Learning Natural Consistency Representation for Face Forgery Video Detection

Daichi Zhang[1,2,3], Zihao Xiao[4], Shikun Li[1,2], Fanzhao Lin[1,2], Jianmin Li[3], and Shiming Ge[1,2,⋆]

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] Department of Computer Science and Technology, Institute for AI, BNRist, Tsinghua University [4] RealAI
{zhangdaichi20,lishikun19,linfanzhao21}@mails.ucas.ac.cn, lijianmin@mail.tsinghua.edu.cn, zihao.xiao@realai.ai, geshiming@iie.ac.cn

Abstract. Face Forgery videos have elicited critical social public concerns and various detectors have been proposed. However, fully-supervised detectors may lead to easily overfitting to specific forgery methods or videos, and existing self-supervised detectors are strict on auxiliary tasks, such as requiring audio or multi-modalities, leading to limited generalization and robustness. In this paper, we examine whether we can address this issue by leveraging visual-only real face videos. To this end, we propose to learn the Natural Consistency representation (NACO) of real face videos in a self-supervised manner, which is inspired by the observation that fake videos struggle to maintain the natural spatiotemporal consistency even under unknown forgery methods and different perturbations. Our NACO first extracts spatial features of each frame by CNNs then integrates them into Transformer to learn the long-range spatiotemporal representation, leveraging the advantages of CNNs and Transformer on local spatial receptive field and long-term memory respectively. Furthermore, a Spatial Predictive Module (SPM) and a Temporal Contrastive Module (TCM) are introduced to enhance the natural consistency representation learning. The SPM aims to predict random masked spatial features from spatiotemporal representation, and the TCM regularizes the latent distance of spatiotemporal representation by shuffling the natural order to disturb the consistency, which could both force our NACO more sensitive to the natural spatiotemporal consistency. After the representation learning stage, a MLP head is fine-tuned to perform the usual forgery video classification task. Extensive experiments show that our method outperforms other state-of-the-art competitors with impressive generalization and robustness.

Keywords: Face forgery video detection · Natural consistency · Spatiotemporal representation · Self-supervised learning

⋆ Corresponding author

1 Introduction

Recently, there has been a significant advancement in face forgery technology [27, 28, 34, 48], particularly since the emergence of generative adversarial networks (GANs) [17]. The generated and manipulated faces are almost indistinguishable to the naked eye, but can be easily produced by available online tools like Deepfakes[1] and FaceSwap[2]. This accessibility enables perpetrators easily using these techniques to generate forgery videos to mislead the public, defame celebrities, or even fabricate evidence, which could result in severe social, political, and security threats [46]. Hence, how to develop effective face forgery video detectors is crucial to prevent the malicious applications of these techniques.

Various detectors have been proposed to address the face forgery video detection task. Existing supervised detectors mainly focuses on specific patterns which are discriminative in forgery and real videos, such as frequency domain information [39], specific artifacts [11,32,51], and spatiotemporal clues [1,22,31, 36,55,57,58,65]. However, these fully-supervised detectors may easily overfit to specific forgery methods or videos in training datasets and the artifacts they rely on may be corrupted under perturbations, leading to limited generalization and robustness. Recent FTCN [65] explores temporal coherence by setting spatial kernel size to one and AltFreezing [55] proposes to learn spatial and temporal features respectively to improve the generalization. But FTCN ignores all the clues from spatial dimension and AltFreezing neglects the connection between spatial and temporal domains, which may hinder their performance. Other self-supervised detectors aim to explore model-agnostic self-supervised representation to detect, such as LipForensics [19] pretrains on lipreading dataset focusing on mouth movements and the RealForensics [18] pretrains on visual-audio modalities to avoid overfitting. However, LipForensics requires labeled datasets for pretraining and only focuses on the mouth region, RealForensics requires both visual-audio modalities for pretraining and performs visual multi-task during finetuning, which are strict on the auxiliary tasks and the pretaining datasets that may limit their scalability and performance.

With all the concerns above in mind, we raise the question: whether we can learn a general and robust representation on visual-only real videos in a self-supervised manner by fully leveraging their spatiotemporal clues, without requiring any fake videos, thus avoiding overfitting to specific forgery methods or videos. To achieve this, we are inspired by the observation that fake videos struggle to preserve the natural spatiotemporal consistency in real face videos, defined as the semantic-level spatiotemporal coherence in natural face videos (the opposite as inconsistency), such as facial movements and expression changes. This high-level natural consistency is corrupted even under unknown forgery methods and different perturbations, providing the insight to leverage this as clue to improve generalization and robustness. As shown in Fig. 1 (a), the real videos exhibit natural spatiotemporal consistency while the fake videos generated by

[1] https://github.com/deepfakes/faceswap
[2] https://github.com/MarekKowalski/FaceSwap

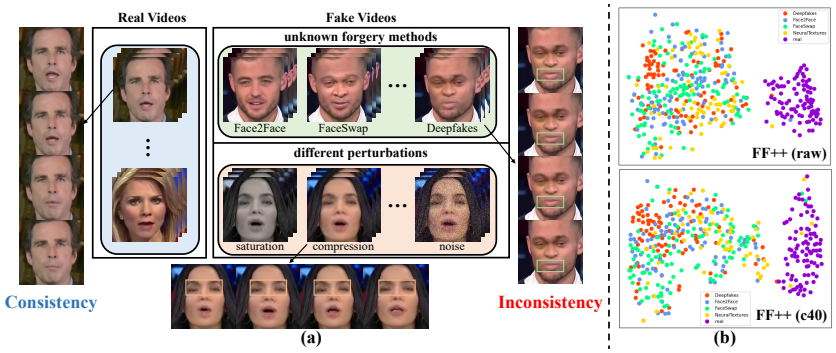

Fig. 1: (a) Real face videos exhibit natural spatiotemporal consistency while fake videos generated from unknown forgery methods or under different perturbations both show inconsistencies. (b) t-SNE [25] visualization of our NACO on uncompressed (raw) and heavily compressed (c40) FF++ which includes four different forgery methods.

unknown forgery methods or under different perturbations both show inconsistency. If we could leverage this natural consistency of real face videos, we could embed all real face videos into one compact cluster in latent space while all other fake videos (from unknown generation methods or different perturbations) are embedded into another cluster with a clear discrepancy margin, such as shown in Fig. 1 (b), achieving general and robust face forgery video detection.

