Deepfake Video Detection with Spatiotemporal Dropout Transformer

Daichi Zhang
Institute of Information Engineering, Chinese Academy of Science
School of Cyber Security, University of Chinese Academy of Sciences
zhangdaichi@iie.ac.cn

Fanzhao Lin
Institute of Information Engineering, Chinese Academy of Science
School of Cyber Security, University of Chinese Academy of Sciences
linfanzhao@iie.ac.cn

Yingying Hua
Institute of Information Engineering, Chinese Academy of Science
School of Cyber Security, University of Chinese Academy of Sciences
huayingying@iie.ac.cn

Pengju Wang
Institute of Information Engineering, Chinese Academy of Science
School of Cyber Security, University of Chinese Academy of Sciences
wangpengju@iie.ac.cn

Dan Zeng
School of Communication and Information Engineering, Shanghai University
dzeng@shu.edu.cn

Shiming Ge*
Institute of Information Engineering, Chinese Academy of Science
School of Cyber Security, University of Chinese Academy of Sciences
geshiming@iie.ac.cn

ABSTRACT

While the abuse of deepfake technology has caused serious concerns recently, how to detect deepfake videos is still a challenge due to the high photo-realistic synthesis of each frame. Existing image-level approaches often focus on single frame and ignore the spatiotemporal cues hidden in deepfake videos, resulting in poor generalization and robustness. The key of a video-level detector is to fully exploit the spatiotemporal inconsistency distributed in local facial regions across different frames in deepfake videos. Inspired by that, this paper proposes a simple yet effective patch-level approach to facilitate deepfake video detection via spatiotemporal dropout transformer. The approach reorganizes each input video into *bag of patches* that is then fed into a vision transformer to achieve robust representation. Specifically, a spatiotemporal dropout operation is proposed to fully explore patch-level spatiotemporal cues and serve as effective data augmentation to further enhance model's robustness and generalization ability. The operation is flexible and can be easily plugged into existing vision transformers. Extensive experiments demonstrate the effectiveness of our approach against 25 state-of-the-arts with impressive robustness, generalizability, and representation ability.

CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**.

KEYWORDS

deepfake video detection, spatiotemporal dropout, vision transformer, data augmentation

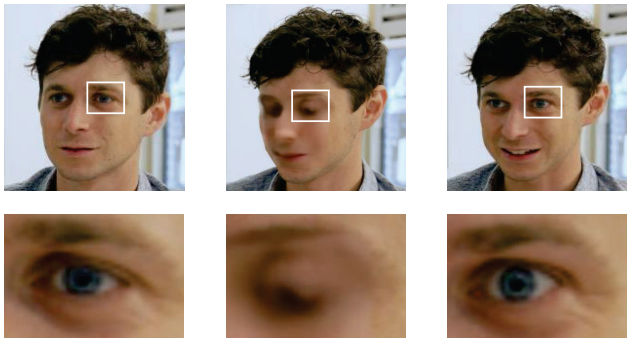

Figure 1: The detection of deepfake videos is challenging due to high photo-realistic frame synthesis. Thus, our approach leverages patch-level spatiotemporal inconsistency in facial regions across frames to facilitate deepfake video detection.

ACM Reference Format:
Daichi Zhang, Fanzhao Lin, Yingying Hua, Pengju Wang, Dan Zeng, and Shiming Ge. 2022. Deepfake Video Detection with Spatiotemporal Dropout Transformer. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Oct. 10–14, 2022, Lisboa, Portugal.* ACM, Lisbon, Portugal, 9 pages. https://doi.org/10.1145/3503161.3547913

1 INTRODUCTION

Deepfake [41] often refers to the technique that generates the images and videos by swapping the faces of source and target persons [25], manipulating the original face attributes [32], or synthesizing an entire face that does not exist [42]. With the rapid development of face generation and manipulation methods, especially after generative adversarial networks (GANs) were proposed [16], deepfake videos can be easily produced by accessible online tools, such as FaceSwap[1] and Deepfakes[2], but can barely be distinguished by the human eyes, leading to significant threaten to the

*Shiming Ge is the corresponding author (geshiming@iie.ac.cn).

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3547913

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

public social, cyber and even political security, such as fabricating evidence and ruining political discourse [41]. Thus, it is very critical to develop effective solutions to detect deepfake videos.

Existing approaches usually formulate the deepfake video detection task as a binary classification problem and can be divided into two major categories: image-level and video-level approaches. Image-level approaches perform frame-wise detection by mining the pattern difference between real and fake images [6, 27, 38, 39]. Generally, these approaches can well exploit the spatial cues in the frames but neglect the temporal cues in the video. Thus, image-level approaches may be limited when detecting deepfake videos. Typically, advanced deepfake approaches may generate extremely genuine facial images without leaving spatial defect but they cannot properly avoid temporal inconsistency since deepfake videos are always generated frame-by-frame. Unlike image-level approaches, recent video-level approaches [2, 17, 50] focus on the sequence patterns and aim to explore the spatiotemporal inconsistency to detect. However, this spatiotemporal inconsistency is distributed dynamically in different local regions and frames, which is extremely difficult to be captured, as shown in Fig. 1, only the eye regions in different frames are visibly inconsistent. Existing video-level approaches are not specially designed for this inconsistency and can not properly capture the spatiotemporal cues which hide in deepfake videos, which makes these approaches poorly generalized and vulnerable, even cannot achieve comparable performance to image-level approaches. Therefore, a key of deepfake video detection approach is to effectively learn the discriminative representations to describe the spatiotemporal inconsistency.

Towards this end, this paper proposes an effective patch-level deepfake video detection framework, named spatiotemporal dropout transformer to capture the spatiotemporal inconsistency effectively. In the approach, the input video is first extracted into facial frame sequence then each frame is grid-wisely cropped into non-overlapping facial patches, which are subsequently reorganized into *bag of patches* and fed into a vision transformer to learn the discriminative representations describing the dynamical spatiotemporal cues in local facial regions across different frames as well as achieving robust representation capacity.

Specifically, a spatiotemporal dropout (STD) operation is designed to fully explore the spatiotemporal inconsistency at patch-level. The STD operation performs temporal dropout and spatial dropout step-wisely. During the temporal dropout, we randomly drop part of the extracted frames after we obtain the frame sequence from input video. Then during the spatial dropout, we randomly drop part of facial patches grid-wisely cropped from each remaining frame and reorganize the remaining patches as a *bag of patches* to train a vision transformer. The bag of patches still contains all facial regions in the original face to preserve the whole information but the data to be processed is largely reduced since the dropout operation. In this way, the inconsistency distributed in local facial patches across different frames in deepfake video is fully represented and explored. And since the dropout operation is random, massive different bag of patches instances can be generated from the same input video in different training iterations, which also serves as data augmentation for more generalized and robust detection. Moreover, the STD operation is flexible and could be plugged into existing vision transformers (ViTs).

