ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746888

ADT: ANTI-DEEPFAKE TRANSFORMER

Ping Wang[†], *Kunlin Liu*[†], *Wenbo Zhou*[†,⋆], *Hang Zhou*[‡], *Honggu Liu*[†], *Weiming Zhang*[†], *Nenghai Yu*[†]
[†] University of Science and Technology of China
[‡] Simon Fraser University

ABSTRACT

Recently almost all the mainstream deepfake detection methods use Convolutional Neural Networks (CNN) as their backbone. However, due to the overreliance on local texture information which is usually determined by forgery methods of training data, these CNN-based methods cannot generalize well to unseen data. To get out of the predicament of prior methods, in this paper, we propose a novel transformer-based framework to model both global and local information and analyze anomalies of face images. In particular, we design attention leading module, multi-forensics module and variant residual connections for deepfake detection, and leverage token-level contrast loss for more detailed supervision. Experiments on almost all popular public deepfake datasets demonstrate that our method achieves state-of-the-art performance in cross-dataset evaluation and comparable performance in intra-dataset evaluation.

Index Terms— Deepfake Detection, Transferability, Face Forensics, Vision Transformer.

1. INTRODUCTION

Deepfakes are synthetic media in which a person in an existing image or video is replaced with someone else's likeness. With the rapid development of Variational Auto-Encoders (VAE) [1], and Generative Adversarial Networks (GAN) [2], deepfake generation techniques have been updating iterations with an incredible speed. Unfortunately, they can be easily used for malicious purposes. By now, it has become almost impossible for humans to distinguish whether some media are credible or not. Deepfake significantly threatens the reputation of celebrities, even may cause political crises.

For security concerns, a series of deepfake detection methods have been proposed in recent years. Among them, most [3, 4, 5, 6, 7] are designed based on Convolutional Neural Networks (CNN), showing significant power. They can achieve perfect performances on FaceForensics++ [3] dataset when train and test on it. Nevertheless, their accuracy drops heavily when test on other datasets, such as Celeb-DF [4]. That's because CNN-based methods distinguish fake media by learning local texture information, which is divergent among datasets. In other words, pure texture information is not commonly applicable evidence for deepfake detection.

Although deepfake media are complex and diverse, they have a common problem: there are always some defects that are normal locally but abnormal from a global perspective. For instance, mismatched facial expressions and head postures, inconsistent color and textures, unnatural blur of eyes and teeth, etc cannot be recognized if only given a local part but can be recognized with the help of global information. So the local areas of interest should be determined according to the global semantics, and modeling long-distance dependencies in the spatial domain are necessary. But it is not direct for the convolutional attention mechanism, especially when the kernel is small. Global pooling may be a choice for assembling global information, but it will average the fragile forgery tracks, resulting in a loss of distinguishability. Therefore, new detection models based on other frameworks are sorely needed.

Recently, vision transformer (ViT) [8] achieved massive success in classical classification tass. It applies transformer to a sequence of image patches with an innate attention mechanism which effectively broaden the receptive field, thus facilitate the capturing of global information. Extended works such as object detection and fine-grained classification [9, 10] further confirm its ability, giving us lots of inspiration. We realize that transformer might be a good way to solve the hard problem of modeling long-distance information. Actually, there are already some works based on transformer, such as [11, 12, 13]. But their naive strategies to using transformer finally limits their detection performance, especially transferability.

In this paper, we propose a novel deepfake detection framework, Anti-Deepfake Transformer (ADT). ADT consists of four cascaded trans-blocks that include three stacked transformer layers to model both global and local information. And we design Variant Residual Connections (VRC) between adjacent trans-blocks to ensure capturing enough texture information. Besides, we design an Attention Leading Module (ALM) to help the network focus on the most valuable and distinguishable regions (That's the area most likely to be modified) while ignoring redundant information, such as the background. Moreover, we design a Multi-Forensics Module (MFM) to combine the features from different levels. In the training process, we leverage a contrast loss to further improve the performance in token-level.