To this end, we propose to learn the Natural Consistency (NACO) representation of visual-only real face videos in a self-supervised manner. We first model the spatiotemporal representation of videos by initially extracting spatial features from each frame by CNNs then integrating the spatial feature sequence into Transformer to learn the long-term spatiotemporal representation. The intuition behind this design is we assume the natural clues of real videos exist in both single frame and long-range sequence. And due to the low information density [20], pixel space could not provide crucial information for natural consistency representation learning. But CNNs and Transformer have their inherent advantages in local spatial receptive field and long-term memory respectively [9, 21, 40, 53]. By incorporating both, we could explore the spatiotemporal representation from local spatial and long-range dimensions to enhance the representation learning.

Further, two specifically designed self-supervised tasks, Spatial Predictive Module (SPM) and Temporal Contrastive Module (TCM), are introduced to enhance the natural consistency learning. The SPM aims to reconstruct random masked spatial features from learned spatiotemporal representation by incorporating a CNN decoder. The TCM regularizes the latent distance of spatiotemporal representation pairs by shuffling the original frame order, which disturbs the natural consistency. The intuition for these two modules is that we assume a desired NACO representation could use the spatiotemporal context to predict the missing information for SPM and should be sensitive to the frame order which indicates the natural consistency for TCM. Leveraging the self-supervised

natural consistency learning, our detector does not so easily overfit the forgery methods or videos in training datasets as in supervised methods. Finally, a MLP head is fine-tuned guided by NACO representation to perform the usual forgery video classification task.

In brief, our contributions are summarized as follows: (1) We propose to learn the Natural Consistency representation (NACO) of visual-only real face videos for general and robust face forgery detection, which leverages the advantages of CNNs and Transformer on local spatial receptive field for single frame and long-term memory for frame sequence respectively. (2) Two specifically designed self-supervised tasks are introduced, including Spatial Predictive Module (SPM) and Temporal Contrastive Module (TCM), which both serve to enhance the natural consistency learning on real face videos. (3) Extensive experiments on public datasets demonstrate the superiority of our proposed method over the state-of-the-art competitors with impressive generalization and robustness.

2 Related Work

Face Forgery Video Detection. Since high-fidelity face forgery videos cause severe threats to society, how to detect them has become an urgent and essential issue. Recent deep-learning based detectors have achieved impressive performance for both fully-supervised and self-supervised detectors. Existing fully-supervised detectors naively train a supervised binary classifier based on specific patterns to distinguish between real and fake videos, such as frequency [39], artifacts [11,23,32,51], and spatiotemporal clues [1,22,31,36,55,57,58,65]. However, these specific patterns can be easily corrupted under common perturbations or eliminated by other unknown forgery methods, leading to limited generalization and robustness. Other self-supervised detectors aim to leverage model-agnostic representation to detect, such as LipForensics [19] and RealForensics [18]. However, LipForensics requires labeled datasets for pretraining and only focuses on the mouth region, RealForensics requires both visual-audio modalities for pretraining and performs visual multi-task during finetuning, which are strict on the auxiliary tasks and the pertaining datasets that may limit their scalability and performance. In contrast, we aim to leverage the natural consistency of visual-only real face videos without additional requirements in both local receptive spatial and long-range spatiotemporal dimensions to learn the model-agnostic representation for general and robust detection.

Self-Supervised Learning. The most common strategy in self-supervised learning is designing auxiliary tasks to introduce supervision without labels [16,20,37, 56,61]. Contrastive learning, which pulls positive pairs closer and pushes negative pairs away in latent space, has achieved impressive performance in self-supervised representation learning [5,30,38,54,62]. Masked models also demonstrate impressive representation capacity by masking part of the input and forcing the model to predict them by leveraging the context [20, 26, 49, 56]. Some self-supervised detectors focus on specific pattern clues [3, 4, 14, 29, 44, 60], which are susceptible to perturbations and unknown forgery methods. Others require large labeled

or multi-modal datasets for pretraining [12, 18, 19, 64], which are strict on the auxiliary tasks and the pertaining datasets that may limit their generalization and scalability. Different from them, we focus on the natural consistency of visual-only real videos without additional requirements for both pretraining and finetuning phases with two specifically designed auxiliary tasks (SPM and TCM) in both spatiotemporal dimensions.

Vision Transformer. Transformers have achieved impressive performance in various natural language processing tasks by introducing self-attention mechanism [50]. Inspired by this, researchers in computer vision field also seek to explore its potential applications. The vision transformer (ViT) [9] could be the first to apply transformer to computer vision tasks by treating image patches as token sequence. Since then, various ViT-based works have been proposed for different vision tasks, such as semantic segmentation [10, 45] and object detection [9]. ViT has also been applied to face forgery detection task [8,12,59,64,65]. However, these methods either directly adopt ViT to train an end-to-end supervised classifier while ignoring its self-supervised representation capacity, especially on long-range memory [8, 59, 65], or train in a self-supervised way on multi-modalities [12,64]. In contrast, our method incorporates Transformer with CNNs to achieve visual-only natural consistency representation learning in both local spatial receptive [13, 15, 21] and long-term spatiotemporal dimensions.

3 Method

In this section, we discuss our proposed NACO which consists of two stages, as illustrated in Fig. 2. The first stage aims to learn the NACO representation by first modeling the spatiotemporal representation of videos and then employing two designed self-supervised tasks (SPM and TCM) for natural consistency learning on visual-only real face videos. The NACO representations are then used to guide the binary forgery video classification task in the second stage by optimizing a MLP head. The details of each stage are presented below.

