2025 19th International Conference on Automatic Face and Gesture Recognition (FG)

Spatiotemporal Face Alignment for Generalizable Deepfake Detection

Alejandro Cobo[1], Roberto Valle[1], José M. Buenaposada[2], Luis Baumela[1]
[1] ETSI Informáticos, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, Boadilla del Monte, 28660, Madrid, Spain
[2] ETSII, Universidad Rey Juan Carlos, Calle Tulipán s/n, Móstoles, 28933, Madrid, Spain

2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG) | 979-8-3315-5341-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/FG61629.2025.11099346

Abstract— **Deepfake detection has progressively become a topic of interest in recent years due to the proliferation of automated facial forgery generation techniques that are able to produce manipulated media indistinguishable for the human eye. One of the most difficult aspects in deepfake detection is generalization to unseen manipulation techniques, which is a key factor to make a method useful in real world applications. In this paper, we propose a new multi-task network termed SFA, which leverages spatiotemporal features extracted from video inputs to provide more robust predictions compared to image-only models, as well as a face alignment task that helps the network to identify anomalous facial movements in the temporal dimension. We show that this multi-task approach improves generalization compared to the single-task baseline, and succeeds in producing results on par with the current state-of-the-art using different cross-dataset and cross-manipulation benchmarks.**

I. INTRODUCTION

Over the last decade, deep learning has proven to be an exceptionally useful tool to solve a wide variety of problems. One of these applications includes generative models [10], which can produce novel data from a learned distribution. “Deepfakes” are images or videos of human faces generated or manipulated from real data by deep learning techniques. With the increasing quality of these manipulations and the ease of use for the general public, they have raised high concerns due to the serious risk of malicious misusage they convey.

Thus, deepfake detection has quickly become a popular topic of research. Early works try to identify forgery artifacts in the pixel or frequency domains [17], [22], [26], but perform poorly on cross-dataset benchmarks. Other methods try to generate pseudo-fakes [5], [19], [29] to improve generalization, but can only detect spatial artifacts and miss temporal clues useful to detect manipulated videos that have been generated frame-by-frame.

Recently, many works have also used spatiotemporal information from video data to detect artifacts in the temporal dimension [14], [13], [38], [44]. Although there is some work on deepfake detection using facial landmarks [32], there is a lack of research to leverage facial alignment tasks to detect anomalous facial movements by modeling facial movement over time.

In this work, we develop a simple but effective multi-task network termed SFA (Spatiotemporal Face Alignment for deepfake detection), which combines spatiotemporal features extracted with a video backbone with face alignment, showing how both tasks can work together and lead to better generalization to unseen manipulations. To track the movement of the face in the video, we train the network to create *motion heatmaps* [6]. This is a novel approach in deepfake detection, as previous literature usually employs 2D landmark coordinates obtained by an external network to train recurrent or graph-based networks [32]. The main advantage of our approach is that landmark annotations are not needed at inference time.

In summary, the main contributions of this paper are as follows:

1) We combine deepfake detection with face alignment in videos, using both tasks to improve generalization to unseen manipulation methods not present in the training dataset. In contrast with previous work, we train the network to perform face alignment instead of relying on landmark annotations at inference time.
2) We compare single- and multi-channel motion heat maps to perform face alignment on videos.

II. RELATED WORK

Since the uprising of deep learning-based methods that can easily and realistically manipulate the identity and appearance of a face in images and videos [9], [21], [27], the field of deepfake detection has also developed significantly. One of the more desirable properties of forgery detection methods is the ability to generalize well against manipulation methods not seen in the training data. Some studies [40], [41] hint that a key factor in achieving good generalization lies in avoiding the use of features that may encourage overfitting to particular manipulation methods of the training dataset, such as the identity of the subject. Thus, generalizable methods learn to discard facial features that are not relevant to the task of deepfake detection.