Our main contributions can be summarized as three folds. First, we propose an effective patch-level deepfake video detection framework, Spatiotemporal Dropout Transformer. In our approach, input videos are reorganized as *bag of patches* instances which are then fed into a vision transformer to achieve strong representation capacity. Second, we design a simple yet effective spatiotemporal dropout operation, which can fully explore the patch-level spatiotemporal inconsistency hidden in deepfake videos and also serve as an effective data augmentation to further improve the model's robustness and generalization ability. Besides, our STD operation can be plugged into existing ViTs. Third, we conduct extensive experiments on three public benchmarks to demonstrate that our approach outperforms 25 state-of-the-arts with impressive robustness, generalizablity, and representation ability.

2 RELATED WORK

Deepfake Generation. Deepfake refers to the techniques of synthesizing or manipulating human face images or videos[44]. Early deepfake is generated by hand-crafted features designed by researchers such as face landmarks, and some post-processing methods are utilized to make the generating artifacts invisible. For example, [15] designs a face reenactment system based on face matching and [10] proposes a 3D multilinear model to track the face movement in video and minimize the blending boundaries. Since these traditional generation methods often suffer from generating artifacts and visual quality, deep learning-based deepfake generation methods are developed for more realistic face synthesis. For example, generative adversarial networks (GANs) [16] have enabled plenty of high-quality face manipulation [26, 43] and face synthesis [8, 22, 23]. Although these methods can achieve high-quality generation results, they work in a frame-by-frame way to generate deepfake videos, and the spatiotemporal cues are difficult to eliminate, which can serve as an effective discriminative clue.

Deepfake Detection. Since deepfake has brought severe threats to society, a variety of deepfake detection approaches have been proposed. Early detection approaches focus on hand-crafted features which are limited at that time, such as [6, 38] utilize image statistic features to detect. With the development of deep learning, researchers begin to utilize DNNs to perform deepfake detection, which can be further divided into two categories: image-level approaches and video-level approaches. Image-level approaches focus on extracting discriminative image-level features for deepfake detection, such as [39] utilizes the XceptionNet to detect deepfakes and [27] detects the blending boundary of two face images. All these approaches can achieve impressive performance on image-level deepfake detection but ignore the temporal cues hidden in deepfake videos, resulting in poor performance when detecting deepfake videos. Video-level approaches pay more attention to the sequence feature and many general video models have been applied to detect deepfake videos, such as [2, 17, 50]. However, these models usually are not specially designed for deepfake video detection task and are not capable to properly capture the spatiotemporal inconsistency dynamically distributed in local facial regions across different frames. For example, TD-3DCNN [50] only consider the inter-frame inconsistency in frame level while ignoring the intra-frame inconsistency cues in spatial domain.

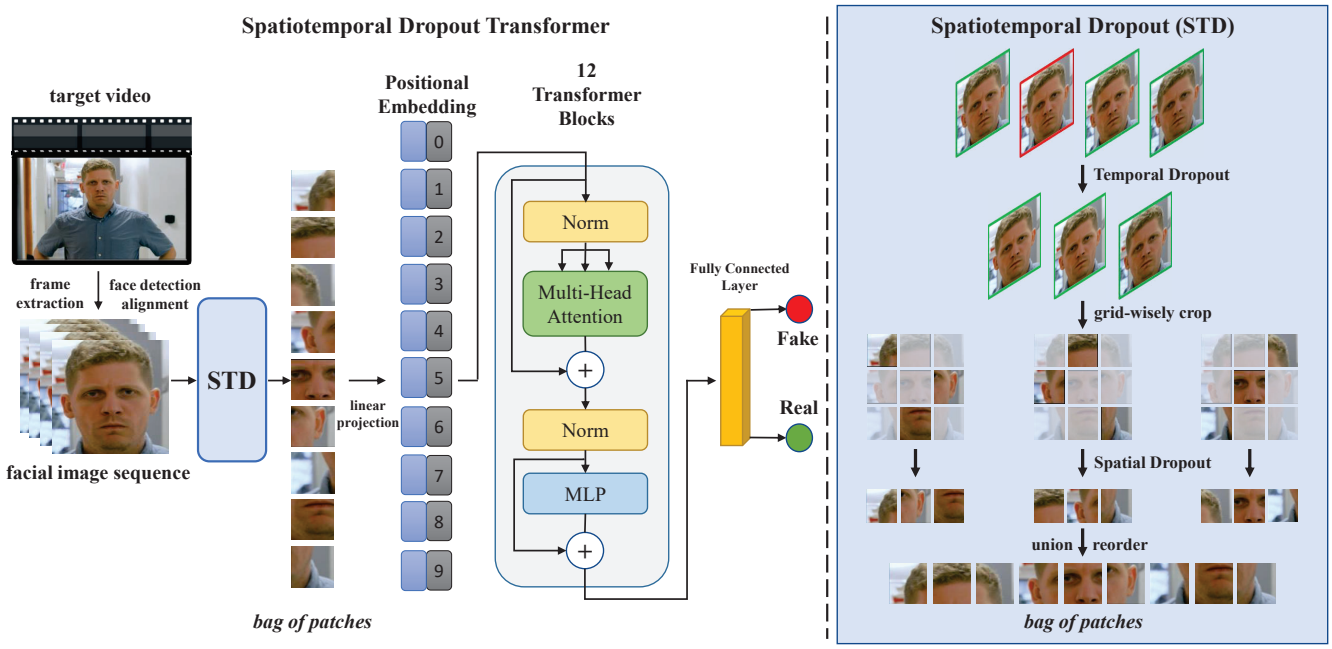

Figure 2: The framework of Spatiotemporal Dropout Transformer for deepfake video detection. We first employ frame extraction, face detection and alignment to get facial image sequence $\mathcal{F} = \{\mathbf{x}_i\}_{i=1}^{n}$. Then each facial image $\mathbf{x}_i \in \mathcal{F}$ is processed by STD operation to get facial patch set $P_{i'} = \{\mathbf{p}_{i',j}\}$ and further reorganized as a *bag of patches* instance $\mathcal{P}$ which is then fed into our vision transformer backbone to learn the discriminative representation and detect.

