AltFreezing for More General Video Face Forgery Detection

Zhendong Wang [1,*] Jianmin Bao[2,*] Wengang Zhou[1,3,†] Weilun Wang[1] Houqiang Li[1,3,†]
[1] CAS Key Laboratory of GIPAS, EEIS Department, University of Science and Technology of China
[2] Microsoft Research Asia
[3] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{zhendongwang,wwlustc}@mail.ustc.edu.cn
jianbao@microsoft.com, {zhwg,lihq}@ustc.edu.cn

Abstract

*Existing face forgery detection models try to discriminate fake images by detecting only spatial artifacts (*e.g.*, generative artifacts, blending) or mainly temporal artifacts (*e.g.*, flickering, discontinuity). They may experience significant performance degradation when facing out-domain artifacts. In this paper, we propose to capture both spatial and temporal artifacts in one model for face forgery detection. A simple idea is to leverage a spatiotemporal model (3D ConvNet). However, we find that it may easily rely on one type of artifact and ignore the other. To address this issue, we present a novel training strategy called AltFreezing for more general face forgery detection. The AltFreezing aims to encourage the model to detect both spatial and temporal artifacts. It divides the weights of a spatiotemporal network into two groups: spatial-related and temporal-related. Then the two groups of weights are alternately frozen during the training process so that the model can learn spatial and temporal features to distinguish real or fake videos. Furthermore, we introduce various video-level data augmentation methods to improve the generalization capability of the forgery detection model. Extensive experiments show that our framework outperforms existing methods in terms of generalization to unseen manipulations and datasets.*

1. Introduction

With the recent rapid development of face generation and manipulation techniques [30, 31, 45–49, 56], it has become very easy to modify and manipulate the identities or attributes given a face video. This brings many important and impressive applications for movie-making, funny video generation, and so on. However, these techniques can also be abused for malicious purposes, creating serious crisis of trust and security in our society. Therefore, how to detect video face forgeries has become a hot research topic recently.

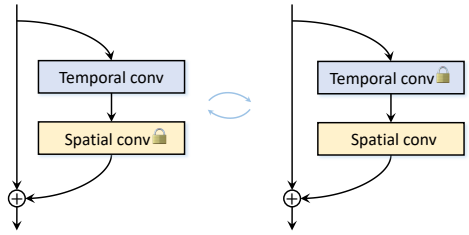

Figure 1. **Illustration of AltFreezing training strategy** in a building block of the spatiotemporal network. The convolutional kernels of the spatiotemporal network are divided into two groups: *temporal-based* and *spatial-based*. Two groups of weights are alternately frozen during training. With the help of the alternate freezing (*AltFreezing*) strategy, our model can capture both spatial and temporal artifacts to distinguish between fake and real videos.

To date, one successful line of research [10, 32, 34, 39, 42, 44, 50] tries to discriminate fake images by detecting "spatial" artifacts in the generated images (*e.g.*, checkboard, unnaturalness, and characteristic artifacts underlying the generative model). While these methods achieve impressive results in searching spatial-related artifacts, they ignore the temporal coherence of a video and fail to capture "temporal" artifacts like flicking and discontinuity in the video face forgeries. Some recent works [25,43,54] notice this issue and try to address it by leveraging temporal clues. Although they achieve encouraging results in detecting unnatural artifacts at the temporal level, the resulting models are not sufficiently capable of finding spatial-related artifacts.

In this paper, we attempt to capture both spatial and temporal artifacts for general video face forgery detection. Generally, a well-trained spatiotemporal network (3D ConvNet) has the capability of searching both spatial and temporal artifacts. However, we find that naïve training may cause it to easily rely on spatial artifacts while ignoring temporal artifacts to make a decision, causing a poor generalization

*Equal contribution.
†Corresponding authors.

capability. This is because spatial artifacts are usually more obvious than temporal incoherence, naïvely optimizing a 3D convolutional network makes it easily rely on spatial artifacts.

So the question is how to enable the spatiotemporal network to capture both spatial and temporal artifacts. To this end, we propose a novel training strategy called *AltFreezing*. As shown in Fig. 1, the key idea is to alternately freeze spatial- and temporal-related weights during training. Specifically, a spatiotemporal network [9] is built upon 3D resblocks, which consist of spatial convolution with kernel size as $1 \times K_h \times K_w$ and temporal convolution with kernel size as $K_t \times 1 \times 1$. These spatial and temporal convolutional kernels are responsible for capturing spatial- and temporal-level features, respectively. Our AltFreezing strategy encourages the two groups of weights to be updated alternately so that both spatial and temporal artifacts can be conquered.

Furthermore, we propose a set of video-level fake video argumentation methods for generating fake videos for training. These methods could be divided into two groups. The first is fake clips that only involve temporal artifacts wherein we just randomly drop and repeat frames for real clips. The second is clips with only spatial artifacts that are obtained by blending a region from one real clip to another real clip. These augmentation methods are the first to take the temporal dimension into consideration and generate spatial-only and temporal-only fake videos. With these augmentations, the spatiotemporal model is further encouraged to capture both spatial and temporal artifacts.

Equipped with the above-mentioned two techniques, we achieve state-of-the-art performance in various challenging face forgery detection scenarios, including generalization capability to unseen forgeries, and robustness to various perturbations. We also provide a comprehensive analysis of our method to verify the effectiveness of our proposed framework.

Our main contributions are three-fold as follows.

- We propose to explore both spatial and temporal artifacts for video face forgery detection. To achieve this, a novel training strategy called AltFreezing is proposed.
- We introduce video-level fake data augmentation methods to encourage the model to capture a more general representation of different types of forgeries.
- Extensive experiments on five benchmark datasets including both cross-manipulation and cross-dataset evaluations demonstrate that the proposed method sets new state-of-the-art performance.