The methods can be categorized into image and video-based models. Image-based methods include [19], [24], [43], which try to focus the attention of the network on certain parts of the face that are more likely to contain forgery artifacts. Some methods leverage frequency information to complement or substitute spatial information [23], [26], [30], [37] and are more robust to video compression. Another popular approach is to use contrastive learning techniques [11], [12], [16], [17], [31] to improve generalization against unseen manipulation methods. Cao et al. [3] train an auto-encoder network to reconstruct only real faces, focusing on

This work was supported by project ETHICFACE PID2022-137581OB-I00 funded by MCIU/AEI/10.13039/501100011033/ FEDER, UE

979-8-3315-5341-8/25/$31.00 ©2025 IEEE

anomalies found in the decoder features to detect deepfakes. Other methods generate pseudo-fakes as training data [5], [19], [29], [38] by blending 2 images of the same person or even the same image with a slightly modified version of itself. They are more difficult to detect, forcing the network to extract more robust features.

On the other hand, some methods use video data to extract more informative spatiotemporal features or to track some biological signals over time inherent to real faces, such as heart rate [7] or action units [1], [33]. Sun et al. [32] use a sequence of facial landmarks in a video clip to detect anomalous movements generated from deepfakes. Haliassos et al. [13], [14] employ pre-trained networks on audio-visual data to generate robust features suitable to detect manipulated videos. Some methods [38], [42], [44] show the proper training methodology of spatiotemporal deepfake detectors, since spatial information is generally more dominant and easy to spot in deepfake detection, and the temporal features tend to be ignored by the network if trained naively. Xu et al. [39] leverage a pre-trained image model and adapt it to video data by stacking several frames into a thumbnail image.

Similarly to [32], we incorporate facial landmark information to detect manipulated videos. However, relying on 2D landmark coordinates only as in [32] can greatly decrease the ability to generalize to unseen manipulations, since the network easily overfits to the particular fake movements of the training data. Instead, we use face alignment as a complementary task in a video model to act as regularization and help the network extract more useful features for cross-dataset and cross-manipulation scenarios. Another difference is that this task is incorporated into the network, so landmark annotations are not needed at inference time.

III. METHOD

Our framework consists of a multi-task video transformer network that leverages spatiotemporal features extracted from videos, coupled with a face alignment task that improves the single-task baseline. Firstly, the new task helps the network produce a set of more informative feature maps for deepfake detection, since the focus is placed on specific parts of the face, acting as regularization. Secondly, it encourages the network to pay more attention to the temporal dimension, since it must track the position of the landmarks throughout the video clip.

A. Classification

An overview of our model is shown in Fig. 1. The input video frames $F \in \mathbb{R}^{C\times T\times H\times W}$ are split into non-overlapping 3-dimensional cubes and projected into tokens $\mathbf{x} \in \mathbb{R}^{P\times D}$, where C refers to color channels, T is the number of frames, H and W are frame height and width, respectively, $P = \frac{T}{2}\cdot\frac{H}{16}\cdot\frac{W}{16}$ is the number of tokens and D is the token dimension.

A set of *task tokens* $\mathbf{x}_{task} \in \mathbb{R}^{(1+L)\times D}$ are appended to $\mathbf{x}$, one for the classification task and L for the face alignment task (one per facial landmark). This new set of tokens $\mathbf{x}' = \mathbf{x} \cup \mathbf{x}_{task}$ is then fed into K sequential transformer layers. From the output of the last layer, we extract the subset of transformed task tokens $\mathbf{x}'_{task}$.

Finally, the transformed classification token in $\mathbf{x}'_{task}$ is fed into a fully connected layer that outputs the probability that the input video contains a deepfake. This task is optimized with a binary cross-entropy loss, $\mathcal{L}_{bce}$.

B. Face alignment

Our framework incorporates a face alignment task that helps the network identify the structure of the face and track its movement in the input video clip. We represent this movement via motion heatmaps [6], which are a temporal aggregation of single-frame probability maps of facial landmark locations. The direction of this movement can be encoded with multi-channel heatmaps (see Fig. 2). We compare single and multi-channel approaches in our ablation study.

Given a set of L landmarks, we define a set of L task tokens, each representing the movement of a facial landmark in the video. At the output of the last encoder layer, we extract the L tokens and perform a linear projection followed by a sigmoid activation that converts each token into a motion heatmap $H \in \mathbb{R}^{M\times H'\times W'}$, where M is the number of motion channels, and H' and W' are height and width of the heatmap.