Vision Transformer. Transformers [45] have achieved impressive performance in natural language processing (NLP) tasks due to their strong representation capacity, such as BERT [11]. Recently, researchers have proved that transformers can also achieve excellent performance on a variety of computer vision tasks. Specifically, vision transformer (ViT) [14] utilizes the same self-attention mechanism and crops an image into a sequence of flattened patches as the input token sequence used in the NLP task to train the transformer encoder for different downstream tasks, such as classification and object detection [4, 14]. Various ViT architectures have been proposed [3, 31, 47] recently and related experiments results further demonstrate ViT also has remarkable representation capacity when dealing with images and videos [19]. Many previous video-level detectors choose traditional CNN as their backbones, which may restrict their performance since CNN's limited representation capacity compared to transformer, such as [50]. To enhance the model's representation capacity in deepfake video detection, more powerful backbones are needed. There are also some existing works that apply ViT to deepfake detection tasks, such as [9, 20, 48], but these approaches focus more on designing ViT architecture or combining with other approaches without exploring the intrinsic characteristics of deepfake video detection task, such as the spatiotemporal cues. Therefore, it is necessary to design an effective and flexible framework to incorporate ViT in deepfake video detection which can make full use of the dynamical cues distributed in local regions across frames (*e.g.*, patch-level spatiotemporal cues in deepfake videos) to further improve model performance.

3 THE APPROACH

3.1 Problem Formulation

In our approach, the objective of deepfake video detection is to learn a discriminative binary classifier ϕ to identify a video clip consisting of n frames $\mathcal{V} = \{\mathbf{f}_i\}_{i=1}^{n}$ into real or fake. Thus, learning ϕ can be formulated as an energy minimization problem that can be solved by:

$$\mathbb{W}^* = \arg\min_{\mathbb{W}} \sum_{\mathcal{V} \in \mathcal{D}} \mathbb{E}(\phi(\mathbb{W}; \mathcal{V}), l), \tag{1}$$

where $\mathbb{W}^*$ is the learned optimal output of the detector parameters, $\mathcal{D}$ is the training set, $l \in \{0, 1\}$ is the video label and $\mathbb{E}$ is an energy function to measure detection loss.

Considering that the spatiotemporal inconsistency existing in deepfake videos is dynamically distributed in different local facial regions across different frames, a key to be addressed by the detector is providing a flexible and exhaustive way to fully explore the spatiotemporal cues and aggregate them to form and learn a typical discriminative features to detect. To achieve that, the detector ϕ should be able to accept dynamical input instances containing spatiotemporal cues to learn a powerful backbone with stronger representation capacity. Towards this end, we take a vision transformer as the backbone and incorporate a simple yet effective Spatiotemporal Dropout operation to fully explore the dynamical spatiotemporal inconsistency across different frames at patch-level, as presented in Fig. 2 and introduced in the following.

3.2 Spatiotemporal Dropout

The Spatiotemporal Dropout operation aims to generate the dynamical input instances from input videos, which contain spatiotemporal inconsistency cues distributed in different facial regions across different frames. Therefore, we perform it in an efficient step-wise manner, including *Temporal Dropout* and *Spatial Dropout*.

Temporal Dropout. For each video, we first randomly sample n consistent raw frames $\{\mathbf{f}_i\}_{i=1}^{n}$ and employ face detection and alignment to each frame $\mathbf{f}_i$ to get a facial image sequence $\mathcal{F} = \{\mathbf{x}_i\}_{i=1}^{n}$. Then we randomly discard part of facial images following a uniform distribution, with defined temporal dropout rate α and remaining $(1-\alpha)\times n$ facial images, which can be formulated as:

$$\begin{aligned} \mathcal{F}_T &= \subseteq (\mathcal{F}, \alpha) \\ &= \subseteq (\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}, \alpha) \\ &= \{\mathbf{x}_{k_1}, \mathbf{x}_{k_1+1}, ..., \mathbf{x}_{k_1+(1-\alpha)*n-1}\}, \\ &\quad k_1 \sim \mathcal{U}(1, \alpha\times n+1), k_1 \in \mathbb{Z}, \alpha \in (0,1) \end{aligned} \tag{2}$$

where $\subseteq$ means the subset operation, α is the temporal dropout rate, k_1 is a random start index following a uniform distribution, and $\mathcal{F}$ is the preprocessed facial image sequence. The uniform distribution can prevent center bias and we finally get a sparse sequence $\mathcal{F}_T$.

Spatial Dropout. For each remaining facial image $\mathbf{x}_{i'} \in \mathcal{F}_T$, we grid-wisely crop $\mathbf{x}_{i'}$ into m regular non-overlapping facial patches $P_{i'} = \{\mathbf{p}_{i',j}\}_{j=1}^{m}$, then randomly discard part of patches following a uniform distribution, with defined spatial dropout rate β and remaining $(1-\beta)\times m$ facial patches, which can be formulated as:

$$\begin{aligned} P_{i',S} &= \subseteq (P_{i'}, \beta) \\ &= \subseteq (\{\mathbf{p}_{i',1}, \mathbf{p}_{i',2}, ..., \mathbf{p}_{i',m}\}, \beta) \\ &= \{\mathbf{p}_{i',k_2}, \mathbf{p}_{i',k_2+1}, ..., \mathbf{p}_{i',k_2+(1-\beta)*m-1}\}, \\ &\quad i' \in [k_1, k_1+(1-\alpha)\times n-1], \\ &\quad k_2 \sim \mathcal{U}(1, \beta\times m+1), k_2 \in \mathbb{Z}, \beta \in (0,1) \end{aligned} \tag{3}$$

where $\subseteq$ means the subset operation, β is the spatial dropout rate, k_2 is a random start index following a uniform distribution and $P_{i'}$ is the cropped facial patch set of $\mathbf{x}_{i'}$. The uniform distribution can prevent center bias and we finally get a sparse patch set $P_{i',S}$.

Bag of Patches. After we get all facial patch sets $\{P_{i',S}\}_{i'=k_1}^{k_1+(1-\alpha)*n-1}$, we can generate our bag of patches instance $\mathcal{P}$ by first collecting all $P_{i',S}$ together:

$$\mathcal{P} = \cup\{P_{i',S}\}_{i'=k_1}^{k_1+(1-\alpha)*n-1}, \tag{4}$$

where $\cup$ means the union of sets. To guarantee each bag of patches contains all facial regions of original face, we ensure the index of each patch $\mathbf{p}_{i',j} \in P_{i',S}$ without repetition by controlling k_2 in Eq.(3) and reorganize the patch order in $\mathcal{P}$ in original face patch order.