The key contributions of this paper are threefold as below:

1. We propose a novel deepfake detection framework, Anti-Deepfake Transformer (ADT), which pays attention to both global and local information and makes up for the shortcomings of CNN-based methods.
2. We design Attention Leading Module (ALM), Variant Residual Connection (VRC) and Multi-Forensics Module (MFM) to take full advantage of ADT and introduce contrast loss to further improve its performance.
3. Extensive experiments demonstrate that ADT could maintain considerable performance in the intra-dataset evaluation and achieve state-of-the-art in the cross-dataset evaluation in deepfake detection.

2. METHOD

2.1. Anti-Deepfake Transformer Pipeline

We show our framework in Figure 1. First we split images into small patches and project them into the embedding space. Then we add learnable position embeddings and input them into trans-blocks connected by variant residuals. And then, we apply ALM to select the

⋆ Corresponding author: welbeckz@ustc.edu.cn

most valuable tokens from all the tokens output by the final trans-blocks. After that these selected tokens are input into a single transformer layer to get sub-classification. Finally, we merge the four sub-results as the final prediction.

Sliding Patch Sequences. Denote input images to the network as $\mathbf{I}$. We first split $\mathbf{I}$ into a patch sequence $\mathbf{I}_p$ through preprocessing. In consideration of keeping the original neighbor structure, we choose to use overlapping patches sequences as input instead of non-overlapping patches. To be specific, we denote the input image with resolution $H \times W$, the size of image patch as P, and the stride of sliding windows as S. Thus images will be split into N patches:

$$N = [(H - P)//S + 1] \times [(W - P)//S + 1] \tag{1}$$

Patch Embedding. We map the vectorized patches $\mathbf{I}_p$ into a latent D-dimensional embedding space using a trainable linear projection, and then a learnable position embedding is added to retain positional information as follows:

$$\mathbf{Z}_0 = [\mathbf{I}_{class}; \mathbf{I}_p^1\mathbf{E}, \mathbf{I}_p^2\mathbf{E}, ..., \mathbf{I}_p^N\mathbf{E}] + \mathbf{E}_{pos} \tag{2}$$

where N is the number of image patches, $\mathbf{E} \in \mathbb{R}^{P^2 \times (C \cdot D)}$ is the patch embedding projection, and $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$ denotes the position embedding.

Trans-block. Our trans-block contains three transformer layers which consist of multi-head self-attention and multi-layer perception blocks. Formally, the output of each layer can be written as follows:

$$\begin{aligned} \mathbf{Z}_l^{'} &= \text{MSA}(\text{LN}(\mathbf{Z}_{l-1})) + \mathbf{Z}_{l-1} \\ \mathbf{Z}_l &= \text{MLP}(\text{LN}(\mathbf{Z}_l^{'})) + \mathbf{Z}_l^{'} \end{aligned} \tag{3}$$

where $\mathbf{LN}(\cdot)$ denotes the layer normalization and $\mathbf{Z}_l$ is the encoded image representation of l-th layer, and $l \in 1, 2, ..., L$. Just like ViT, we use the first token of transformer layer $\mathbf{Z}_L^0$ as the representation of features and forward it to a classifier head for classification. Suppose that all the layers have C self-attention heads then the hidden layer features and attention weights can be expressed as follows:

$$\begin{aligned} \mathbf{Z}_l &= [\mathbf{Z}_l^0; \mathbf{Z}_l^1, \mathbf{Z}_l^2, \ldots, \mathbf{Z}_l^N] \\ \mathbf{A}_l &= [[a_l^{0_0}; a_l^{0_1}, .., a_l^{0_N}], .., [a_l^{C_0}; a_l^{C_1}, .., a_l^{C_N}]] \end{aligned} \tag{4}$$

Attention Leading Module As we mentioned earlier, analyzing the most discriminative information is a crucial step. But the previous work [14] pointed out that the raw attention weights do not necessarily correspond to the relative importance of input tokens especially for higher layers of a transformer-based model. Therefore, we cannot evaluate the importance of features directly from final attention weights. To ensure the correspondence between the input token and the attention weight as much as possible by fusing the attention weights of all the previous transformer layers. Specifically, we recursively apply a matrix multiplication to the raw attention weights after softmax in all the layers as:

$$\mathbf{A}_{final} = \prod_{l=0}^{L-1} \text{Softmax}(\mathbf{A}_l) \tag{5}$$

Then we find the index of the largest attention weight from $\mathbf{A}_{final}$ and denoted as $M_1, M_2, ..., M_C$, where C is the number of self-attention heads. And softmax is introduced when calculating $\mathbf{A}_{final}$. Finally, we take the token we selected and the classification token concatenate together as the input sequence of the higher layer, expressed as the following form:

$$\mathbf{Z}_{final} = [\mathbf{Z}_{L-1}^0; \mathbf{Z}_{L-1}^{M_1}, \mathbf{Z}_{L-1}^{M_2}, \ldots, \mathbf{Z}_{L-1}^{M_C}] \tag{6}$$