These heatmaps are optimized via the intersection-over-union loss:

$$\mathcal{L}_{iou} = 1 - \frac{1}{N\cdot L}\sum_{n,l}^{N,L}\frac{s\left(y_{nl}\cdot p_{nl}\right)}{s\left(y_{nl}^2\right) + s\left(p_{nl}^2\right) - s\left(y_{nl}\cdot p_{nl}\right)} \quad (1)$$

$$s\left(h\right) = \sum_{i=1}^{M}\sum_{j=1}^{H'}\sum_{k=1}^{W'} h_{ijk} \quad (2)$$

where N is the number of video clips in a training batch, and $y_{nl} \in \mathbb{R}^{M\times H'\times W'}$ and $p_{nl} \in \mathbb{R}^{M\times H'\times W'}$ are, respectively, the ground-truth and predicted motion heatmaps for the l-th landmark in the n-th video of the input batch.

The final loss is the sum of the classification and heatmap regression losses, $\mathcal{L} = \mathcal{L}_{bce} + \lambda\mathcal{L}_{iou}$, where λ is the weight of the heatmap regression task, and is set empirically.

IV. EXPERIMENTS

A. Experimental settings

Datasets. To clearly assess the performance of deepfake detection models, it is important to use manipulation methods not seen in the training data. The most common cross-dataset benchmark used in the literature sets **FaceForensics++** [27] as the training dataset. It consists of 1000 videos obtained from YouTube and 4 manipulation methods (Deepfakes[1], FaceSwap[2], Face2Face [35], NeuralTextures [34]) applied to them, for a total of 4000 manipulated videos.

For cross-dataset evaluation, we employ **CelebDF-v2** [21], which consists of YouTube videos tampered with a more

[1] https://github.com/deepfakes/faceswap

[2] https://github.com/MarekKowalski/FaceSwap

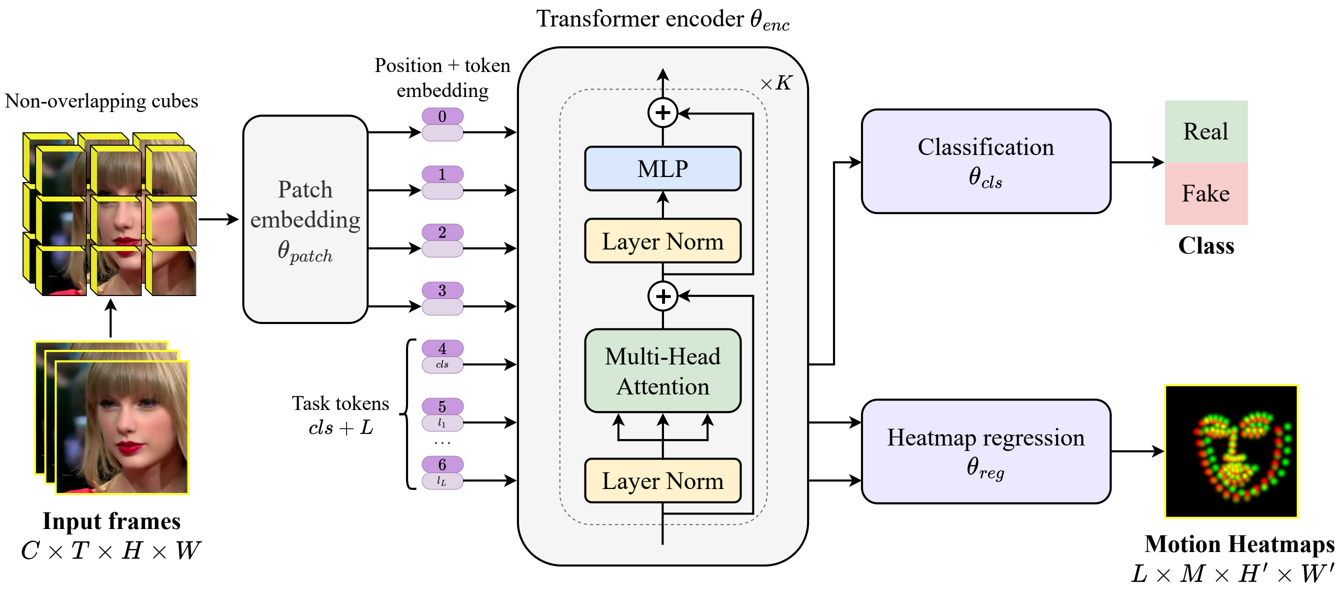
Transformer encoder θ_{enc}
Non-overlapping cubes
Position + token embedding
Patch embedding θ_{patch}
0
1
2
3
4
cls
5
l_1
6
l_L
Task tokens $cls + L$
$\times K$
MLP
Layer Norm
Multi-Head Attention
Layer Norm
Classification θ_{cls}
Real
Fake
Class
Heatmap regression θ_{reg}
Input frames
$C \times T \times H \times W$
Motion Heatmaps
$L \times M \times H' \times W'$