The whole algorithm of our spatiotemporal dropout can be described as follows:

- Step 1: Extract and randomly sample n consistent raw frames from input video $\mathcal{V} = \{\mathbf{f}_i\}_{i=1}^{n}$.
- Step 2: Employ face detection and alignment to each frame $\mathbf{f}_i$ to get facial image sequence $\mathcal{F} = \{\mathbf{x}_i\}_{i=1}^{n}$.
- Step 3: Randomly discard $\alpha\times n$ images from $\mathcal{F}$ through temporal dropout to get $\mathcal{F}_T$.
- Step 4: Grid-wisely crop each remaining facial image $\mathbf{x}_{i'} \in \mathcal{F}_T$ into $m = row \times col$ patches to get facial patch set $P_{i'}$.
- Step 5: Randomly discard $\beta\times m$ patches from each facial patch set $P_{i'}$ to get $P_{i',S}$.
- Step 6: Collect and reorder all facial patch set $P_{i',S}$ to get one bag of patches instance $\mathcal{P}$.

Different from many existing video-level approaches which focus on single frames or the general 3D feature of video, our STD is specially designed to capture the patch-level spatiotemporal cues hidden in deepfake videos. By applying STD operation during training, the spatiotemporal inconsistency in facial regions across different frames is fully explored and learned by our ViT backbone. By introducing the dropout operation, the data to be processed is largely reduced. Besides, since the dropout operation is random, we can generate massive different bag of patches instances $\mathcal{P}$ from the same input video $\mathcal{V}$ in each different training iteration, exhaustively exploiting the dynamical spatiotemporal cues and also serving as data augmentation to further improve our model's robustness and generalization ability. Moreover, the STD operation is flexible and could be plugged into existing ViTs, which is further discussed in the Experiments section.

3.3 Overall Architecture

After obtaining massive bag of patches instances $\mathcal{P}$ through our STD operation above, we feed them into our ViT backbone to optimize the final classifier ϕ to perform detection, which turns Eq.(1) into following formulation:

$$\mathbb{W}^* = \arg\min_{\mathbb{W}} \sum_{\mathcal{V}\in\mathcal{D}} \sum_{\mathcal{P}\in STD(\mathcal{V})} \mathbb{E}(\phi(\mathbb{W};\mathcal{P}), l), \tag{5}$$

where $\mathbb{W}^*$ is the learned optimal output of the detector parameters, $\mathcal{V}$ is the video clip in training set D, $l \in \{0,1\}$ is the video label and $\mathbb{E}$ is an energy function to measure detection loss.

The overall architecture of our proposed approaches is presented in Fig. 2. During the training process, we first employ frame extraction, face detection, and alignment to input video $\mathcal{V}$ to get facial image sequence $\mathcal{F} = \{\mathbf{x}_i\}_{i=1}^{n}$. Then the facial image sequence $\mathcal{F}$ is processed by our STD operation to generate bag of patches instance $\mathcal{P}$, which is then linearly projected with positional embedding and fed into our transformer encoder to achieve a stronger representation capacity. Since the dropout operation is random, massive different bag of patches instances are generated in different training iterations from the same input video, exhaustively mining the dynamic spatiotemporal cues at patch-level and also serving as data augmentation. The representations learned by the transformer encoder are input to a fully connected layer and output the prediction of being fake or real. For our ViT backbone, we choose the most basic ViT-Base-16 model presented in [14] which contains 12 transformer blocks consisting of two normalization layers, one multi-head attention block, and one MLP head. A *Binary Cross Entropy* loss function is employed as our criteria and energy function in Eq.(5). During inference, test videos are processed with the same procedure in training and the model would output the prediction of being fake or real.

4 EXPERIMENTS

We conduct experiments and present a systematic analysis to demonstrate the effectiveness of our proposed Spatiotemporal Dropout Transformer (STDT). First, we make comparisons with 25 state-of-the-arts on three popular benchmarks. Then, we conduct visualizations and experiments under different perturbations and across different datasets to demonstrate its robustness, cross-dataset generalization and representation ability. Finally, a series of ablation analysis are performed to investigate the impact of each key components of our approach.

4.1 Experimental Settings

Datasets. We evaluate our model on FaceForensics++ (FF++) [39], DFDC [12] and Celeb-DF(v2) [29] datasets. FF++ contains 1,000 original videos and corresponding 5×1,000 manipulated videos by using five different generation methods (including Deepfakes, Face2Face, FaceShifter, FaceSwap, and NeuralTextures) with different compression rates (raw, c23, and c40 from no to high compression), where we choose the raw data and select 1,600 for training, 200 for validation and 200 for testing for each subset. DFDC (the Deepfake Detection Challenge dataset) contains original videos recorded from 430 hired actors and over 400G fake videos are synthesized by several deepfake generation methods. Among all videos, we randomly select 6,261 for training, 800 validation, and 781 for testing. Celeb-DF(v2) contains 590 original videos covering different ages, genders, and ethnics, and 5,639 corresponding synthesized videos (4,807 for training, 1,203 for validation, and 518 for testing).

Implementation Details. We use FFmpeg[3] to extract all frames of original videos and choose a pre-trained MobileNet[4] as the face detector for all datasets. The extracted face images are then aligned and resized into 384×384 shape. For our vision transformer backbone, we choose the most basic ViT-Base-16 model described in [14] as our backbone. Specifically, our ViT-Base-16 backbone contains 12 layers with the dropout rate set to 0.1 and 12 self-attention heads with the attention dropout rate set to 0.0. And we choose *Binary Cross Entropy* (*BCE*) as our loss and energy function in Eq.(5). The embedded vector dimension of projected flatten tokens is 768 and the dimension of MLP header is 3072. Our ViT was pretrained on ImageNet and we adopt SGD optimizer and OneCylcleLR strategy. The global learning rate is set to 10^{-3} and the weight decay is set to 10^{-4}. We set the raw frame sequence length n to 24, temporal dropout rate α to 1/4, spatial dropout rate β to 17/18 and each face is grid-wisely cropped into 6×6 patches. Furthermore, no additional augmentation methods are employed during training for a fair evaluation of our STD operation.