2. Related Work

In the past few years, face forgery detection has been an emerging research area with the fast development of generative models and manipulation techniques. In this section, we briefly revisit the development of face forgery detection.

2.1. Image Face Forgery Detection

Earlier face forgery detection methods [4, 8, 10, 22, 26, 41, 50] mainly focus on spatial artifacts of manipulated images, and directly train a binary classifier based on CNN or MLP as the detector. Later Rossler *et al*. [42] suggest that an unconstrained Xception [11] network can achieve an impressive performance. some works pay more attention to special types of artifacts, such as frequency [28, 29, 37, 40], blending artifacts [20, 32, 44], resolution difference [35], and so on. Moreover, there are some works [7, 14, 24, 27, 39, 51] aiming to localize the forged regions and make a decision based on the predicted regions. A more recent work ICT [17, 18] tries to leverage identity information for detecting face forgeries.

2.2. Video Face Forgery Detection

Recent works [6, 13, 16, 25, 33, 38, 43] start to take temporal cues into consideration for face forgery detection. CNN-GRU [43] employs a GRU module after CNN to introduce the temporal information. In [6, 16], a 3D ConvNet is directly trained to detect spatial and temporal artifacts. Some studies introduce prior knowledge to benefit video face forgery detection, such as eye blinking [33], lip motion [25], and emotion [38]. Amerini *et al*. [6] suggest that predicting optical flow between frames helps deepfake detection.

There are A part of works [19, 21, 54] which tend to focus on representation learning. STIL [21] considers both the spatial and temporal inconsistency and designs a spatiotemporal inconsistency Learning framework for deepfake video detection. RealForensics [19] introduces audio information and leverages self-supervised learning for representation learning. A recent work FTCN [54] explores directly training a fully temporal 3D ConvNets with an attached temporal Transformer. However, detecting without spatial information may harm the generalization capability. In this work, we aim to bring both spatial and temporal features for more general face forgery detection.

2.3. Generalization to Unseen Manipulations

With the rapid development of face generation and manipulation techniques, many previous face forgery detection methods [10, 42, 50] cannot well address unseen manipulations and datasets. Recent studies have noticed this challenge and focus on improving the generalization capability of the model. FWA [35] targets the artifacts in affine face warping as the distinctive feature to detect the forgery. Face X-ray [32] proposes that detecting blending boundaries in images can make a general detector, which sets up a new paradigm of synthesizing images for generalizable face forgery detection. SBI [44] inherits the detecting boundaries thought proposed by Face X-ray [32] and suggests that

blending from single pristine images is more suitable. Another work SLADD [20] proposes to dynamically synthesize forged images by adversarial learning.

Besides image-level face forgery detection, there are also works [19, 21, 25, 54] paying attention to video-level face forgery detection. LipForensics [25] uses a network pretrained on a LipReading dataset [12] and then makes a prediction based on the mouth region, which relies on audio data. RealForensics [19] also introduces audio information and leverages self-supervised learning to learn a better representation of forgery discrimination. FTCN [54] takes full advantage of temporal incoherence to detect the forged videos, based on an assumption that detecting forgeries in the temporal dimension is more general. In this work, we make no assumption or hypothesis. Instead, we design a novel training strategy to make full use of spatial and temporal information to make a prediction without extra data.

2.4. Data Synthesis for Face Forgery Detection

Synthesizing or augmenting data is a classic method to improve the diversity and amount of training datasets in deep learning. In face forgery detection, several works start from the data synthesis viewpoint to seek a more general detector. FWA [35] proposes to synthesize fake data by blurring facial regions based on the assumption that current deepfake algorithms can only generate images of limited resolutions. Face X-ray [32], I2G [53], SLADD [20], and SBI [44] propose to synthesize fake images by blending two images based on the thought of most manipulated images may produce blending boundary artifacts. Although those blending artifact detection methods achieve promising performance on generalization experiments, until recently, there is not a very effective video-level data synthesis method in face forgery detection. In this work, we aim to design video-level data augmentation methods which are more suitable for encouraging spatiotemporal networks to learn better spatial and temporal representation.

3. Method

3.1. Motivation

Artifacts in forged face images can be roughly divided into two types: spatial-related (*e.g.*, generative artifacts, blending, and *etc.*) and temporal-related artifacts (*e.g.*, flickering and discontinuity). Earlier works [8, 10, 11] mostly focus on searching spatial artifacts. These artifacts can be easily captured by training a deep neural network. However, these image-level face forgery detection methods do not have the capability of capturing temporal-level artifacts.

With the demand for detecting more challenging forgeries, research on how to detect video-level face forgeries attracts more and more attention. Researchers seek to leverage video-level artifacts for detecting fake videos. Among them, the typical works are LipForensics [25] and FTCN [54]. They achieve impressive results in detecting temporal-level artifacts like unnatural lip motion or temporal incoherence. However, they have a strong assumption that focusing on temporal incoherence contributes to a more general detector. Indeed, currently most face manipulation and generation methods [31, 45–49, 57] generate forged videos in a frame-by-frame manner, yielding flicking and incoherent artifacts. However, few but not none, there are also video-level manipulation and generation methods [5, 23, 52], which can produce videos that are more coherent perceptually, making these temporal-based detectors difficult to handle. On the other hand, these generative methods still contain spatial-level artifacts to some extent. Hence, it is urgent to develop a general face forgery detector for capturing both spatial and temporal artifacts.