3.1 Spatiotemporal Representation

We aim to develop a general and robust face forgery video detector by leveraging the natural consistency of real face videos as clues. Hence, the initial step is to model the spatiotemporal representation of video. We first question whether the representation can be learned directly from pixel space. Due to its low information density [20], the pixel space lacks crucial information for natural consistency learning. Further, based on our observations, we intuitively assume that the natural clues in real videos exist in both individual frames and long sequences. Thus, our method aims to model the spatiotemporal representation from two distinct aspects: local receptive spatial (single frame) and long-range spatiotemporal (long-range sequence) dimensions, described in detail as follows:

Local Receptive Spatial Feature. Instead of learning directly from pixel space, we initially introduce a simply designed CNNs to extract the local receptive spatial feature of each frame, which we assume can provide more local spatial

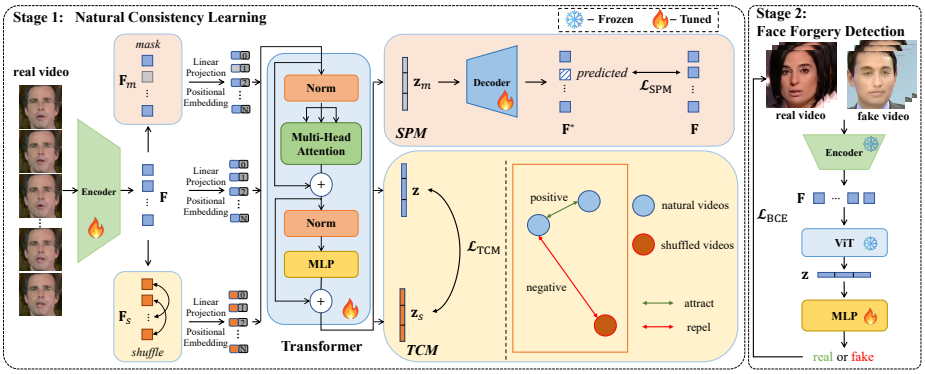

Fig. 2: The pipeline of our proposed NACO. In the first stage, real videos are first extracted into spatial feature sequence $\mathbf{F}$ by CNN encoder, which is fed into the Transformer to learn long-range spatiotemporal representation $\mathbf{z}$. Further, two designed auxiliary tasks: SPM and TCM are introduced to enhance the natural consistency learning on real face videos. In the second stage, the encoder and Transformer are frozen and a fully-connected classification head (two-layer MLP) guided by learned NACO representation is optimized to perform the usual face forgery video classification task.

natural clues by CNN's inherent advantage on local spatial receptive field [21]. We assume access to a real face video dataset $\mathcal{D}_r$. Each sample in $\mathcal{D}_r$ represents a real face video consisting of multiple frames. We sample random clips $\mathbf{X}$ from each video with consistent n frames employing face extraction and alignment to get the input frame sequence of our model, where $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n}$ and $\mathbf{X} \in \mathcal{D}_r$. Then we introduce a simple CNN encoder to extract the spatial feature of each frame $\mathbf{x}_i$, which projects the frame sequence into spatial feature sequence $\mathbf{F}$, formulated as follows:

$$\mathbf{F} = \mathrm{Conv}(\mathbf{X}, \Theta_e) = \mathrm{Conv}(\{\mathbf{x}_i\}_{i=1}^{n}, \Theta_e) = \{\mathbf{f}_i\}_{i=1}^{n}, \tag{1}$$

where the $\{\Theta_e\}$ is the parameters of encoder and the dimension of $\mathbf{f}_i$ is 256.

Long-range Spatiotemporal Representation. We further explore the natural clues in the long-range sequence to learn the long-term spatiotemporal representation of real face videos. To achieve this, we incorporate Transformer for its inherent advantage in long-term memory [9, 40, 53] by integrating the extracted spatial features. Specifically, we reshape $\mathbf{f}_i$ into 16×16 tokens and employ a trainable linear projection matrix $\mathbf{W}$ with positional embedding $\mathbf{E}_{pos}$ to get the input feature sequence and feed them into our Transformer to learn the spatiotemporal representation $\mathbf{z}$ of each video clip, which mainly consists of K standard Transformer Encoder blocks [50], each block contains a multi-head self-attention (MSA) block and an MLP block with the commonly-used LayerNorm (LN) before them. We also use the GELU as the activation function and

the whole process can be formulated as follows:

$$\mathbf{z}_0 = \mathbf{W} \cdot \mathbf{F} + \mathbf{E}_{pos} = \mathbf{W} \cdot [\mathbf{f}_1, \cdots, \mathbf{f}_n]^{\mathrm{T}} + \mathbf{E}_{pos}, \tag{2}$$

$$\mathbf{z}_k = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{k-1})) + \mathbf{z}_{k-1}, \qquad k = 1 \ldots K \tag{3}$$

We denote all the trainable parameters of the Transformer as $\{\Theta_v\}$ and regard the last-layer output $\mathbf{z}_K$ as the learned spatiotemporal representation with dimension of 768. Unless stated otherwise, the mentioned representation $\mathbf{z}$ in the following represents the representation learned from the last-layer $\mathbf{z}_K$, which we denote as $\mathbf{z}$ for simplicity. Furthermore, two designed self-supervised tasks are introduced to enhance the natural consistency representation learning in latent space, described in detail in the following.

3.2 Natural Consistency Learning

We assume that low-level clues such as artifacts could be corrupted by unknown manipulation types or different perturbations, which causes limited generalization and robustness. But the high-level clues, such as the natural spatiotemporal consistency would still exist under both situations, since forgery methods typically struggle to preserve such clues during generation and high-level features are naturally less susceptible to common perturbations. Hence, we aim to leverage the natural consistency of visual-only real videos in a self-supervised manner to develop a general detector that can achieve both high generalization and robustness. Based on the spatiotemporal representation learned from visual-only real videos described above, we further design two self-supervised tasks: Spatial Predictive Module (SPM) and Temporal Contrastive Module (TCM) to enhance natural consistency learning, described in detail below.