4.2 State-of-the-Art Comparison

To demonstrate the advantage of our approach, we make comparisons with 25 state-of-the-art approaches on Celeb-DF(v2), DFDC and FaceForensics++ Deepfakes subset. These approaches include image-level detectors which use the average frame-level scores as the final prediction, video-level detectors such as TD-3DCNN [50] and RNN [40], and recent approaches based on ViT [9, 20, 48]. Training and testing on the same datasets aim to explore the model's capacity of capturing the deepfake cues in deepfake videos. The area under the curve (AUC) score is used as the metric and the results are presented in Tab. 1.

From the table, we can easily observe that our STDT consistently achieves the highest AUC score on all three datasets, ie, 99.8% on FaceForensics++, 99.1% on DFDC and 97.2% on Celeb-DF(v2), which is at least 0.3%, 0.6% and 2.8% higher than the state-of-the-arts, demonstrating the effectiveness of our proposed approach.

Table 1: AUC (%) comparisons with 25 state-of-the-art approaches on three popular benchmarks.

Approach	FF++	DFDC	Celeb-DF	Year
Two-stream [52]	70.7	61.4	53.8	2017
MesoNet [1]	84.7	75.3	54.8	2018
Head-Pose [49]	-	55.9	54.6	2019
Vis-Art [34]	78.0	66.2	55.1	2019
Multi-Task [36]	76.3	53.6	54.3	2019
Warp-Art [28]	93.0	75.5	64.6	2019
XceptionNet [39]	95.5	69.7	65.5	2019
CapsuleNet [37]	96.6	53.3	57.5	2019
CNN-RNN [40]	80.8	68.9	69.8	2019
CNN-Spot [46]	65.7	72.1	75.6	2020
X-Ray [27]	92.8	65.5	79.5	2020
TwoBranch [33]	93.2	-	76.6	2020
PatchBased [5]	57.8	65.6	69.9	2020
AudioVis [35]	-	84.4	-	2020
TD-3DCNN [50]	72.2	79.0	88.8	2021
MAT [51]	97.6	-	-	2021
Lips [18]	97.1	73.5	82.4	2021
DIANet [21]	90.4	90.5	70.4	2021
SPSL [30]	96.9	66.2	-	2021
FD^2Net [53]	99.5	66.1	-	2021
ConvViT [48]	-	91.0	-	2021
EffViT [9]	-	91.9	-	2021
DistViT [20]	-	97.8	-	2021
VFD [7]	-	98.5	-	2022
ICT [13]	98.6	-	94.4	2022
STDT (Ours)	**99.8**	**99.1**	**97.2**	2022

Table 2: Intra-datasets results on Celeb-DF(v2), DFDC and five subsets of FaceForensics++ (Deepfakes, Face2Face, FaceShifter, FaceSwap and NeuralTextures).

Datasets	ACC(%)	AUC(%)	REC(%)	PRE(%)	F1(%)
Celeb-DF	91.70	97.21	95.01	92.55	93.76
DFDC	97.44	99.14	98.63	98.33	98.48
Deepfakes	97.97	99.76	95.70	100.0	97.80
Face2Face	98.01	99.09	97.94	97.94	97.94
FaceShifter	98.64	99.88	98.28	99.13	98.70
FaceSwap	98.61	99.89	99.11	98.23	98.67
NeuralTextures	91.88	97.98	97.98	97.50	97.74

[3]https://ffmpeg.org/
[4]https://github.com/yeephycho/tensorflow-face-detection

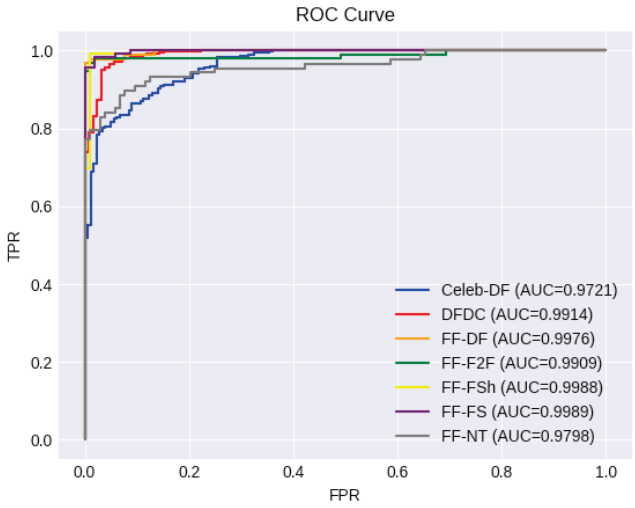

Figure 3: ROC curves on Celeb-DF(v2), DFDC and five subsets of FaceForensics++ datasets (Deepfakes, Face2Face, FaceShifter, FaceSwap and NeuralTextures).

Table 3: Robustness experiments on Celeb-DF(v2) under five representative augmentation methods.

Augmentation	ACC(%)	AUC(%)	REC(%)	PRE(%)	F1(%)
Flip	90.54	96.56	95.01	90.99	92.95
Blur	89.01	94.69	92.94	90.54	91.73
Bright	87.84	94.13	92.06	89.68	90.86
Compress	89.38	95.74	96.76	88.20	92.29
Gaussian Noise	87.45	94.25	96.76	85.90	91.01
Origin	91.70	97.21	95.01	92.55	93.76

Specifically, DFDC and Celeb-DF contain higher quality videos with a variety of scenarios, people groups and generation methods and the FF++ contains videos from the Internet, indicating our approach can still achieve impressive performance in a variety of situations. The main reason is that our approach focuses on capturing the patch-level inconsistency from both inter-frame and intra-frames rather than inter-frame inconsistency like [50] .

Additionally, as the FF++ dataset contains several subsets generated by different deepfake methods, we conduct intra-dataset experiments on each subset respectively to further evaluate our proposed approach. We choose the accuracy (ACC), the area under the curve (AUC), recall (REC), precision (PRE), and F1 score as our evaluation metrics. The results are presented in Tab. 2 together with Celeb-DF(v2) and DFDC. From the table, it can be observed that our proposed approach achieves impressive performance on all three datasets with all ACC higher than 90%, AUC higher than 97%, demonstrating the effectiveness and the ability of our approach on handling various deepfake generation methods. Especially on the Deepfakes, Face2Face, FaceShifter, and FaceSwap subset of FF++, we achieve all AUC higher than 99%. The corresponding ROC curves on each dataset are also presented in Fig. 3 for better comprehension. After analyzing some samples from these datasets, we find that all these different generation methods will result in subtle visual artifacts around some facial regions across frames, which may be disguised by the video quality or so subtle that could be easily ignored. These spatiotemporal cues are hard to be captured by human eyes, but the results demonstrate that our approach can effectively capture these inconsistencies and distinguish the deepfake videos generated by various methods with impressive performance.