Algorithm 1 Pseudocode of AltFreezing in Pytorch.

```
# F: a 3D spatiotemporal network
# V, y: video clips, labels
# I_s, I_t: iterations of freezing spatial, temporal
    kernels

def st_optimizer(network):
   # splitting params into
   # spatial-related and temporal-related
   params_s, params_t = st_split(network)
   # alternate optimizer
   return SGD(params_s,...), SGD(params_t,...)

count = 0
optim_s, optim_t = st_optimizer(F)
for V, y in loader: # load a minibatch
   optim_t.zero_grad() # zero gradient
   optim_s.zero_grad() # zero gradient
   V = aug(V) # random augmentation
   pred = F(V) # network prediction
   loss = CrossEntropyLoss(pred, y) # compute loss
   loss.backward() # compute gradient
   if count %(I_s+I_t)<I_s: # spatial freezing
      optim_t.step() # temporal optimization
   else: # temporal freezing
      optim_s.step() # spatial optimization
   count+=1
```

A spatial-temporal network is theoretically capable of capturing both spatial and temporal artifacts. However, we observe that if we naïvely train a spatiotemporal network with a binary classification loss, the network will rely on one type of "easy" artifact to distinguish real or fake. It makes the detector cannot completely to find all the artifacts for classification. This will cause the detector to have a terrible generalization capability to unseen deepfake datasets or manipulation methods.

To address this issue, we propose a novel training strategy named *AltFreezing*, to encourage the model to capture both spatial and temporal artifacts. We assume that capturing both the spatial- and temporal-level artifacts can yield a strong generalization capability for unseen datasets and manipulation methods.

Moreover, we notice that data augmentation plays an increasingly important role in improving the generalization

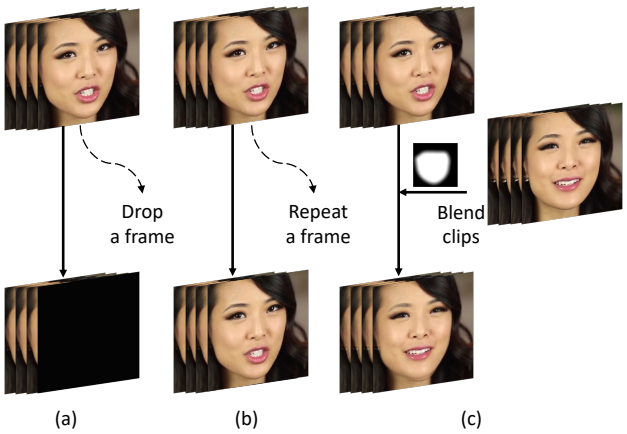

Figure 2. **Illustration the proposed Fake Clip Generation.** For each video clip during training, we randomly use a) temporal dropout, b) temporal repeat, and c) clip-level blending to generate a fake clip for generating fake samples. Temporal dropout and repeat can introduce fake clips with challenging temporal incoherence. Clip-level blending can generate fake clips which only contain challenging spatial artifacts.

capability of face forgery detection. However, most of them are at the image level. To encourage the spatiotemporal network to encompass a strong generalization capability, we further introduce some video-level augmentation techniques.

3.2. AltFreezing

Our AltFreezing is a simple modification to the standard spatiotemporal network updating mechanism. It first divides the weights of the network into two groups. Then the two groups of weights are alternately frozen during the training process. In other words, AltFreezing updates the weights of two groups in turn. Take a typical 3D ConvNet, 3D ResNet-50 (R50) [9] as an example. The convolutional weights of 3D R50 can be mainly divided into spatial-based (*i.e.*, $1 \times K_h \times K_w$ convolutional kernels) and temporal-based (*i.e.*, $K_t \times 1 \times 1$ convolutional kernels). Note that, $1 \times 1 \times 1$ convolutional layers, linear layers, batch normalization layers, and other modules with parameters are regarded as both related considering these layers do not have a receptive field on both temporal and spatial dimensions. After splitting, AltFreezing starts to freeze the two groups of weights alternately during the training stage. Specifically, when the spatial-related weights are frozen, the network will strive to search temporal artifacts to distinguish between real and fake. Similarly, when the temporal-related weights are frozen, the network will struggle to search spatial artifacts to discriminate between real and fake.

Suppose that the weights θ of the spatiotemporal network F are divided into θ_S and θ_T. Given training data (input video clips V, real/fake labels y), our goal of training a video face forgery detector is to minimize the loss function $\mathcal{L}(\mathbf{F}(V;\theta_S,\theta_T),y)$ by optimizing the weights θ_S and θ_T of spatiotemporal network $\mathbf{F}$. The update of θ_S is formulated as follows,

$$\theta_S \leftarrow \theta_S - \alpha \frac{\partial \mathcal{L}(\mathbf{F}(I;\theta_S,\theta_T),y)}{\partial \theta_S}, \tag{1}$$

where α is the learning rate. Correspondingly, the update of θ_T is formulated as follows,

$$\theta_T \leftarrow \theta_T - \alpha \frac{\partial \mathcal{L}(\mathbf{F}(I;\theta_S,\theta_T),y)}{\partial \theta_T}. \tag{2}$$

Moreover, we can control the ratio of iterations in freezing spatial weights and temporal weights I_s:I_t to encourage the network to pay more attention to spatial or temporal artifacts. Within a cycle, we first freeze the spatial weights I_s iterations then we freeze the temporal weights I_t iterations. We find that spatial artifacts are usually easy to learn, so I_s is set to be larger than I_t. The whole algorithm of AltFreezing is summarised in Algorithm 1. With the help of the alternately freezing strategy, the network cannot easily converge by only focusing on one single type of artifact. By switching between spatial and temporal weights, the final trained network is enabled with an ability to capture both spatial and temporal artifacts for more general face forgery detection.

