2021.

[Zhao *et al.*, 2021b] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *ICCV*, pages 15023–15033, 2021.

[Zhao *et al.*, 2023] Hongjie Zhao, Beibei Liu, Yongjian Hu, Jicheng Li, and Chang-Tsun Li. Hybrid domain meta-learning network for face forgery detection and localization in deepfakes. In *IJCNN*, pages 1–8, 2023.

[Zheng *et al.*, 2021] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *ICCV*, pages 15044–15054, 2021.

[Zi *et al.*, 2020] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. WildDeepfake: A challenging real-world dataset for deepfake detection. In *ACM MM*, pages 2382–2390, 2020.

[Ţânţaru *et al.*, 2024] Dragoş-Constantin Ţânţaru, Elisabeta Oneaţă, and Dan Oneaţă. Weakly-supervised deepfake localization in diffusion-generated images. In *WACV*, pages 6258–6268, 2024.