ALM ensures that the corresponding relationship between the attention weight and the input tokens is forwarded to the higher layers of the model. And it also helps the model to focus on the most valuable area for deepfake detection.

Variant Residual Connection. Texture information is always an important clue for deepfake detection. But our method is a pure transformer-based method. It's not easy for such architecture to capture enough texture information. To address this issue, we adopt variant residual connections among adjacent trans-blocks. It is noteworthy that residual connection proposed by Resnet [15] leverage the addition while we leverage subtraction for learned features. Denote the four trans-blocks as $\mathbf{T}_i$:

$$\mathbf{X}_{T_{i+1}} = \text{F}(\mathbf{X}_{T_i}) - \mathbf{X}_{T_i} \tag{7}$$

where i represents index of trans-block, the $\mathbf{X}_{T_i}$ means input of i-th trans-block, and F is formal description of trans-block.

Multi-Forensics Module. Considering the complexity and diversity of deepfake media, we believe that the detection model should not only focus on those high-layer features but also low-layer features, and allow all the features from different levels participate in the final decision. And we believe that for an exemplar deepfake detection method, the corresponding features of the real face should show the nature of tending to "True" at every level, and vice versa. So we propose Multi-Forensics Module (MFM) to obtain more convincing and exhaustive prediction. Denote $\mathbf{T}_i$ as i-th trans-block, then $\mathbf{Z}_{T_i}$ represents the tokens output by the i-th block:

$$\mathbf{Z}_{T_i} = [\mathbf{Z}_{T_i}^0; \mathbf{Z}_{T_i}^1, \mathbf{Z}_{T_i}^2, \ldots, \mathbf{Z}_{T_i}^{M_C}] \tag{8}$$

It's noteworthy that $\mathbf{Z}_{T_i}$ is already processed by ALM here. Then we input $\mathbf{Z}_{T_i}$ into an additional transformer layer, and take linear classification on the output classification tokens. As shown in Figure 1, we can get sub-prediction of classification tokens from four levels as $\mathbf{S_i}$, then:

$$\mathbf{Pred} = \text{Mean}(\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, \mathbf{S}_4). \tag{9}$$

Finally, the prediction $\mathbf{Pred}$ is the mean value of them all.

2.2. Training Losses

Training such a deep transformer network require strong and detail supervision. We leverage the combination of classification loss(cross-entropy loss) $\mathcal{L}_{cls}$ and token-level contrast loss $\mathcal{L}_{con}$ as training losses.

The features only supervised by softmax loss or cross-entropy loss are not discriminative enough, since the differences between real and fake faces might be imperceptible. So we introduce a token-level contrast loss, $\mathcal{L}_{con}$ as follows, which aims to minimize the similarity of classification tokens corresponding to different labels and maximize the similarity of classification tokens of samples with the same label, to further enhance the supervision.

$$\begin{aligned} \mathcal{L}_{con} = \frac{1}{N^2} \sum_{i}^{N} \Big(& \sum_{j:y_i=y_j}^{N} (1 - \frac{\mathbf{Z}_i \cdot \mathbf{Z}_j}{\|\mathbf{Z}_i\| \, \|\mathbf{Z}_j\|}) \\ & + \sum_{j:y_i \neq y_j}^{N} \max(\frac{\mathbf{Z}_i \cdot \mathbf{Z}_j}{\|\mathbf{Z}_i\| \, \|\mathbf{Z}_j\|} - \alpha, 0) \Big) \end{aligned} \tag{10}$$

where $\mathbf{Z}_i$ and $\mathbf{Z}_j$ are classification tokens of the last transformer layer which are pre-processed with ℓ_2 normalization, $\frac{\mathbf{Z}_i \cdot \mathbf{Z}_j}{\|\mathbf{Z}_i\| \|\mathbf{Z}_j\|}$ is the cosine similarity of $\mathbf{Z}_i$ and $\mathbf{Z}_j$ and α is a constant margin to balance the contribution of the second item.