3.3. Fake Clip Generation

Some recent methods [15, 32, 35, 44] leverage data augmentations to encourage a more general representation learning for detecting face forgeries. However, these augmentations are only at the image level. Until recently in the face forgery detection area, little attention is paid to video-level augmentations which are actually more compatible with 3D ConvNets. To learn better video-level representation, we propose a set of fake video synthetic methods including temporal-level and spatial-level augmentations.

As shown in Fig. 2, we propose three video-based augmentations, *i.e.*, temporal dropout, temporal repeat, and clip-level blending. The first two types are temporal-related fake clip generation. Temporal dropout (Fig. 2 (a)) means one or multiple random frames of the video clip are dropped, which is a strong imitation of cutting video frames. After dropping, frames after the dropped frames are shifted forward, and the empty frames are set to 0. For performing temporal repeat (Fig. 2 (b)), one or multiple random frames are repeated, which is a strong simulation of inserting frames into an original video. Then the frames after the repeated frames are shifted backward, and the extra frames are removed. These two temporal-based augmentations can help the spatiotemporal network to capture temporal artifacts.

On the contrary, the proposed clip-level blending (Fig. 2 (c)) is spatial-related fake clip generation. Concretely, we first randomly choose two clips from a single video or two videos, in which one serves as the foreground clip V_f and the other serves as the background clip V_b, After

that, we generate a random mask M delimiting the manipulated region, with each pixel having a greyscale value between 0.0 and 1.0. Then we use the mask to blend each frame from the foreground clip to its corresponding frame of the background clip by:

$$V^i = V_f^i * M + V_b^i * (1 - M), \quad (3)$$

where $i = 1, 2, \cdots, L$ is the i-th frame of the clip, L is the length of the clip. Since V_f and V_b are real clips that are temporally coherent, the resulting clip V is also temporally coherent since the blend operation does not corrupt temporal coherence. Thus V only contains spatial-related artifacts, *i.e.*, the blending boundary around the mask M. Our method is different from the previous image-level blending methods [20, 32, 44], which process each image independently yielding temporal incoherence.

Incorporating the video-level fake clip augmentations with AltFreezing, our spatiotemporal network can capture more general spatial and temporal artifacts for face forgery detection. Finally, our model is trained with a simple binary cross-entropy loss, which is formulated as follows,

$$\mathcal{L}(\widetilde{y}, y) = -\sum_{i=1}^{N}(y^i * \log(\widetilde{y}^i) + (1 - y^i) * \log(1 - \widetilde{y}^i)), \quad (4)$$

where N denotes mini-batch size, y is the label, and $\widetilde{y}$ is the prediction of the network.

4. Experiment

4.1. Setup

Datasets. (1) **FaceForensics++** (FF++) [42] consists of 1,000 real videos and 4,000 fake videos. The fake videos are generated by four manipulation methods (Deepfake (DF) [2], Face2Face (F2F) [47], FaceSwap (FS) [3], NeuralTexture(NT) [45]). (2) **CelebDF v2** (CDF) [36] is a new face-swapping dataset including 5,639 synthetic videos and 890 real videos downloaded from YouTube. In our experiments, its 518 testing videos are used for evaluation. (3) **Deepfake Detection** (DFD) [1] contains over 3,000 manipulated videos from 28 actors in various scenes. (4) **FaceShifter** (FSh) [31] and **DeeperForensics** (DFo) [30] generate high-fidelity face-swapping videos based on the real videos from FF++. In our experiments, we use the training split of FF++ as the training data by default. Unless stated otherwise, we use the c23 version of FF++, following recent literatures [25, 54].

Evaluation Metrics. Following previous methods [25, 32, 44, 54], we report the area under the receiver operating characteristic curve (AUC) to evaluate the performance. Following [25, 54], we report video-level AUC for fair comparisons. And for image-based methods, we average the frame-level predictions as the corresponding video-level prediction.

Method	CDF	DFD	FSh	DFo	Avg
Xception [42]	73.7	–	72.0	84.5	–
CNN-aug [50]	75.6	–	65.7	74.4	–
PatchForensics [10]	69.6	–	57.8	81.8	–
Multi-task [39]	75.7	–	66.0	77.7	–
FWA [34]	69.5	–	65.5	50.2	–
Two-branch [37]	76.7	–	–	–	–
Face X-ray [32]	79.5	95.4	92.8	86.8	88.6
SLADD [20]	79.7	–	–	–	–
SBI-R50* [44]	85.7	94.0	78.2	91.4	87.3
CNN-GRU [43]	69.8	–	80.8	74.1	–
STIL [21]	75.6	–	–	–	–
LipForensics-Scratch [25]	62.5	–	84.7	84.8	–
LipForensics [25]	82.4	–	97.1	97.6	–
RealForensics-Scratch [19]	69.4	–	87.9	89.3	–
RealForensics [19]	86.9	–	**99.7**	**99.3**	–
FTCN [54]	86.9	94.4	98.8	98.8	94.7
AltFreezing (ours)	**89.5**	**98.5**	99.4	**99.3**	**96.7**

Table 1. **Generalization to unseen datasets.** We report the video-level AUC (%) on four unseen datasets: Celeb-DF-v2 (CDF), DeepFake Detection (DFD), FaceShifter (FSh), and DeeperForensics (DFo). The models are trained on FF++ and tested on these unseen datasets. * denotes our reproduction with the official code, due to its unfair experiments using the raw version of training data of FF++. The results of other methods are from [25].