Spatial Predictive Module (SPM). We assume that an effective natural consistency representation should learn the relationship between consistent spatial features at different timestamps and could use the context to predict the missing information, preventing from relying on specific input. Some previous works have also demonstrated this [20, 56]. Thus, we initially randomly mask parts of extracted local spatial features $\mathbf{f}_i$ before feeding them into Transformer. Then we introduce a CNN decoder to predict the masked spatial features.

To make this process clearer, we first define one operation between two sequence: $\mathbf{A} \ominus \mathbf{B}$ which indicates the indices at which the element is not masked in $\mathbf{A}$ but masked in $\mathbf{B}$. For example, if $\mathbf{A} = \{\mathbf{a}\ \mathbf{b}\ \mathbf{c}\ \mathbf{d}\}$ and $\mathbf{B} = \{\mathbf{a}$ [mask] $\mathbf{c}$ [mask] $\}$, then we have $\mathbf{A} \ominus \mathbf{B} = \{2, 4\}$.

Then for extracted local spatial feature sequence $\mathbf{F}$, we first random mask parts of $\mathbf{F}$ with a ratio α to get the masked local spatial feature sequence $\mathbf{F}_m = \mathrm{Mask}(\mathbf{F}, \alpha)$, where there are total $n \times \alpha$ spatial features $\mathbf{f}_i$ are randomly masked by setting the tensor value to zeros. Then the $\mathbf{F}_m$ is fed into Transformer to learn the spatiotemporal representation $\mathbf{z}_m$ of the masked sequence.

Further, a CNN decoder is introduced to predict the masked local spatial feature from the learned masked spatiotemporal representation $\mathbf{z}_m$. We calculate the distance between the original and predicted sequence to regularize the

prediction process. Specifically, we only compute the distance on masked spatial features, which can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{SPM}} &= \|\mathbf{F}^{*}-\mathbf{F}\| \\ &= \|\mathrm{Conv}(\mathbf{z}_m,\Theta_d)-\mathbf{F}\| \\ &= \sum_{i\in\mathbf{F}\ominus\mathbf{F}_m}\|\mathbf{f}_i^{*}-\mathbf{f}_i\|,\end{aligned} \tag{4}$$

where $\mathbf{F}^{*}$ is the predicted local spatial feature sequence and $\{\Theta_d\}$ is the parameters of the CNN decoder. We use the mean squared error (MSE loss) as the distance function.

Temporal Contrastive Module (TCM). We further hypothesize that a desired natural consistency representation should be sensitive to the frame sequence order, which also indicates the natural consistency. To address this, we aim to regularize the distance of spatiotemporal representation in latent space using contrastive learning by disturbing the original frame order. We initially disturb the natural consistency in real face videos by random shuffling the frame sequence order, which is also equivalent to shuffling the extracted local spatial features order to get $\mathbf{F}_s = \mathrm{Shuffle}(\mathbf{F})$, where $\mathbf{F}_s \neq \mathbf{F}$.

Then we introduce contrastive learning to regularize the distance in latent space. The key question here is how to construct the positive and negative pairs for contrastive learning. We first define the similarity of two different spatiotemporal representation $(\mathbf{z}_i, \mathbf{z}_j)$ learned by the Transformer, which can be formulated as follows:

$$\mathrm{sim}\left(\mathbf{z}_i,\mathbf{z}_j\right)=\frac{\mathbf{z}_i\cdot\mathbf{z}_j}{\max(\|\mathbf{z}_i\|_2\cdot\|\mathbf{z}_j\|_2,\epsilon)}, \tag{5}$$

where hyper-parameter ϵ is set to $1e-8$. Then for a natural unshuffled local spatial feature sequence $\mathbf{F}$ and corresponding spatiotemporal representation $\mathbf{z}$, we consider the representation learned from other natural unshuffled input as positive pairs $\mathbf{z}^{+}$, while the $\mathbf{z}_s$ from shuffled spatial feature sequence $\mathbf{F}_s$ as negative pairs $\mathbf{z}^{-}$. Thus, we can formulate the contrastive loss $\mathcal{L}_{\mathrm{TCM}}$ within mini-batch samples as follows:

$$\mathcal{L}_{\mathrm{TCM}}=-\log\frac{\exp\left(\mathrm{sim}\left(\mathbf{z},\mathbf{z}^{+}\right)/\tau\right)}{\exp\left(\mathrm{sim}\left(\mathbf{z},\mathbf{z}^{+}\right)/\tau\right)+\sum\exp\left(\mathrm{sim}\left(\mathbf{z},\mathbf{z}^{-}\right)/\tau\right)}, \tag{6}$$

where the temperature τ is set to 0.5. Specifically, for each input video clip during each iteration, we shuffle its order to get one negative sample.

Therefore, the total loss function for the self-supervised natural consistency representation learning stage can be formulated as follows:

$$\mathcal{L}=\lambda_1\mathcal{L}_{\mathrm{SPM}}+\lambda_2\mathcal{L}_{\mathrm{TCM}}, \tag{7}$$

where $\{\lambda_1, \lambda_2\}$ are the hyper-parameter weights to balance the SPM and TCM loss. Optimizing by these two designed tasks, we finally obtain the desired Natural Consistency representation (NACO) by $\mathbf{z}^{*} = f(\mathbf{z}, \Theta_e, \Theta_v, \Theta_d)$, where $f(\cdot)$ represents the natural consistency learning process.

3.3 Face Forgery Detection

After the natural consistency representation learning, we freeze the backbone including the encoder and Transformer, and add a fully-connected classification head (two-layer MLP) which takes the learned NACO representation to classify real and fake videos. Since fake videos can't preserve the natural consistency as real videos, they should be discriminated by our NACO generally. Given an input video $\widehat{\mathbf{X}}$, the detection can be formulated as follows:

$$y = \mathrm{MLP}(\hat{\mathbf{z}}^*, \Theta_c) = \mathrm{MLP}(g(\widehat{\mathbf{X}}, \Theta_e, \Theta_v), \Theta_c), \tag{8}$$

where the $\{\Theta_e, \Theta_v, \Theta_c\}$ denote the parameters of the encoder, Transformer, and classification head, and $g(\cdot)$ means the process to obtain NACO representation from $\{\Theta_e, \Theta_v\}$ during fine-tuning. Noticing that only $\{\Theta_c\}$ is optimized in this stage while others are frozen. Then given a labeled forgery video dataset which includes both real and fake videos,we choose the vanilla binary cross-entropy loss $\mathcal{L}_{\mathrm{BCE}}$ to supervise the forgery classification task.