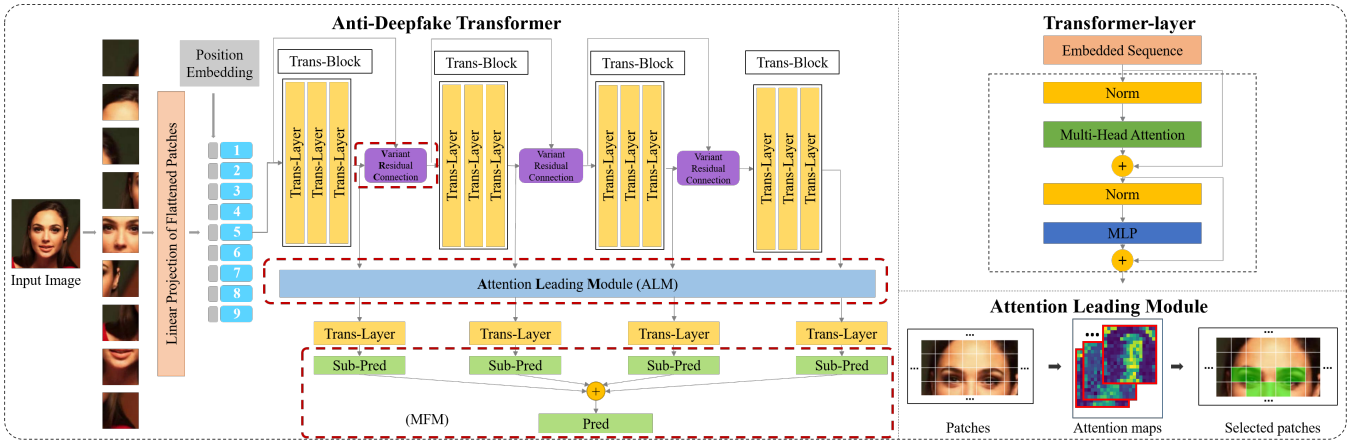

Fig. 1. Left is framework of ADT, right is structure of transformer layer and brief description of the Attention Leading Module.

Table 1. Intra-Dataset evaluation results (ACC(%) and AUC(%)) on FaceForensics++ dataset with high-quality and low-quality settings.

Methods	HQ		LQ	
	ACC	AUC	ACC	AUC
MesoNet [6]	83.10	-	70.47	-
Face X-Ray [23]	-	87.35	-	61.60
Xception [7]	92.39	94.86	80.32	81.76
Two-Branch [24]	-	98.70	-	86.59
SPSL [25]	91.50	95.30	81.57	82.82
F^3-Net [26]	97.52	98.10	90.43	93.30
Multi-attentional [27]	97.60	99.29	88.69	90.40
M2TR [28]	98.23	99.84	92.35	94.22
Long-distance [12]	95.81	98.49	99.51	99.88
BOLF[29]	-	-	-	-
Ours	92.05	96.30	81.48	82.52

3. EXPERIMENT

3.1. Datasets

Same as related works of deepfake detection, we first conduct our experiments on the most popular two benchmark deepfake datasets:FaceForensics++ (FF++) [3] and Celeb-DF [4]. FF++ consists of five kinds of common deepfake generation methods [16, 17, 18, 19, 20]. Celeb-DF is the most challenging dataset to almost all the current methods. To further evaluate transferability, we use test set of DeepFake Detection Challenge (DFDC) [21], FaceShifter [20] and DeeperForensics [22] as evaluation dataset.

3.2. Implementation and Hyper-Parameters

In our experiments, we resize target images to 256×256 and then augment the data (random cropping and random horizontal flipping for training and center cropping for testing).Then We split target image to patches of size 16×16 and the stride of sliding window is set to 12. So the H, W, P, S in Equation 1 is 256, 256, 16, 12 respectively. And the constant margin α in Equation 10 is set to 0.4, which is selected through experimental verification. Since we utilized transformer as our base layer, we load intermediate weights from ViT-B_16 model pretrained on ImageNet21k and the batch size is set to 16. SGD optimizer is employed with a momentum of 0.9. The learning rate is initialized as 0.03. We adopt cosine annealing as the scheduler of optimizer. All the experiments are performed with two Nvidia GeForce RTX 2080Ti GPUs using the PyTorch toolbox and APEX with FP16 training.