Model	#params	Arch	FSh	DFo	Avg
LipForensics-Scratch [25]	36.0M	R18+MS-TCN	84.7	84.8	84.8
LipForensics [25]	36.0M	R18+MS-TCN	97.1	97.6	97.4
FTCN [54]	**26.6M**	3D R50+TT	98.8	98.8	98.8
AltFreezing (ours)	27.2M	3D R50	**99.4**	**99.3**	**99.4**

Table 2. **Comparison with video-level state-of-the-art methods in terms of parameters and architectures.** Video-level AUC(%) is reported on FSh and DFo using models trained on FF++ [42]. "MS-TCN" means multi-scale temporal convolutional network. "TT" means temporal Transformer. Note that, 3D R50 used in FTCN [54] is without spatial kernels.

Implementation details. We use 3D ResNet50 [9] as our network and train it for 1k epochs with the SGD optimizer. The batch size is 32. We random sample concussive 32 frames of each video during training. The initial learning rate is 0.05, decayed to 0 at the ending epoch following the curve of cosine annealing. For data augmentations, RandomHorizontalFlip, RandomCutOut, and AddGaussianNoise are also applied besides the proposed video-level fake video synthetic augmentations.

4.2. Generalization to Unseen Datasets

In real-world scenarios, there is usually a gap between the tested forged videos and fake videos from the training dataset. Therefore, the generalization capability of models to unseen datasets is critical. To evaluate the generalization ability of our model, we use the original videos and all four types of fake videos in FF++ [42] as the training data, then evaluate the performance on CDF [36], DFD [1], FSh [31],

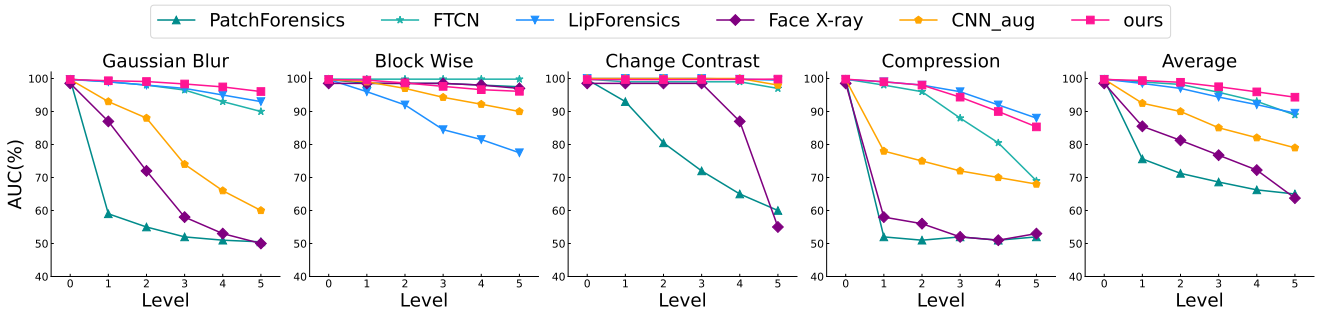

Figure 3. **Robustness to unseen perturbations.** Video-level AUC(%) is reported under five different degradation levels of four particular types of perturbations [30]. "Average" AUC score denotes the mean of each perturbation at each level.

and DFo [30].

We report the AUC results in Tab. 1. We observe that our model achieves the best performance on CDF (89.5%), DFD (98.5%), and DFo (99.3%), and competitive performance on FSh (99.4%). It is worth noting that all other methods perform unsatisfactorily on CDF, while our model obtains an AUC of 89.5%, which shows our model's strong generalization capability. And observing from the average AUC comparison, our method achieves a significant improvement of 2% AUC score compared to previous video-level state-of-the-art method [54], 94.7% $\rightarrow$ 96.7%. Moreover, we also compare the parameters and architectures in Tab. 2. We observe that our method achieves the best performance with a simple 3D R50 network, which further indicates that our model is a simple but more general face forgery detector.

4.3. Generalization to Unseen Manipulations

For a general face forgery detector, they usually do not know which manipulation is applied to the tested videos. It is important to have a strong generalization capability to unseen manipulations. Following previous works [25, 54], we conduct the experiments on FF++ [42] with a leave-one-out setting. There are four types of forged face videos, *i.e.*, DF, F2F, FS, and NT in FF++. We choose three of the forgery subsets as the training set. The remaining subset is used for evaluating the generalization capability of the model.

In Tab. 3, we show comparisons of our method with other state-of-the-art methods. The AUC scores demonstrate that our model can achieve impressive performance on the whole FF++ test set (average AUC: 98.6%), especially on the subsets DF (99.8%) and FS (99.7%) compared to previous methods. On F2F and NT, our results of ours are slightly lower than LipForensics [25]. One possible explanation is that LipForensics [25] employs a pre-trained model with a strong prior knowledge of the mouth region, which is beneficial for unseen manipulation detection. While our model is trained from scratch, without using pre-trained models. Nonetheless, our model achieves better performance over LipForensics in terms of the average AUC on the four types of manipulation methods.

Method	Train on remaining three				
	DF	FS	F2F	NT	Avg
Xception [42]	93.9	51.2	86.8	79.7	77.9
CNN-aug [50]	87.5	56.3	80.1	67.8	72.9
PatchForensics [10]	94.0	60.5	87.3	84.8	81.7
Face X-ray [32]	99.5	93.2	94.5	92.5	94.9
CNN-GRU [43]	97.6	47.6	85.8	86.6	79.4
LipForensics-Scratch [25]	93.0	56.7	98.8	98.3	86.7
LipForensics [25]	99.7	90.1	**99.7**	**99.1**	97.1
FTCN* [54]	**99.8**	99.6	98.2	95.6	98.3
AltFreezing (ours)	**99.8**	**99.7**	98.6	96.2	**98.6**

Table 3. **Generalization to unseen manipulations.** We report the video-level AUC (%) on the FF++ dataset, which consists of four manipulation methods (DF, FS, F2F, NT). The experiments obey the leave-one-out rule as [25, 54]. The three subsets of fake videos are used as the training data, the other one serves as the testing data. * denotes our reproduction without a temporal Transformer. The results of other methods are from [25].