Table 4: Cross-datasets generalization results on Celeb-DF(v2) (C-DF), DFDC and FaceForensics++ Deepfakes (FF-DF).

Train	Test	ACC(%)	AUC(%)	REC(%)	PRE(%)	F1(%)
C-DF	DFDC	83.87	70.14	100.0	83.87	91.23
	FF-DF	74.62	79.26	87.10	62.79	72.97
DFDC	C-DF	71.43	73.60	95.59	70.96	81.45
	FF-DF	80.71	87.95	92.47	69.36	79.26
FF-DF	C-DF	68.92	69.78	96.47	68.76	80.29
	DFDC	83.87	66.99	100.0	83.87	91.23

4.3 Performance Analysis

Robustness. To evaluate the robustness of our proposed approach, we train our model on the original datasets and test them under several augmentation and perturbation methods. We choose five representative methods, including Flip, Blur, Bright, Compress, and Gaussian Noise. Specifically, we utilize 10x10 kernel size in Blur, set compression ratio to 4 in Compress (where file storage becomes 1/4 after compression), and set the mean and variance of Gaussian Noise to 0.1 and 0.01. Besides, we train the models without any additional augmentation methods. The results are presented in Tab. 3. From the table, we could observe that the performance under different augmentation methods is nearly the same as origin, such as the ACC and AUC under Flip and Compress. This provides clear evidence for the impressive robustness of our proposed approach on handling real-world scenarios such as compressed videos on the Internet. Two video examples and corresponding output predictions during testing are presented in Fig. 4 for better comprehension.

Cross-dataset Generalizability. To further demonstrate our approach's generalizability, we perform cross-dataset experiments on three datasets by training on one dataset then testing on the other two. The results are presented in Tab. 4. From the results, our approach exhibits competitive AUC scores compared to those opponents in Tab 1. For example, the model training on DFDC and testing on Celeb-DF and FF++, and the model training on Celeb-DF and testing on DFDC both outperform plenty of approaches listed in Tab. 1. This implies that our approach achieves an impressive generalization ability by capturing the spatial and temporal inconsistency generally existing in deepfake videos instead of focusing on the specific generation pattern or artifact of certain deepfake methods. Moreover, in some real-world scenarios where we may not access to the specific generated data, we can directly utilize the model trained on other accessible data to perform detection based on the proven cross-dataset generalizability. Besides, if we can collect a small amount of data, fine-tuning the trained model could achieve better performance.

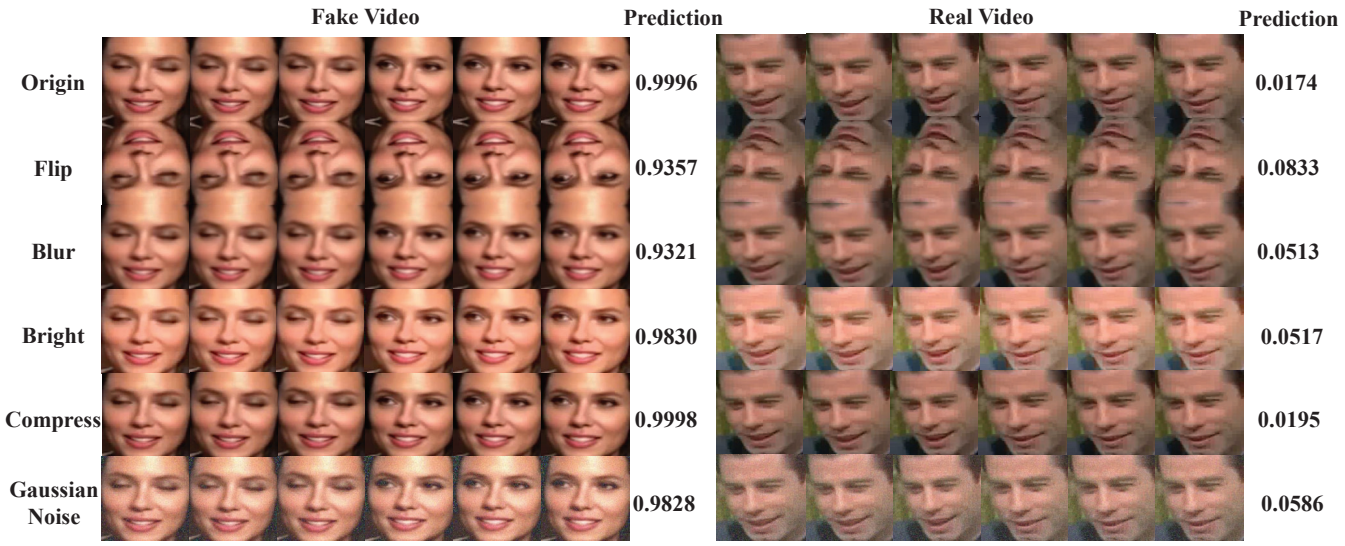

Figure 4: Two video examples from Celeb-DF(v2) with five augmentation methods and the predicted possibilities of being fake.