3.3. Comparison with Previous Methods

We compare our framework with state-of-the-art methods in deepfake detection. First, we train and test the performance of our model on FF++, and further we test the cross-dataset performance of our model on Celeb-DF and other popular datasets to evaluate its transferability. Like previous methods, we use ACC (Accuracy) and AUC (Area Under Receiver Operating Characteristic Curve) as main evaluation metrics.

Intra-Dataset Evaluation We conduct in-dataset evaluation on processed images in FF++, and we directly use the results reported in their papers for fair comparison. As shown in Table 1, our method can achieve competitive performance compared with previous methods. Since CNN has strong ability to capture sufficient texture information, most CNN-based methods achieve perfect performance. Though we proposed several methods to enhance our methods, due to its structure, our framework cannot learn such texture-level features as CNN, so it's difficult to accurately grasp the inherent texture introduced by specific generation method, which limits the performance of ADT on specific dataset.

Cross-Dataset Evaluation We train our model on FF++(all generation methods), then test it on Celeb-DF and DF(Deepfakes in FF++). The image-level experimental results are shown in Table 2, where we also list the performance of previous competitive detection methods. It can be seen that the performance of our model is very impressive, outperforms in the comparation of transferability with all existing popular works.

Furthermore, we evaluate video-level performance on Celeb-DF, DFDC, Faceshifter and DeeperForensics after training our model on FF++ (all generation methods). We align the experiment and settings with many competitive works in deepfake detection at the video level, and test our models on videos in corresponding dataset. The results are shown in Table 3. ADT almost outperforms on all datasets except DeeperForensics. Thanks to capturing more common artifacts, ADT is not as seriously overfit to training set as pre-

Table 2. Cross-Dataset Evaluation (AUC (%)) on images in Celeb-DF. Results for some other methods are from [4].

Method	FF++ (DF)	Celeb-DF
MesoNet [6]	84.70	54.80
Face X-Ray [23]	-	-
Xception-c23 [30]	99.7	65.3
Two-Branch [24]	93.20	73.40
SPSL [25]	96.94	76.88
F^3-Net [26]	97.97	65.17
Multi-Attention [27]	99.80	67.44
M2TR [28]	99.50	65.70
Long-distance [12]	**99.97**	70.33
BOLF[29]	-	78.26
Ours	98.71	**84.97**

Table 3. Cross-Dataset evaluation results (AUC(%)) on videos in Celeb-DF, DFDC, FaceShifter and Deeper(DeeperForensics). Results for other methods are from [31].

Method	Celeb-DF	DFDC	FaceShifter	Deeper
Xception [3]	73.7	70.9	72.0	84.5
CNN-aug [5]	75.6	72.1	65.7	74.4
Patch-based [32]	69.6	65.6	57.8	81.8
Face X-Ray [23]	79.5	65.5	92.8	86.8
Multi-task [33]	75.7	68.1	66.0	77.7
DSP-FWA [34]	69.5	67.3	65.5	50.2
Two-Branch [24]	76.7	-	-	-
LipForensics [31]	82.4	73.5	97.1	**97.6**
Ours	**89.0**	**76.2**	**98.0**	96.7

vious methods, shows very impressive transferability, and achieves excellent results on these complex datasets that never involved in training stage.

3.4. Ablation Study

To illustrate the effectiveness of proposed modules, we conduct several ablation studies. We take pure stacked transformers which consists of four cascaded trans-blocks as our baseline and then add proposed modules step by step. We train these models on FF++ and test on images in Celeb-DF. Results in Table 4 shows that the proposed contrast loss greatly improve the AUC score. Besides, ALM and VRC also help the model gain more transferability. Although MFM seems not to improve the AUC score effectively, it provides a fascinating perspective for deepfake detection. In general, all the data confirms that the introduced modules and methods are indeed practical for our framework.