Fig. 1: Overview of our method. A pre-trained video transformer encoder [2] is given a set of learnable task tokens, which are used for a deepfake detection task and a motion heatmap regression task. The latter helps to extract more discriminative features for better generalization as well as forcing the network to focus on the movement of the face in the video input, leveraging temporal clues to detect anomalous facial movements.

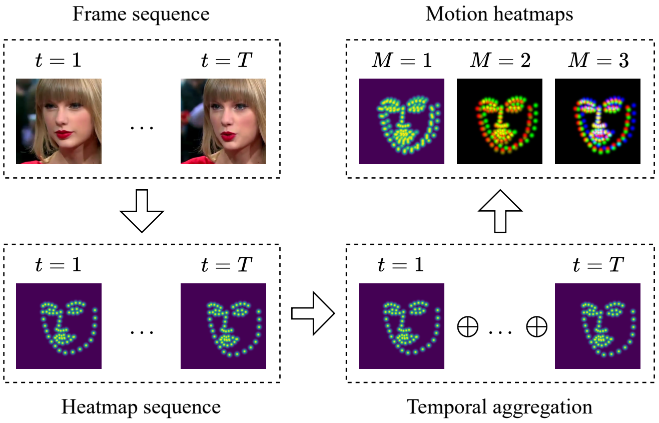

Fig. 2: Ground-truth motion heatmap generation process for different number of channels M [6]. For visual clarity, heatmaps for all facial landmarks are shown together.

advanced face swapping algorithm for a total of 518 testing videos; and **DFDCP** [9], with 780 testing video clips generated with two in-house face swapping methods and randomly subjected to data augmentation.

As cross-manipulation datasets, we use **FaceShifter** [18], which applies a two-stage occlusion-aware face swapping network, and **DeeperForensics-1.0** [15], which uses a Variational Auto-Encoder that generates face-swap images as well as optical flow information, improving temporal continuity. Both benchmarks are based on the same real testing videos as FaceForensics++.

Evaluation metrics. Following previous work, we report the video level area under the receiver operating characteristic curve (AUC) in our tables, that is, the average AUC between all the clips in a video. This ensures a fair comparison between image and video-based methods.

Implementation details. We first detect faces on each video with the RetinaFace detector [8], and expand each bounding box by a factor of 1.3 around its center. We align the faces to the center of the frame by moving the center of the bounding box to the landmark corresponding to the tip of the nose detected by RetinaFace. To obtain the landmarks used for generating the ground-truth heatmaps in Eq. 1, we employ a state-of-the-art face alignment method [25]. The transformer encoder used in our experiments is pre-trained on a large-scale facial video reconstruction task [2]. Our network processes video clips of 16 consecutive 224 × 224 frames, and we use $H' = W' = 64$ in Eq. 2 as the output size of our heatmap regression head. The weight of the heatmap regression loss function λ is set to 1.0. We use a batch size of 8 and over-sample the real class on the training set to match the number of fake videos. We use the Adam optimizer with a maximum learning rate of $7.07 \cdot 10^{-6}$, obtained empirically, and a cosine annealing learning rate scheduler with linear warm-up. Data augmentation includes hue and brightness manipulation, affine transformations, image compression, Gaussian blur and cutout, applied uniformly to all frames of a video clip.