4 Experiment

4.1 Experimental Settings

Datasets. We choose the large-scale real face video dataset: VoxCeleb2 [6] for pretraining, which contains over 1 million utterances from 6,112 celebrities, extracted from videos on YouTube. For the face forgery video detection task, we choose following datasets for finetuning and evaluation: (1) FaceForensics++(FF++) [41] contains 1,000 real videos and 4,000 fake videos with three different compression levels (raw/c23/c40, from uncompressed to heavily compressed) generated from four different manipulation methods, including two face swapping methods, Deepfakes [3] (DF) and FaceSwap [4] (FS), and two face reenactment methods, Face2Face [48] (F2F) and NeuralTextures [47] (NT). Unless stated otherwise, we use the mildly compressed version of the dataset (c23). (2) Celeb-DF-v2(CDF) [33] is a challenging dataset including 590 real videos and 5,639 fake videos. (3) DFDC [7] is a subset of the Deepfake Detection Challenge Dataset [5], where each video is recorded in challenging environments. (4) FaceShifter(FSh) [28] is the recent high-fidelity face swapping method that has been applied to the real videos of FF++. (5) DeeperForensics(DFo) [24] contains real videos recorded in difficult real-world scenarios and high-fidelity forgery videos based on the real videos from FF++.

Evaluation metrics. Following recent works [18,19,55,65], we report the video-level Area Under the Receiver Operating Characteristic Curve (AUC(%)) and Accuracy (ACC(%)) to compare with prior works. For frame-level detectors, the metrics are averaged over sampled video frames.

[3] https://github.com/deepfakes/faceswap
[4] https://github.com/MarekKowalski/FaceSwap
[5] https://www.kaggle.com/c/deepfake-detection-challenge/data

Implementation details. For each video, we sample continuous 20 frames reshaped into 224×224 from a random offset as input sequence and repeat three times to get the mean results, which should be beneficial for different video lengths. Then we employ face extraction and alignment using tool [6]. The encoder consists of three convolutional layers with 3×3 kernel and 64, 128, 256 channels and one average pooling layer which extract spatial features of 256 dimensions, and the decoder is one convolutional layer with 1×3 kernel which projects the spatiotemporal representation of 768 dimensions into 256 spatial features. We choose the basic ViT-Base architecture described in [9] with $K = 12$ blocks, whose final output is 768 dimensions representation. We use a two-layer MLP including one hidden layer of 256 dimensions and one output layer with Softmax activation as our classification head. We employ the Adam optimizer with an initial learning rate of $5e-4$ and a weight decay of $1e-4$. The batch size is set to 64 and we empirically set $\alpha = 1/2$ in SPM, $\lambda_1 = 1.0$ and $\lambda_2 = 0.5$ in Eq. (7).

4.2 Experimental Results

Table 1: Generalization to unseen datasets. Video-level AUC (%) is reported. Other methods' results are from [18] and their original papers.

Method	CDF	DFDC	FSh	DFo	*Avg*
Xception [41]	73.7	70.9	72.0	84.5	75.3
CNN-aug [51]	75.6	72.1	65.7	74.4	72.0
Patch-based [2]	69.6	65.6	57.8	81.8	68.7
Two-branch [35]	76.7	-	-	-	-
Face X-ray [29]	79.5	65.5	92.8	86.8	81.2
Multi-task [36]	75.5	68.1	66.0	77.7	71.9
DSP-FWA [32]	69.5	67.3	65.5	50.2	63.1
CNN-GRU [42]	69.8	68.9	80.8	74.1	73.4
LipForensics-scratch [19]	62.5	65.5	84.7	84.8	74.4
LipForensics [19]	82.4	73.5	97.1	97.6	87.7
FTCN [65]	<u>86.9</u>	74.0	98.8	98.8	89.6
RealForensics-scratch [18]	69.4	68.1	87.9	89.3	78.7
RealForensics [18]	<u>86.9</u>	<u>75.9</u>	**99.7**	<u>99.3</u>	<u>90.5</u>
ISTVT [63]	84.1	74.2	99.3	98.6	89.1
NoiseDF [52]	75.9	63.9	-	70.9	-
AltFreezing [55]	**89.5**	-	<u>99.4</u>	<u>99.3</u>	-
NACO-scratch (ours)	67.9	69.4	88.3	89.5	78.8
NACO (ours)	**89.5**	**76.7**	<u>99.4</u>	**99.5**	**91.2**

Generalization to unseen datasets. We first evaluate our method's generalization by training on FF++ then evaluating on other four unseen challenging datasets as presented in Tab. 1 (scratch means directly fine-tuning without pre-training). We observe that both frame-level detectors, such as [41,51] and simply

[6] https://github.com/1adrianb/face-alignment

designed supervised video-level detectors [42] lead to limited generalization. But the video-level detectors focusing on spatiotemporal features achieve better performance, such as [18, 19, 55, 65] and our NACO, providing additional evidence that high-level spatiotemporal clues are the key for generalization.

Furthermore, our NACO outperforms other state-of-the-art supervised [55, 65] and self-supervised video-level detectors [55, 65] with the highest average 91.2% AUC score across the four unseen datasets, indicating the superiority by leveraging natural consistency of real face videos. And the performance is also better when training from scratch compared to LipForensics and RealForensics. The promising results suggest the potential of our method to detect more challenging unseen forgery videos in the future.

Table 2: Generalization to unseen manipulations. Video-level AUC (%) on each subset of FF++ is reported. Other methods' results are from [18] and original papers.