3.5. Qualitative Analysis

To explore the difference between our method and traditional CNN-based methods, we respectively visualize the attention map of ADT and Gradient-weighted Class Activation Mapping (Grad-CAM) of the Xception in Figure 2. It can be seen that compared to Xception, ADT can more accurately focus on the abnormal regions in deepfake images, which are the forged areas poorly coordinated with the entire face image. This can provide deepfake detection and face forensics more meaningful information instead of only classification results.

Table 4. Ablation study on proposed modules. Cross-Dataset evaluation results (ACC (%) and AUC (%)) on Celeb-DF.

Baseline	Contrast Loss	ALM	VRC	MFM	ACC	AUC
✓					79.15	74.67
✓	✓				80.48	81.23
✓	✓	✓			81.56	83.17
✓	✓	✓	✓		81.76	85.05
✓	✓	✓	✓	✓	82.44	84.97

Fig. 2. The first row are deepfake images.Their attention maps of ADT and Grad-CAM results of Xception are shown in the second and third row.

4. CONCLUSION

In this paper, we propose a pure transformer-based framework for deepfake detection. It aims to expose inconsistency between local and global information. Extensive experiments demonstrate that ADT achieves the state-of-the-art transferability among almost all the public datasets, confirming that ADT can capture more common artifacts than existed methods. And we hope ADT can inspire others to explore the potential of transformer in deepfake detection field.

5. ACKNOWLEDGMENT

This work was supported in part by the Natural Science Foundation of China under Grant U20B2047, 62072421, 62002334, 62102386 and 62121002, Exploration Fund Project of University of Science and Technology of China under Grant YD3480002001, and by Fundamental Research Funds for the Central Universities under Grant WK2100000011. This work was also partly supported by Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China.

6. REFERENCES

[1] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin, "Variational autoencoder for deep learning of images, labels and captions," 2016.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, X. Bing, and Y. Bengio, "Generative adversarial nets," *MIT Press*, 2014.

[3] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *International Conference on Computer Vision (ICCV)*, 2019.

[4] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, 2020.

[5] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8692–8701.

[6] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*.

[7] Francois Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," .

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "Transtrack: Multiple-object tracking with transformer," 2020.

[10] Ju He, Jieneng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, Changhu Wang, and Alan Yuille, "Transfg: A transformer architecture for fine-grained recognition," *arXiv preprint arXiv:2103.07976*.

[11] Deressa Wodajo and Solomon Atnafu, "Deepfake video detection using convolutional vision transformer," .

[12] Wei Lu, Lingyi Liu, Junwei Luo, Xianfeng Zhao, Yicong Zhou, and Jiwu Huang, "Detection of deepfake videos using long distance attention," .

[13] Young Jin Heo, Young Ju Choi, Young-Woon Lee, and Byung-Gyu Kim, "Deepfake detection scheme based on vision transformer and distillation," .

[14] Samira Abnar and Willem Zuidema, "Quantifying attention flow in transformers," 2020.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[16] "Deepfakes," https://github.com/deepfakes/faceswap, [Accessed: 2020-09-02].

[17] "Faceswap," https://github.com/MarekKowalski/FaceSwap, [Accessed: 2020-09-03].

[18] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] Justus Thies, Michael Zollhöfer, and Matthias Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, 2019.

[20] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," .

[21] B. Dolhansky, R Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (dfdc) preview dataset," 2019.

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[23] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[24] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Computer Vision – ECCV 2020*.

[25] H. Liu, X. Li, W. Zhou, Y. Chen, and N. Yu, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," 2021.

[26] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Computer Vision – ECCV 2020*.

[27] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu, "Multi-attentional deepfake detection," *arXiv preprint arXiv:2103.02406*, 2021.

[28] Junke Wang, Zuxuan Wu, Jingjing Chen, and Yu-Gang Jiang, "M2tr: Multi-modal multi-scale transformers for deepfake detection," *arXiv preprint arXiv:2104.09770*.

[29] C. Miao, Q. Chu, W. Li, T. Gong, W. Zhuang, and N. Yu, "Towards generalizable and robust face manipulation detection via bag-of-local-feature," 2021.

[30] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1–11.

[31] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," 2021.

[32] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola, "What makes fake images detectable? understanding properties that generalize," in *European Conference on Computer Vision*, 2020.

[33] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2019, pp. 1–8.

[34] Yuezun Li and Siwei Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.