B. Comparison with the state-of-the-art

Table I shows a comparison of our method with several state-of-the-art image (first half) and video-based (second half) deepfake detectors. All methods are trained on FaceForensics++ high-quality subset (FF++ HQ), and tested on CelebDF-v2 (CDF), Deepfake Detection Challenge Preview

TABLE I: Comparison with the state-of-the-art in terms of video-level AUC (%). All models are trained on FaceForensics++ HQ and evaluated on CelebDF-v2 (CDF), Deepfake detection challenge preview (DFDCP), FaceShifter (FSh) and DeeperForensics (DFo). * indicates results computed by us with official model weights, otherwise taken from [38], [39].

Method	**Cross-dataset**		**Cross-manipulation**		**Avg.**
	CDF	**DFDCP**	**FSh**	**DFo**	
FWA [20]	69.50	-	65.50	50.20	-
PatchForensics [4]	69.60	-	57.80	81.80	-
Xception [27]	73.70	-	72.00	84.50	-
CNN-aug [36]	75.60	-	65.70	74.40	-
Multi-Att [43]	75.70	-	66.00	77.70	-
Face X-Ray [19]	79.50	**80.92**	92.80	86.80	85.01
SLADD [5]	79.70	-	-	-	-
CNN-GRU [28]	69.80	-	80.80	74.10	-
LipForensics [14]	82.40	-	97.10	97.60	-
ISTVT [42]	84.10	74.20	99.30	98.60	-
FTCN [44]	86.90	74.00	98.80	98.80	89.63
RealForensics [13]	86.90	-	99.70	99.30	-
AltFreezing [38]	89.50	70.91 *	99.40	99.30	89.78
TALL-Swin [39]	**90.79**	-	99.67	**99.62**	-
SFA (ours)	89.52	80.58	**99.84**	99.24	**92.30**

TABLE II: Ablation study with different number of channels of motion heatmaps M. All metrics represent video-level AUC (%).

Method	**Cross-dataset**		**Cross-manipulation**		**Avg.**
	CDF	**DFDCP**	**FSh**	**DFo**	
Baseline	87.17	78.45	99.55	98.50	90.92
$M = 1$	**89.70**	79.84	99.82	99.17	92.13
$M = 2$	89.52	**80.58**	**99.84**	**99.24**	**92.30**

(DFDCP), FaceShifter (FSh) and DeeperForensics-1.0 (DFo) datasets. Note that some of the methods are trained on pseudo-fakes instead of the original fake videos of FF++. **Best** and second best results are shown in bold and underlined, respectively.

We can see that our method can achieve very competitive results in all testing datasets, and even establishes new state-of-the-art on FaceShifter. This shows that modeling the movement of the face as a high-level source of information incorporated in the model has a positive effect on the deepfake detection task. Overall, results show the generalization capabilities of our method to unseen domains and manipulation techniques not seen in the training dataset.

C. Ablation study

Table II shows the performance of our method when using different number of channels in the motion heatmaps. First, compared to the baseline detector, which includes only a classification token and lacks the face alignment task, we can see that adding the heatmap regression task consistently improves the average results considering all test datasets. This highlights the usefulness of the face alignment task, as the network is forced to extract more robust features and track the movement of the face in the input video. This is important to detect forgery clues in the temporal dimension, as deepfake videos are usually crafted frame-by-frame without considering any temporal consistency.

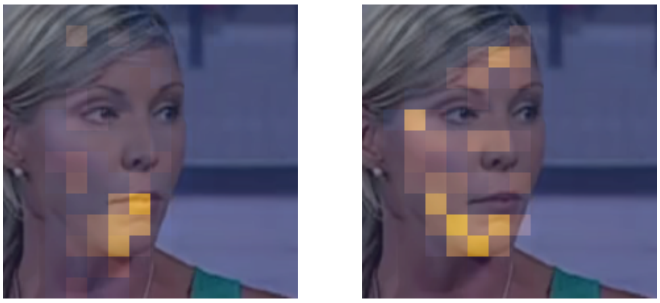
(a) Single-task.
(b) Multi-task.