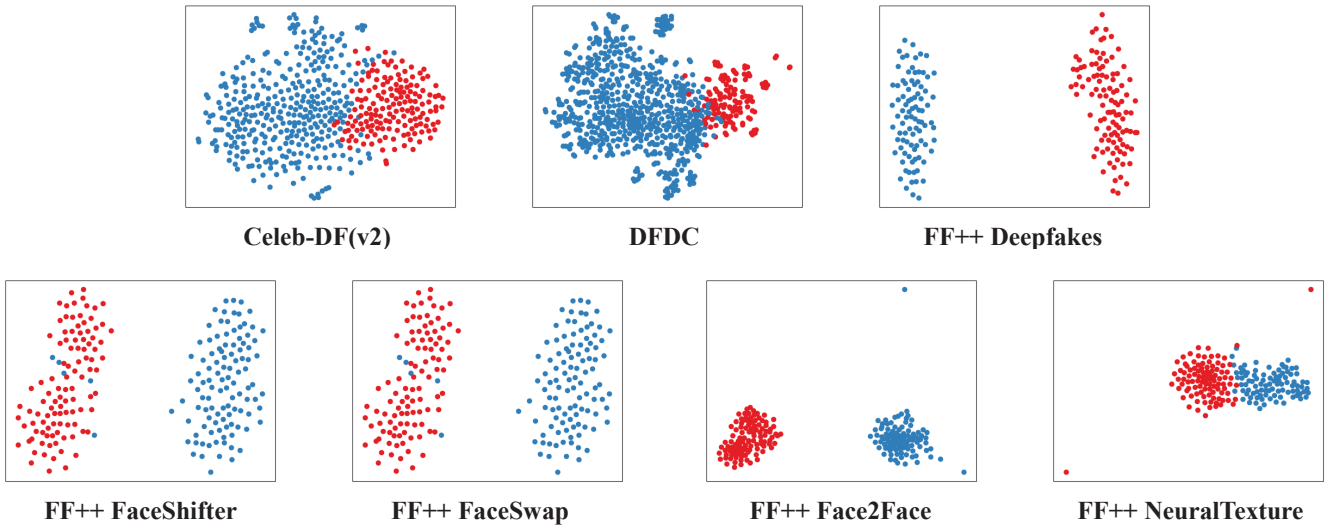

Figure 5: Visualization of representations on Celeb-DF(v2), DFDC and FF++ (five subsets) via t-SNE. Red: Real, Blue: Fake.

Representation. To further demonstrate the representation ability of our approach, we utilize t-SNE [24] to visualize the representations learned from trained transformer encoder on three datasets and each subsets. The visualization results are presented in Fig. 5, which shows that the real and fake videos generated by different methods are distinctly clustered in latent space, proving the strong representation ability of our approach.

4.4 Ablation Analysis

To systemically evaluate the key components of our approach, we conduct ablation experiments on Celeb-DF(v2) dataset from three aspects and present a complete analysis in the following.

Dropout Operation. To check the effect of dropout operation, we first investigate four variants: 1) no dropout (-), 2) spatial dropout (S), 3) temporal dropout (T), and 4) spatiotemporal dropout (S+T). Here, spatial dropout discards random patches in frames but does not discard frames, while temporal dropout just randomly drops some frames from extracted sequence. The results are presented in Tab. 5, and we can observe that the accuracy and AUC can be improved when applying spatial or temporal dropout. Specially, combining both spatial and temporal dropout achieves the best performance due to the consideration of spatiotemporal inconsistency, demonstrating the effectiveness of spatiotemporal dropout. To further verify that, we conduct an experimental comparison

Table 5: Ablation analysis on operation effectiveness on Celeb-DF(v2): no-dropout (-), temporal dropout (T), spatial dropout (S), and spatiotemporal dropout (S+T).

Dropout	ACC(%)	AUC(%)	REC(%)	PRE(%)	F1(%)
-	87.65	93.43	94.71	85.70	90.96
S	88.42	95.01	97.94	86.27	91.74
T	89.58	95.56	92.65	91.57	92.11
S+T	91.70	97.21	95.01	92.55	93.76

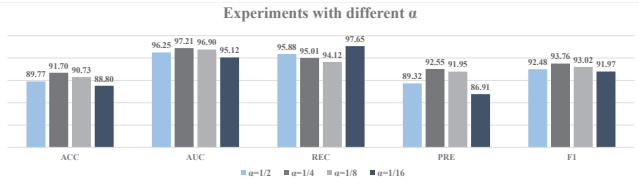

Figure 6: Ablation analysis on temporal dropout rate α on Celeb-DF(v2). We set α to 1/2, 1/4, 1/8, and 1/16 with other hyper-parameters fixed.

with TD-3DCNN [50] by using different dropout operations and network architectures, and report the results in Tab 6 where remarkable improvement is achieved by our spatiotemporal dropout.

Table 6: The AUC (%) comparisons with [50] under different dropout operations and network architectures on Celeb-DF(v2), DFDC and FaceForensics++(FF++).

Architecture	Dropout	Celeb-DF	DFDC	FF++
3DCNN	TD [50]	88.83	78.97	72.22
	STD	93.39	85.87	92.70
ViT	TD [50]	95.56	97.39	97.41
	STD	97.21	99.14	99.76

Dropout Rate. To further explore the effects of spatial and temporal dropout respectively, we evaluate our approach on Celeb-DF(v2) by setting different dropout rate α and β with other hyper-parameters fixed. We define $\beta^{'} = 1 - \beta$ for convenience and the results are presented in Fig. 6 and Fig. 7. From the results, we can find that different α and β result in different detection performance on accuracy and AUC, and the influence is significant, *i.e.*, the AUC is about 2% higher when α set to 1/4 compared to 1/16, and about 1.5% higher when $\beta^{'}$ set to 1/18 compared to 1/9. Besides, we can find that the accuracy and AUC decline when setting α or β too low or high, *i.e.*, the ACC and AUC both declined when changing α from 1/4 to 1/2 and 1/8, or changing $\beta^{'}$ from 1/18 to 1/9 and 1/36. We analyze this is because the spatiotemporal inconsistency is crushed or damaged when setting α or β too low or high, dropping too many or few frames and patches. This indicates that the value of α and β should be carefully considered.

ViT Backbones. We evaluate our model's performance under four ViT backbones (ViT-B16, ViT-B32, ViT-L16 and ViT-L32 [14]) on

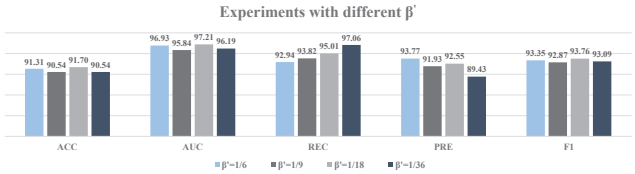

Figure 7: Ablation analysis on Spatial Dropout rate $\beta^{'} = 1 - \beta$ on Celeb-DF(v2). We set $\beta^{'}$ to 1/6, 1/9, 1/18, and 1/36 with other hyper-parameters fixed.

Table 7: Ablation analysis of four different ViT backbones on Celeb-DF(v2) with (√) and without (-) STD.

Backbone	STD	ACC(%)	AUC(%)	REC(%)	PRE(%)	F1(%)
ViT-B16	-	82.43	92.63	79.41	92.79	85.58
	√	91.70	97.21	95.01	92.55	93.76
ViT-B32	-	75.29	87.68	99.71	72.75	84.12
	√	85.52	95.34	98.53	82.72	89.93
ViT-L16	-	76.83	89.27	97.06	75.01	84.62
	√	86.87	94.26	91.47	88.57	90.14
ViT-L32	-	71.43	83.13	98.82	70.01	81.95
	√	86.49	96.95	97.35	84.44	90.44

Celeb-DF(v2) to further demonstrate the effectiveness and flexibility of our spatiotemporal dropout. The results are presented in Tab. 7. From the table, we can find that the performance is consistently improved when incorporating STD, averagely improving ACC by 11.15% and AUC by 7.76%. This conclusively demonstrates the effectiveness of our STD as well as its flexibility to plug into various ViT architectures. Besides, the results shown in Tab. 6 also come to this conclusion when employing in CNN architectures.