Method	Train on remaining three				*Avg*
	DF	FS	F2F	NT	
Xception [41]	93.9	51.2	86.8	79.7	77.9
CNN-aug [51]	87.5	56.3	80.1	67.8	72.9
Patch-based [2]	94.0	60.5	87.3	84.8	81.7
Face X-ray [29]	99.5	93.2	94.5	92.5	94.9
CNN-GRU [42]	97.6	47.6	85.8	86.6	79.4
LipForensics-scratch [19]	93.0	56.7	98.8	98.3	86.7
LipForensics [19]	99.7	90.1	<u>99.7</u>	99.1	97.1
FTCN [65]	<u>99.9</u>	**99.9**	<u>99.7</u>	<u>99.2</u>	**99.7**
RealForensics-scratch [18]	98.8	87.9	98.7	88.6	93.5
RealForensics [18]	**100.**	97.1	<u>99.7</u>	<u>99.2</u>	<u>99.0</u>
AltFreezing [55]	99.8	<u>99.7</u>	98.6	96.2	<u>98.6</u>
NACO-scratch (ours)	98.9	88.3	97.9	89.1	93.6
NACO (ours)	<u>99.9</u>	<u>99.7</u>	**99.8**	**99.4**	**99.7**

Generalization to unseen manipulations. We further evaluate our method's generalization to unknown manipulation methods on each subset of FF++ by training on three and evaluating on the remaining one as shown in Tab. 2. We observe that both frame-level detectors that focus on low-level artifacts [41] and naive supervised video-level detectors [42] suffer significant drops when detecting unknown forgery types. But the detectors which leverage high-level spatiotemporal clues achieve better performance, such as [18, 19, 55, 65] and our NACO. This also supports our motivation that the inconsistency still exists in unknown forgery types, regardless of face swapping or reenactment. Moreover, our method outperforms other state-of-the-art supervised [55, 65] and self-supervised video-level methods [18, 19] with achieving 99.7% average AUC, indicating the effectiveness of leveraging natural consistency representation in visual-only real face videos to defend unseen manipulation methods. Besides, the generalization is also better when training from scratch compared to [18, 19].

Furthermore, we also compare the parameters and architectures in Tab. 3. We observe that our method which only optimizes a MLP head achieves the minimum finetuning parameters with impressive generalization.

Table 3: Trainable parameters for usual face forgery detection task and generalization comparisons. Video-level AUC (%) is reported when trained on FF++. 2D/3D means the CNN architectures and TF means the Transformer.

Method	#params	Arch	FSh	DFo
LipForensics-scratch [19]	36.0M	2D+MS-TCN	84.7	84.8
LipForensics [19]	24.8M	2D+MS-TCN	97.1	97.6
FTCN [65]	26.6M	3D+TF	98.8	98.8
RealForensics-scratch [18]	21.4M	2D+CSN	87.9	89.3
RealForensics [18]	21.4M	2D+CSN	**99.7**	99.3
AltFreezing [55]	27.2M	3D	99.4	99.3
NACO (ours)	**4.7M**	2D+TF	99.4	**99.5**

Table 4: Robustness to unseen perturbations. Video-level AUC (%) on FF++ under seven different perturbations described in [24]. Other methods' results are from [18] and * denotes our reproduction.

Method	Clean	Saturation	Contrast	Block	Noise	Blur	Pixel	Compress	*Avg/Drop*
Xception [41]	99.8	99.3	98.6	**99.7**	53.8	60.2	74.2	62.1	78.3/-21.5
CNN-aug [51]	99.8	99.3	99.1	95.2	54.7	76.5	91.2	72.5	84.1/-15.7
Patch-based [2]	99.9	84.3	74.2	99.2	50.0	54.4	56.7	53.4	67.5/-32.4
X-Ray [29]	99.8	97.6	88.5	99.1	49.8	63.8	88.6	55.2	77.5/-22.3
CNN-GRU [42]	99.9	99.0	98.8	97.9	47.9	71.5	86.5	74.5	82.3/-17.6
LipForensics [19]	99.9	**99.9**	99.6	87.4	73.8	96.1	95.6	95.6	92.6/-7.3
FTCN [65]	99.4	99.4	96.7	97.1	53.1	95.8	98.2	86.4	89.5/-9.9
RealForensics [18]	99.8	99.8	99.6	98.9	79.7	95.3	98.4	**97.6**	95.6/-4.2
AltFreezing* [55]	99.9	99.5	**99.8**	97.1	75.2	**97.4**	98.1	92.6	94.2/-5.7
NACO (ours)	99.9	98.9	98.2	98.4	**86.5**	96.2	**98.6**	96.7	**96.2/-3.7**

Robustness to unseen perturbations. We further investigate our model's robustness against various unseen perturbations by training on uncompressed videos and evaluating on videos added perturbations. We consider the following seven perturbations described in [24]: saturation, contrast, block-wise, Gaussian noise, Gaussian blur, pixelation, and video compression. Each perturbation is applied under five different severity levels. We report the average video-level AUC scores across all severity levels and make comparisons in Tab. 4. We find our NACO suffers significantly less than frame-level detectors, such as [29, 41]. This indicates leveraging high-level natural consistency leads to more robust detection than relying on low-level clues corrupted in perturbations, which is consistent with our initial motivation. Furthermore, our method also outperforms other recent supervised and self-supervised video-level SOTAs [18, 19, 55, 65], with the minimum AUC drop of -3.7%, especially on Gaussian noise, blur, and pixelation,

which disturb more on consistency. This also provides additional evidence of our method's superior robustness to unseen perturbations by leveraging the natural consistency of real videos.

4.3 Ablation Study

We conduct further ablations by fine-tuning on FF++ (c23) and testing on FF++ (c40) and Celeb-DF (CDF) to evaluate the robustness and generalization.
Number of real samples. We first investigate how the quantity of real face videos used for natural consistency learning affects our model's performance. We regard the whole original number of real samples as 1.0 and reduce the scale with an interval of 0.2, noticing that setting the data scale to 0.0 means we directly fine-tune the model with random initialization without pretraining stage. The results are presented in Fig. 3. We can see that our method benefits from a large number of real samples, which proves the effectiveness of our proposed NACO by leveraging the natural consistency representation on visual-only real face videos.