4.4. Robustness to Unseen Perturbations

Besides the generalization to unseen datasets and manipulations, the robustness to unseen perturbations is also a concerning problem in real-world scenes. Following [30], we evaluate the robustness of our model to unseen perturbations considering four different degradation types, *i.e.*, Gaussian blur, Block-wise distortion, Change contrast, and Video compression. Each perturbation is operated at five levels to evaluate the robustness of models under different-level different-type distortion. We show the AUC results on these unseen perturbations in Fig. 3, using the model trained on FF++. We observe that our method outperforms previous methods a lot at every level on average, which indicates that our method is much more robust and generalizable. Especially on serious degradations (level 4, 5 in Fig. 3), the AUC of our model surpasses others a lot, *i.e.*, about 3% improvement on level 4 and 4% improvement on level 5.

4.5. Ablation Studies

In this section, we perform ablation studies to verify the effectiveness of the proposed AltFreezing. We do not utilize

Model	Train on FF++			Avg
	FF++	CDF	FSh	
3D R50	99.3	81.8	99.2	93.4
3D R50 (freeze S. kernels)	99.5	76.8	98.9	91.7
3D R50 (freeze T. kernels)	99.4	80.6	**99.4**	93.1
3D R50 (AltFreezing)	**99.7**	**86.4**	99.3	**95.1**

Table 4. **Ablation study of variants of AltFreezing.** Video-level AUC(%) is reported. "S." means spatial and "T." means temporal.

Model	Train on FF++		Avg
	Temporal Set	Spatial Set	
FTCN [54]	74.8	75.8	75.3
3D R50	76.5	71.5	74.0
3D R50 (AltFreezing)	**80.6**	**84.5**	**82.6**

Table 5. **Effect of AltFreezing when facing more hard cases.** Video-level AUC(%) is reported on our synthetic datasets. The temporal Set is a synthetic dataset with only temporal incoherence. Spatial Set is a synthetic dataset with only spatial artifacts. The models are all trained on FF++ [42].

fake clip generation techniques without special notations. All the models are trained on FF++ [42], and tested in FF++ [42], CelebDF v2 [36], and FaceShifter [31] to evaluate the generalization capability of the models.

Effect of AltFreezing. We design several variants of AltFreezing, We use the 3D Resnet50 [9] (3D R50 for short) as the basic network structure. 1) a vanilla 3D R50 network without any change of network structure or training strategy. 2) In 3D R50 (freeze S. kernels), we split the weights of 3D R50 into spatial-related and temporal-related as AltFreezing. And during training, we fix all the spatial kernels and only update the weights of temporal kernels. 3) Similar to 2), in 3D R50 (freeze T. kernels), we only update the weights of spatial kernels. 4) the proposed AltFreezing with Resnet50 as the backbone.

We report the AUC results of these models in Tab. 4. Compared with the 3D R50 baseline, our proposed AltFreezing training strategy can significantly improve the performance of in-domain face forgery detection and out-domain face forgery detection. We also notice that simply freezing the spatial or temporal weights of the 3D R50 network can not obtain a performance gain and even damage the performance. So the key design is alternately freezing the spatial and temporal weights during training. Moreover, AltFreezing is a plug-and-play training strategy for capturing both spatial and temporal artifacts, which does not introduce any extra computation or parameters.

Does AltFreezing really learn how to capture spatial and temporal artifacts? Although our motivation for AltFreezing is learning to search both spatial and temporal artifacts, the readers might wonder whether our AltFreezing can really achieve this goal. We conduct experiments in more challenging scenes to verify the ability of our model to capture spatial and temporal artifacts. We build two new challenging

Freezing ratio ($I_s : I_t$)	Train on FF++			Avg
	FF++	CDF	FSh	
baseline	99.3	81.8	99.2	93.4
1:1	99.6	82.4	99.2	93.7
5:1	99.5	82.8	**99.7**	94.0
10:1	99.6	83.4	99.1	94.0
20:1	**99.7**	**86.4**	99.3	**95.1**
30:1	99.5	82.1	99.2	93.6

Table 6. **Ablation study of the freezing ratio of AltFreezing.** Video-level AUC(%) is reported. "baseline" means a 3D R50 with end-to-end training.

Aug. level	Train on FF++			Avg
	FF++	CDF	FSh	
none	**99.7**	86.4	99.3	95.1
image	99.6	78.6	**99.4**	92.5
video	**99.7**	**89.5**	99.3	**96.2**

Table 7. **image-level augmentation [44] *vs*. video-level augmentation.** Video-level AUC(%) is reported. "Aug." means augmentation. "image" level augmentation means that the blending used is image level like [32, 44] while keeping other augmentations unchanged. "video" level augmentation means that using the proposed video-level augmentation methods.

Backbone	Train on FF++			Avg
	FF++	CDF	FSh	
3D R50	**99.7**	89.5	99.3	96.2
3D R101	99.6	**90.4**	**99.4**	**96.5**

Table 8. **Different backbone architectures.** Video-level AUC (%) is reported. The models are trained on FF++.

datasets based on the testing set of real data in FF++ [42]. 1) Temporal Set: we aim to build a test set that only contains temporal-related artifacts, we randomly drop or repeat frames from a real video clip that all frames are real with only temporal incoherence introduced. 2) Spatial Set: we aim to build a test set that only contains spatial-related artifacts, we use a random mask to extract all the same region from a clip, then blend the region back into the other clip, since each pixel of these two clips is coherent, the newly generated clip is temporal coherent with only spatial artifacts. It is worth noting that we do not use the proposed fake video augmentation methods described in Sec. 3.3 in this experiment, in order to evaluate the performance of the proposed AltFreezing training strategy.