5 CONCLUSION

In this paper, we propose a spatiotemporal dropout transformer to detect deepfake videos at patch-level. In the approach, an input video is grid-wisely cropped and reorganized as massive *bag of patches* instances which are then fed into a vision transformer to achieve robust representations. A spatiotemporal dropout operation is designed to fully explore the patch-level spatiotemporal inconsistency as well as serving as data augmentation, further improving model's robustness and generalizability. The spatiotemporal dropout operation is flexible and can be plugged into various ViTs. Extensive experiments clearly shows our approach outperforms 25 state-of-the-arts with impressive robustness, generalizability and representation ability. In the future, we will extend our approach to more video understand tasks and also enhance its interpretability to provide a more human-understandable detection result.

ACKNOWLEDGMENTS

This work was partially supported by grants from the National Key Research and Development Plan (Grant No. 2020AAA0140001), Beijing Natural Science Foundation (L192040) and National Natural Science Foundation of China (61772513).

REFERENCES

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

[2] Irene Amerini and Roberto Caldelli. 2020. Exploiting Prediction Error Inconsistencies through LSTM-based Classifiers to Detect Deepfake Videos. In *IH&MMSec*. 97–102.

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *ICCV*. 6836–6846.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *ECCV*. 213–229.

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

[6] Wen Chen, Yun Q. Shi, and Wei Su. 2007. Image splicing detection using 2-D phase congruency and statistical moments of characteristic function. In *Security, Steganography, and Watermarking of Multimedia Contents IX*, Vol. 6505. 65050R.

[7] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Tao Ye, and Liqiang Nie. 2022. Voice-Face Homogeneity Tells Deepfake. *arXiv preprint arXiv:2203.02195* (2022).

[8] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *CVPR*. 8789–8797.

[9] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. 2022. Combining efficientnet and vision transformers for video deepfake detection. In *ICIAP*. 219–229.

[10] Kevin Dale, Kalyan Sunkavalli, Micah K. Johnson, et al. 2011. Video face replacement. *ACM TOG* 30, 6 (2011), 130.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. 4171–4186.

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

[13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. 2022. Protecting Celebrities from DeepFake with Identity Consistency Transformer. In *CVPR*. 9468–9478.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[15] Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormählen, Patrick Pérez, and Christian Theobalt. 2014. Automatic Face Reenactment. In *CVPR*. 4217–4224.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NeurIPS*. 2672–2680.

[17] David Guera and Edward J. Delp. 2018. Deepfake Video Detection Using Recurrent Neural Networks. In *AVSS*. 1–6.

[18] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2021. Lips Don't Lie: A Generalisable and Robust Approach To Face Forgery Detection. In *CVPR*. 5039–5049.

[19] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. 2020. A survey on visual transformer. *arXiv preprint* (2020).

[20] Young Jin Heo, Young Ju Choi, Young-Woon Lee, and Byung-Gyu Kim. 2021. Deepfake Detection Scheme Based on Vision Transformer and Distillation. *arXiv preprint arXiv:2104.01353* (2021).

[21] Ziheng Hu, Hongtao Xie, Yuxin Wang, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. 2021. Dynamic Inconsistency-aware DeepFake Video Detection. In *IJCAI*. 736–742.

[22] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*. 4401–4410.

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *CVPR*. 8107–8116.

[24] Van Der Maaten Laurens and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *JMLR* 9, 2605 (2008), 2579–2605.

[25] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2019. FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping. *arXiv preprint arXiv:1912.13457* (2019).

[26] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2020. Advancing High Fidelity Identity Swapping for Forgery Detection. In *CVPR*. 5074–5083.

[27] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face x-ray for more general face forgery detection. In *CVPR*. 5001–5010.

[28] Yuezun Li and Siwei Lyu. 2019. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In *CVPRW*. 46–52.

[29] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *CVPR*. 3204–3213.

[30] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. 2021. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *CVPR*. 772–781.

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*. 10012–10022.

[32] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. 2018. Attribute-Guided Face Generation Using Conditional CycleGAN. In *ECCV*. 293–308.

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

[34] F. Matern, C. Riess, and M. Stamminger. 2019. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In *WACVW*. 83–92.

[35] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues. In *ACM MM*. 2823–2832.

[36] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. 2019. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *BTAS*. 1–8.

[37] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Use of a Capsule Network to Detect Fake Images and Videos. *arXiv preprint arXiv:1910.12467* (2019).

[38] Xunyu Pan, Xing Zhang, and Siwei Lyu. 2012. Exposing image splicing with inconsistent local noise variances. In *ICCP*. 1–10.

[39] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *ICCV*. 1–11.

[40] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. 2019. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. In *CVPRW*. 80–87.

[41] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: learning lip sync from audio. *ACM TOG* 36, 4 (2017), 95:1–13.

[42] Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: image synthesis using neural textures. *ACM TOG* 38, 4 (2019), 66:1–12.

[43] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*. 2387–2395.

[44] Rubén Tolosana, Sergio Romero-Tapiador, Julian Fiérrez, and Rubén Vera-Rodríguez. 2020. DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance. In *ICPR Workshops*. 442–456.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. 2017. Attention is all you need. In *NeurIPS*. 5998–6008.

[46] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. CNN-generated images are surprisingly easy to spot... for now. In *CVPR*. 8695–8704.

[47] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*. 568–578.

[48] Deressa Wodajo and Solomon Atnafu. 2021. Deepfake Video Detection Using Convolutional Vision Transformer. *arXiv preprint arXiv:2102.11126* (2021).

[49] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing Deep Fakes Using Inconsistent Head Poses. In *ICASSP*. 8261–8265.

[50] Daichi Zhang, Chenyu Li, Fanzhao Lin, Dan Zeng, and Shiming Ge. 2021. Detecting Deepfake Videos with Temporal Dropout 3DCNN. In *IJCAI*. 1288–1294.

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

[52] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. 2017. Two-Stream Neural Networks for Tampered Face Detection. In *CVPRW*. 1831–1839.

[53] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z Li. 2021. Face Forgery Detection by 3D Decomposition. In *CVPR*. 2929–2939.