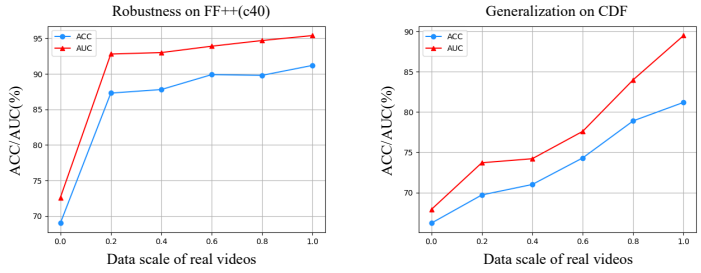

Fig. 3: Comparisons on different number of real samples in natural consistency representation learning stage.

Natural Consistency Learning. We further investigate how the two designed self-supervised tasks (SPM and TCM) for natural consistency learning effect our model's performance by employing each respectively as presented in Tab. 5. We observe that employing both modules achieve the highest performance on both robustness and generalization, average 1.05% and 9.60% AUC improvements, which indicates both designed tasks have positive effect on the natural consistency representation learning on real face videos.

4.4 Forgery Localization

Furthermore, we provide the Grad-CAM [43] response based on the predicted spatial features on five consecutive frames from each subset of FF++ (c23) in Fig. 4. From the results, we observe that the responses of forgery videos focus on the specific face areas, such as the mouth and center face, where exists inconsistency. But the response of real videos is average over the entire facial area

Table 5: Analysis of the two designed self-supervised tasks (SPM and TCM) for natural consistency learning.

$\mathcal{L}_{\text{SPM}}$	$\mathcal{L}_{\text{TCM}}$	FF++ (c40)		CDF	
		ACC (%)	AUC (%)	ACC (%)	AUC (%)
✓	-	90.7	94.6	77.8	80.6
-	✓	87.8	94.1	74.3	79.2
✓	✓	91.2	95.4	81.2	89.5

since there are no inconsistencies in natural real videos. The results indicate that our method can effectively localize the inconsistent forgery areas in fake videos and provide human-trustable explanations for decision results.

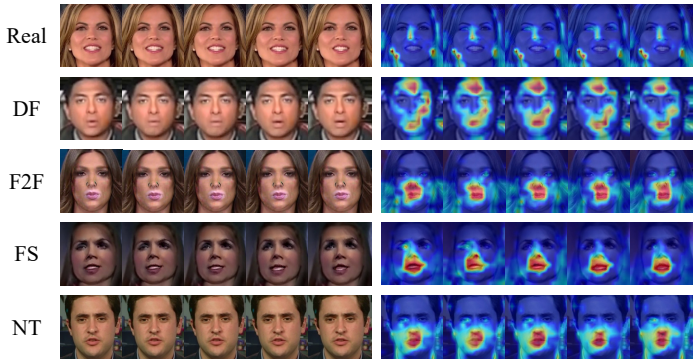

Fig. 4: Forgery localization. Grad-CAM results on five consecutive frames on FF++ (c23). We find that our method can effectively respond to the inconsistencies in fake videos and localize the forgery areas.

5 Conclusion

In this paper, we propose to learn the Natural Consistency representation (NACO) of visual-only real face videos to develop a general and robust face forgery detector. NACO initially extracts spatial features of each single frame by CNNs then integrates them into Transformer to learn long-range spatiotemporal representation. Furthermore, two specifically designed self-supervised tasks, Spatial Predictive Module (SPM) and Temporal Contrastive Module (TCM), are introduced to enhance the natural consistency learning on visual-only real face videos. Extensive experiments have shown that our method achieves impressive generalization to unknown forgery types and robustness to various perturbations. In the future, we aim to apply our model to more complicated samples from real scenarios, such as from the web, and also extend our method to other media forensic tasks, such as audio and multi-modalities.

Acknowledgements

This work was partially supported by grants from the Pioneer R&D Program of Zhejiang Province (2024C01024), Open Research Project of the State Key Laboratory of Media Convergence and Communication, Communication University of China (SKLMCC2022KF004).

References

1. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: WIFS. pp. 1–7 (2018)
2. Chai, L., Bau, D., Lim, S.N., Isola, P.: What makes fake images detectable? understanding properties that generalize. In: ECCV. pp. 103–120 (2020)
3. Chen, L., Zhang, Y., Song, Y., Liu, L., Wang, J.: Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In: CVPR. pp. 18689–18698 (2022)
4. Chen, L., Zhang, Y., Song, Y., Wang, J., Liu, L.: OST: improving generalization of deepfake detection via one-shot test-time training. In: NeurIPS (2022)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020)
6. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: Interspeech. pp. 1086–1090 (2018)
7. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C.: The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854 (2019)
8. Dong, X., Bao, J., Chen, D., Zhang, T., Zhang, W., Yu, N., Chen, D., Wen, F., Guo, B.: Protecting celebrities from deepfake with identity consistency transformer. In: CVPR. pp. 9468–9478 (2022)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
10. Duke, B., Ahmed, A., Wolf, C., Aarabi, P., Taylor, G.W.: Sstvos: Sparse spatiotemporal transformers for video object segmentation. In: CVPR. pp. 5912–5921 (2021)
11. Fei, J., Dai, Y., Yu, P., Shen, T., Xia, Z., Weng, J.: Learning second order local anomaly for general face forgery detection. In: CVPR. pp. 20238–20248 (2022)
12. Feng, C., Chen, Z., Owens, A.: Self-supervised video forensics by audio-visual anomaly detection. In: CVPR. pp. 10491–10503 (2023)
13. Ge, S., Li, J., Ye, Q., Luo, Z.: Detecting masked faces in the wild with lle-cnns. In: CVPR. pp. 2682–2690 (2017)
14. Ge, S., Lin, F., Li, C., Zhang, D., Wang, W., Zeng, D.: Deepfake video detection via predictive representation learning. ACM TOMM **18**(2s), 115:1–115:21 (2022)
15. Ge, S., Zhao, S., Li, C., Li, J.: Low-resolution face recognition in the wild via selective knowledge distillation. IEEE Transactions on Image Processing **28**(4), 2051–2062 (2018)
16. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
17. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: NeurIPS. pp. 2672–2680 (2014)