The evaluation results on these two hand-crafted datasets are reported in Tab. 5. Even though our model does not see any types of artifacts in the Temporal Set and Spatial Set during the training stage. It achieves strong performance on these test sets. Compared with FTCN, which is specially designed for detecting temporal artifacts, our AltFreezing achieves a better performance in detecting temporal artifacts. This validates that spatial convolution is also important for detecting temporal artifacts, restricting all the spatial kernels to 1 is not an optimal choice.

Influence of the freezing ratio in AltFreezing. In the AltFreezing algorithm, the ratio of iterations in freezing spatial

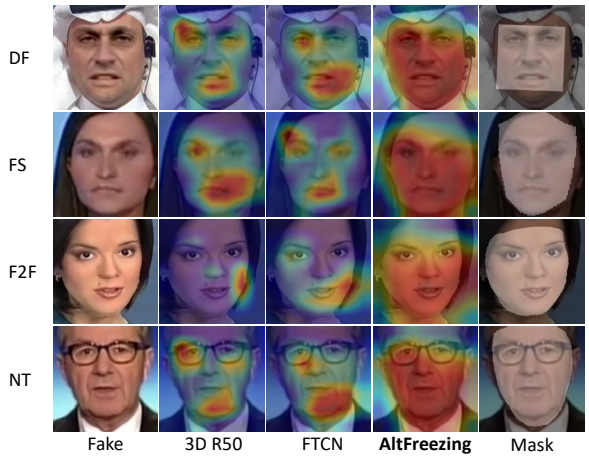

Figure 4. **Visualization of activation maps for fake samples from different manipulation methods.** Warmer color indicates a higher probability of fake. We compare vanilla 3D R50, FTCN, and 3D R50 with our AltFreezing strategy. The activation maps shown here are the mixing of activation heatmaps and the input fake frames. Our AltFreezing could locate the forgery region precisely.

weights and temporal weights I_s:I_t controls the capability of the network on handling spatial-related and temporal-related artifacts. We conduct a comprehensive study about the effect of the freezing iterations ratio. We consider the freezing ratio in the set {1:1, 5:1, 10:1, 20:1, 30:1}, and train our models with a different freezing ratio while keeping other configurations the same.

Tab. 6 shows that AltFreezing's performance initially increases and then decreases as the freezing ratio varies from 1:1 to 30:1. The model achieves the best generalization capability when the freezing ratio is 20:1. It is worth noting that AltFreezing is better than baseline (without AltFreezing) on the generalization ability of the model at all freezing ratios. Recent works FTCN [54] and LipForiensics [25] suggest that temporal information is more important for 3D networks to make a prediction. And combined with our analysis, spatial information may serve as a complement to temporal information for detecting face forgeries. So in our experiments, the freezing iterations of spatial-based kernels are more than temporal-based ones for the 3D ConvNet to capture more temporal artifacts and also involve spatial information.

Effect of the fake clip generation. The proposed fake clip generation method is based on three video-level augmentations to encourage a more general representation learning of video-level forgeries. Here, we study the effect of the data synthesis method. As shown in Tab. 7, our video-level augmentations bring an average AUC improved from 95.1 % to 96.2 %. Especially on CDF [36], with our fake video augmentations, our AltFreezing method gets a +3.1 % absolute boost, 86.4% $\rightarrow$ 89.5%. This indicates that the proposed video-level fake sample synthesis benefits the generalization ability of the network a lot. We also compare our augmentations with recent self-blending image augmentation [44]. We find frame-level augmentation is not suitable to be directly applied to a spatiotemporal network for general face forgery detection. For more detailed experiments and discussions of fake clip generation, please refer to the supplemental material.

Results of advanced architectures. We further conduct an experiment about the effect of more advanced network architectures, as shown in Tab. 8. Using 3D R101 as the backbone network brings further improvement compared to using 3D R50, indicating that a better backbone network yields better detection performance.

Visualization of the captured artifacts. For a more in-depth understanding of how AltFreezing works, we further use Classification activation maps (CAM) [55] to localize which regions are activated to detect artifacts. The visualization results are shown in Fig. 4. Neither the vanilla 3D R50 nor the FTCN [54] can notice the precise regions that are indeed manipulated. 3D R50 focuses on a very limited area for discrimination, which confirms that nïave training a 3D ConvNet leads to a trivial solution. FTCN [54] pays more attention to locations outside of forged areas compared to 3D R50. In contrast, our AltFreezing makes it discriminates between real and fake by focusing predominantly on the manipulated face area. This visualization further identifies that AltFreezing encourages the 3D ConvNet to capture more spatial and temporal artifacts.

5. Conclusion and Discussion

In this paper, we seek to capture both spatial and temporal artifacts in one model for more general face forgery detection. Concretely, we present a training strategy called AltFreezing that separates the spatial and temporal weights into two groups and alternately freezes one group of weights to encourage the model to capture both the spatial and temporal artifacts. Then, we propose a set of video-level fake data augmentations to encourage the model to capture a more general representation of different manipulation types. Extensive experiments verify the effectiveness of the proposed AltFreezing training strategy and video-level data augmentations. We hope that our work can attract more attention to video-level representation learning in the face forgery detection community.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Contract 61836011 and 62021001, and in part by the Fundamental Research Funds for the Central Universities under contract WK3490000007. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution and the Supercomputing Center of USTC.