Fig. 3: Visualization of the attention map of the class token for single- and multi-task networks in a manipulated video from FaceForensics++. The baseline network (3a) fails to recognize the deepfake. The multi-task approach (3b) focuses on different parts of the face and correctly classifies it.

This results also show that using heatmaps that model the direction of the movement (M=2) slightly improves the single-channel approach (M=1). This is notable in DFDCP, where videos have been altered with several augmentations, such as reduced encoding quality and image downsampling. Since the pixel-level data of these videos is less informative, having access to a high-level representation of the face can improve the performance of the network. With more motion channels, the face alignment task becomes more difficult, and the network is forced to increase its importance. In contrast, deepfake videos in CDF can be easily spotted by looking at spatial discrepancies in the face, such as color and resolution differences caused by the face swapping procedure, and high-level representations become less useful.

Additionally, we show in Fig. 3 the effect of the face alignment task on the attention maps of the classification token aggregated over the temporal dimension. We can see that, when the model is trained with both tasks, the attention of the network to relevant parts of the face is more notorious. This is a result of the regularization effect of the face alignment task, as the network is more likely to extract features more related to the structure of the face, disregarding irrelevant parts of the input frames.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we present a novel approach to the challenging task of detecting deepfakes. We showed how a simple framework can leverage face alignment on videos to improve the generalization ability of a baseline detector.

Since face alignment is a useful tool for locating the face and its different parts in the image, we believe that this work can be further extended to improve the interpretability of the deepfake detector by accurately showing which parts of the face have been manipulated. This could also be used as a complement to pseudo-fake generation techniques extended to videos.

VI. ACKNOWLEDGMENTS

José M. Buenaposada and Luis Baumela are members of the Madrid Ellis unit, funded by the Regional Government of Madrid. Alejandro Cobo was funded by a predoctoral contract from Universidad Politécnica de Madrid.

ETHICAL IMPACT STATEMENT

Manipulated media derived from deepfake generation technologies present a serious risk to society as they can be used for malicious purposes, such as face spoofing or identity theft. Because of the fast development of AI, it is very difficult to foresee the actual generalization capabilities of a deepfake detector, and future manipulation techniques can include the detector predictions to generate data more robust to that specific model. Thus, we cannot guarantee the applicability of our work to all deepfake detection contexts in the future, which can lead to false negative predictions and cause a sense of false security among its users.

To address this issue, one strategy to follow is to continually improve the training data with more modern fake generation techniques and retrain the network to keep it up-to-date.

Despite the potential risks, we believe that our research can have a positive impact on society by exposing manipulated media and preventing malicious uses of deepfakes.

REFERENCES

[1] W. Bai, Y. Liu, Z. Zhang, B. Li, and W. Hu. Aunet: Learning relations between action units for face forgery detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 24709–24719, 2023.

[2] Z. Cai, S. Ghosh, K. Stefanov, A. Dhall, J. Cai, H. Rezatofighi, R. Haffari, and M. Hayat. MARLIN: masked autoencoder for facial video representation learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1493–1504, 2023.

[3] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4103–4112, 2022.

[4] L. Chai, D. Bau, S. Lim, and P. Isola. What makes fake images detectable? understanding properties that generalize. In *Proc. European Conference on Computer Vision*, pages 103–120, 2020.

[5] L. Chen, , Y. Zhang, Y. Song, L. Liu, and J. Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 18689–18698, 2022.

[6] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid. Potion: Pose motion representation for action recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7024–7033, 2018.

[7] U. A. Ciftci and I. Demir. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2019.

[8] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint*, arxiv:1905.00641, 2019.

[9] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton-Ferrer. The deepfake detection challenge (DFDC) preview dataset. *arXiv preprint*, arXiv:1910.08854, 2019.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Conference on Neural Information Processing Systems*, pages 2672–2680, 2014.