18. Haliassos, A., Mira, R., Petridis, S., Pantic, M.: Leveraging real talking faces via self-supervision for robust forgery detection. In: CVPR. pp. 14930–14942 (2022)
19. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: CVPR. pp. 5039–5049 (2021)
20. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 15979–15988 (2022)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
22. Hu, Z., Xie, H., Wang, Y., Li, J., Wang, Z., Zhang, Y.: Dynamic inconsistency-aware deepfake video detection. In: IJCAI. pp. 736–742 (2021)
23. Hua, Y., Zhang, D., Wang, P., Ge, S.: Interpretable face manipulation detection via feature whitening. arXiv preprint arXiv:2106.10834 (2021)
24. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In: CVPR. pp. 2886–2895 (2020)
25. Laurens, V.D.M., Hinton, G.: Visualizing data using t-sne. JMLR **9**(2605), 2579–2605 (2008)
26. Li, C., Ge, S., Zhang, D., Li, J.: Look through masks: Towards masked face recognition with de-occlusion distillation. In: ACM MM. pp. 3016–3024 (2020)
27. Li, J., Li, J., Zhang, H., Liu, S., Wang, Z., Xiao, Z., Zheng, K., Zhu, J.: Preim3d: 3d consistent precise image attribute editing from a single image. In: CVPR. pp. 8549–8558 (2023)
28. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Advancing high fidelity identity swapping for forgery detection. In: CVPR. pp. 5074–5083 (2020)
29. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: CVPR. pp. 5001–5010 (2020)
30. Li, S., Xia, X., Ge, S., Liu, T.: Selective-supervised contrastive learning with noisy labels. In: CVPR. pp. 316–325 (2022)
31. Li, Y., Chang, M., Lyu, S.: In ictu oculi: Exposing AI created fake videos by detecting eye blinking. In: WIFS. pp. 1–7 (2018)
32. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. In: CVPRW. pp. 46–52 (2019)
33. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: A large-scale challenging dataset for deepfake forensics. In: CVPR. pp. 3204–3213 (2020)
34. Lu, Y., Tai, Y., Tang, C.: Attribute-guided face generation using conditional cyclegan. In: ECCV. pp. 293–308 (2018)
35. Masi, I., Killekar, A., Mascarenhas, R.M., Gurudatt, S.P., AbdAlmageed, W.: Two-branch recurrent network for isolating deepfakes in videos. In: ECCV. pp. 667–684 (2020)
36. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. In: BTAS. pp. 1–8 (2019)
37. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV. pp. 69–84 (2016)
38. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
39. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: ECCV. pp. 86–103 (2020)
40. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. NeurIPS **32** (2019)

41. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: ICCV. pp. 1–11 (2019)
42. Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., Natarajan, P.: Recurrent convolutional strategies for face manipulation detection in videos. In: CVPRW. pp. 80–87 (2019)
43. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017)
44. Shiohara, K., Yamasaki, T.: Detecting deepfakes with self-blended images. In: CVPR. pp. 18699–18708 (2022)
45. Strudel, R., Pinel, R.G., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: ICCV. pp. 7242–7252 (2021)
46. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM TOG **36**(4), 95:1–13 (2017)
47. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: image synthesis using neural textures. ACM TOG **38**(4), 66:1–12 (2019)
48. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: CVPR. pp. 2387–2395 (2016)
49. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. NeurIPS **35**, 10078–10093 (2022)
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., et al.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)
51. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-generated images are surprisingly easy to spot... for now. In: CVPR. pp. 8695–8704 (2020)
52. Wang, T., Chow, K.P.: Noise based deepfake detection via multi-head relative-interaction. In: AAAI. vol. 37, pp. 14548–14556 (2023)
53. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (2018)
54. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: ICCV. pp. 2794–2802 (2015)
55. Wang, Z., Bao, J., Zhou, W., Wang, W., Li, H.: Altfreezing for more general video face forgery detection. In: CVPR. pp. 4129–4138 (2023)
56. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: CVPR. pp. 9653–9663 (2022)
57. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP. pp. 8261–8265 (2019)
58. Zhang, D., Li, C., Lin, F., Zeng, D., Ge, S.: Detecting Deepfake Videos with Temporal Dropout 3DCNN. In: IJCAI. pp. 1288–1294 (2021)
59. Zhang, D., Lin, F., Hua, Y., Wang, P., Zeng, D., Ge, S.: Deepfake video detection with spatiotemporal dropout transformer. In: ACM MM. pp. 5833–5841 (2022)
60. Zhang, D., Xiao, Z., Li, J., Ge, S.: Self-supervised transformer with domain adaptive reconstruction for general face forgery video detection. arXiv preprint arXiv:2309.04795 (2023)
61. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV. pp. 649–666 (2016)
62. Zhang, T., Qiu, C., Ke, W., Süsstrunk, S., Salzmann, M.: Leverage your local and global representations: A new self-supervised learning strategy. In: CVPR. pp. 16580–16589 (2022)

63. Zhao, C., Wang, C., Hu, G., Chen, H., Liu, C., Tang, J.: Istvt: interpretable spatial-temporal video transformer for deepfake detection. TIFS **18**, 1335–1348 (2023)
64. Zhao, H., Zhou, W., Chen, D., Zhang, W., Yu, N.: Self-supervised transformer for deepfake detection. arXiv preprint arXiv:2203.01265 (2022)
65. Zheng, Y., Bao, J., Chen, D., Zeng, M., Wen, F.: Exploring temporal coherence for more general video face forgery detection. In: ICCV. pp. 15024–15034 (2021)