References

[1] Contributing data to deepfake detection research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. Accessed: 2021-11-13. 5

[2] Deepfakes. https://github.com/deepfakes/faceswap. [Accessed: 2020-09-02]. 5

[3] Faceswap. https://github.com/MarekKowalski/FaceSwap. [Accessed: 2020-09-03]. 5

[4] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *WIFS*, pages 1–7, 2018. 2

[5] Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third time's the charm? image and video editing with stylegan3. *arXiv preprint arXiv:2201.13433*, 2022. 3

[6] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *ICCVW*, pages 0–0, 2019. 2

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

[8] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *ACM IHMS Workshop*, pages 5–10, 2016. 2, 3

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 2, 4, 5, 7

[10] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. *arXiv preprint arXiv:2008.10588*, 2020. 1, 2, 3, 5, 6

[11] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017. 2, 3

[12] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *ACCV*, pages 87–103, 2016. 3

[13] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *ICCV*, pages 15108–15117, 2021. 2

[14] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *CVPR*, pages 5781–5790, 2020. 2

[15] Sowmen Das, Selim Seferbekov, Arup Datta, Md Islam, Md Amin, et al. Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation. In *ICCV*, pages 3776–3785, 2021. 4

[16] Oscar de Lima, Sean Franklin, Shreshtha Basu, Blake Karwoski, and Annet George. Deepfake detection using spatiotemporal convolutional networks. *arXiv preprint arXiv:2006.14749*, 2020. 2

[17] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *CVPR*, pages 9468–9478, 2022. 2

[18] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Identity-driven deepfake detection. *arXiv preprint arXiv:2012.03930*, 2020. 2

[19] Alexandros *et al*. Leveraging real talking faces via self-supervision for robust forgery detection. In *CVPR*, 2022. 2, 3, 5

[20] Chen *et al*. Self-supervised Learning of Adversarial Examples: Towards Good Generalizations for DeepFake Detections. In *CVPR*, 2022. 2, 3, 5

[21] Gu *et al*. Spatiotemporal inconsistency learning for deepfake video detection. In *ACMMM*, 2021. 2, 3, 5

[22] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *TIFS*, 7(3):868–882, 2012. 2

[23] Tsu-Jui Fu, Xin Eric Wang, Scott T Grafton, Miguel P Eckstein, and William Yang Wang. M3l: Language-based video editing via multi-modal multi-level transformers. In *CVPR*, pages 10513–10522, 2022. 3

[24] Wu Haiwei, Zhou Jiantao, Zhang Shile, and Tian Jinyu. Exploring spatial-temporal features for deepfake detection and localization. *arXiv preprint arXiv:2210.15872*, 2022. 2

[25] 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. 1, 2, 3, 5, 6, 8

[26] Chih-Chung Hsu, Chia-Yen Lee, and Yi-Xiu Zhuang. Learning to detect fake face images in the wild. In *IS3C*, pages 388–391, 2018. 2

[27] Ashraful Islam, Chengjiang Long, Arslan Basharat, and Anthony Hoogs. Doa-gan: Dual-order attentive generative adversarial network for image copy-move forgery detection and localization. In *CVPR*, 2020. 2

[28] Yonghyun Jeong, Doyeon Kim, Seungjai Min, Seongho Joe, Youngjune Gwon, and Jongwon Choi. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *WACV*, pages 48–57, 2022. 2

[29] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. Frepgan: Robust deepfake detection using frequency-level perturbations. *arXiv preprint arXiv:2202.03347*, 2022. 2

[30] 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 2886–2895, 2020. 1, 5, 6

[31] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *CVPR*, pages 5074–5083, 2020. 1, 3, 5, 7

[32] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *CVPR*, pages 5001–5010, 2020. 1, 2, 3, 4, 5, 6, 7

[33] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *WIFS*, pages 1–7, 2018. 2

[34] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2, 2018. 1, 5

[35] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *CVPRW*, 2019. 2, 3, 4

[36] 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. 5, 7, 8

[37] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *ECCV*, pages 667–684, 2020. 2, 5

[38] 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. 2

[39] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *BTAS*, pages 1–8, 2019. 1, 2, 5

[40] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, pages 86–103, 2020. 2

[41] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *WIFS*, pages 1–6, 2017. 2

[42] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 1, 2, 5, 6, 7

[43] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *GUI*, 3(1), 2019. 1, 2, 5, 6

[44] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *CVPR*, pages 18720–18729, 2022. 1, 2, 3, 4, 5, 7, 8

[45] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, pages 716–731, 2020. 1, 3, 5

[46] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM TOG*, 38(4):1–12, 2019. 1, 3

[47] 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, 2016. 1, 3, 5

[48] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven realistic facial animation with temporal gans. In *CVPRW*, pages 37–40, 2019. 1, 3

[49] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *IJCV*, pages 1–16, 2019. 1, 3

[50] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, volume 7, 2020. 1, 2, 5, 6

[51] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *CVPR*, 2019. 2

[52] Zhiliang Xu, Zhibin Hong, Changxing Ding, Zhen Zhu, Junyu Han, Jingtuo Liu, and Errui Ding. Mobilefaceswap: A lightweight framework for video face swapping. *arXiv preprint arXiv:2201.03808*, 2022. 3

[53] 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. 3

[54] 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. 1, 2, 3, 5, 6, 7, 8

[55] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 8

[56] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, volume 33, pages 9299–9306, 2019. 1

[57] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *CVPR*, pages 4834–4844, 2021. 3