[11] Q. Gu, S. Chen, T. Yao, Y. Chen, S. Ding, and R. Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *AAAI Conference on Artificial Intelligence*, page 735–743, 2022.

[12] Y. Guo, C. Zhen, and P. Yan. Controllable guide-space for generalizable face forgery detection. In *Proc. International Conference on Computer Vision*, pages 20818–20827, 2023.

[13] A. Haliassos, R. Mira, S. Petridis, and M. Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 14930–14942, 2022.

[14] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5039–5049, 2021.

[15] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2886–2895, 2020.

[16] N. Larue, N. Vu, V. Struc, P. Peer, and V. Christophides. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proc. International Conference on Computer Vision*, pages 20954–20964, 2023.

[17] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 6458–6467, 2021.

[18] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint*, arxiv:1912.13457, 2019.

[19] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face x-ray for more general face forgery detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 5000–5009, 2020.

[20] Y. Li and S. Lyu. Exposing deepfake videos by detecting face warping artifacts. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 46–52, 2019.

[21] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3204–3213, 2020.

[22] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 772–781, 2021.

[23] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *Proc. European Conference on Computer Vision*, pages 667–684, 2020.

[24] D. Nguyen, N. Mejri, I. P. Singh, P. Kuleshova, M. Astrid, A. Kacem, E. Ghorbel, and D. Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 17395–17405, 2024.

[25] A. Prados-Torreblanca, J. M. Buenaposada, and L. Baumela. Shape preserving facial landmarks with graph attention networks. In *British Machine Vision Conference*, 2022.

[26] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proc. European Conference on Computer Vision*, pages 86–103, 2020.

[27] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proc. International Conference on Computer Vision*, 2019.

[28] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 80–87, 2019.

[29] K. Shiohara and T. Yamasaki. Detecting deepfakes with self-blended images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 18699–18708, 2022.

[30] L. Song, Z. Fang, X. Li, X. Dong, Z. Jin, Y. Chen, and S. Lyu. Adaptive face forgery detection in cross domain. In *Proc. European Conference on Computer Vision*, pages 467–484, 2022.

[31] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji. Dual contrastive learning for general face forgery detection. In *AAAI Conference on Artificial Intelligence*, pages 2316–2324, 2022.

[32] Z. Sun, Y. Han, Z. Hua, N. Ruan, and W. Jia. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3609–3618, 2021.

[33] L. Tan, Y. Wang, J. Wang, L. Yang, and Y. G. Xunxun Chen and. Deepfake video detection via facial action dependencies estimation. In *AAAI Conference on Artificial Intelligence*, pages 5276–5284, 2023.

[34] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: image synthesis using neural textures. *ACM Transactions on Graphics*, 38:66:1–66:12, 2019.

[35] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.

[36] S. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 8692–8701, 2020.

[37] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 7278–7287, 2023.

[38] Z. Wang, J. Bao, W. Zhou, W. Wang, and H. Li. Altfreezing for more general video face forgery detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 4129–4138, 2023.

[39] Y. Xu, J. Liang, G. Jia, Z. Yang, Y. Zhang, and R. He. TALL: thumbnail layout for deepfake video detection. In *Proc. International Conference on Computer Vision*, pages 22601–22611, 2023.

[40] Z. Yan, Y. Zhang, Y. Fan, and B. Wu. UCF: uncovering common features for generalizable deepfake detection. In *Proc. International Conference on Computer Vision*, pages 22355–22366, 2023.

[41] K. Yao, J. Wang, B. Diao, and C. Li. Towards understanding the generalization of deepfake detectors from a game-theoretical view. In *Proc. International Conference on Computer Vision*, pages 2031–2041, 2023.

[42] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang. ISTVT: interpretable spatial-temporal video transformer for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 18:1335–1348, 2023.

[43] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu. Multi-attentional deepfake detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2185–2194, 2021.

[44] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen. Exploring temporal coherence for more general video face forgery detection. In *Proc. International Conference on Computer Vision*, pages 15024–15034, 2021.