Dual-PST: Dual-Branch SpatioTemporal-Planar Network for Video Forgery Detection

1st Siyu Liu
School of Computer Science and Technology
Anhui University
Hefei, China
liusiyu0102@gmail.com

2nd Zhida Zhang
MAIS&NLPR, Institute of Automation
Chinese Academy of Sciences
Beijing, China
zhida.zhang@cripac.ia.ac.cn

3rd Junxian Duan
MAIS&NLPR,Institute of Automation
Chinese Academy of Sciences
Beijing, China
junxian.duan@ia.ac.cn

4th Jie Cao*
MAIS&NLPR,Institute of Automation
Chinese Academy of Sciences
Beijing, China
Corresponding author: jie.cao@cripac.ia.ac.cn

5th Aihua Zheng
School of Artificial Intelligence
Anhui University
Hefei, China
ahzheng214@foxmail.com

***Abstract*—With the advancement of generative AI, distinguishing real and AI-generated faces in videos has become increasingly challenging. However, traditional methods struggle to capture local details and temporal dynamics simultaneously, making it difficult to achieve high detection accuracy while maintaining low computational overhead. To address this problem, we propose a Dual-branch SpatioTemporal-Planar Network (Dual-PST) based on the selective state-space model. It is capable of extracting image features and temporal relations simultaneously, while maintaining linear computational consumption. Specifically, we design a Multi-Selective State-Space module (MS3) that can extract global features from image typography consisting of consecutive video frames by scanning them in multiple sequences. To further enhance temporal modeling capabilities, we propose a Sequential Tri-frame Local module, which captures inter-frame temporal relationships and local features by temporally splicing single-frame features. These features are first extracted using MS3 and then further enhanced through inter-frame masking operations. Experimental results show that Dual-PST significantly improves detection accuracy while maintaining low computational complexity and strong model robustness.**

***Index Terms*—Face Forgery Detection, Spatiotemporal Inconsistencies, Selective State-Space Model**

I. Introduction

In recent years, research in face forgery has experienced remarkable advancements [1], [2], [3]. Forgery images may deceive facial recognition systems or allow malicious exploitation, leading to societal trust issues and security risks. Consequently, there is a compelling need to explore reliable methods for face forgery detection.

The core of deepfake detection lies in identifying subtle clues that distinguish real images from synthetic ones. Since deepfake algorithms often operate on a frame-by-frame basis, the generated videos often exhibit both spatial and temporal artifacts [4], [5], [6]. Modeling only one aspect (spatial or temporal) may not be sufficient to cover all types of artifacts.

Although some studies have employed spatiotemporal neural networks to effectively detect temporal inconsistency in videos [7], [8], these methods primarily focus on spatial feature extraction within individual frames and fail to effectively incorporate temporal relationships into local feature extraction, thus overlooking the temporal dimension. In addition, the frame-by-frame processing approach further increases computational complexity, making it difficult for the model to achieve an optimal balance between accuracy and efficiency.

To deal with the above challenges, we propose a **D**ual-branch **S**patio**T**emporal-**P**lanar network (Dual-PST). As shown in Fig. 1, before entering the global and local branches, we divide the given video into segments and randomly selected several consecutive frames (default is 4) arranged into a specific spatiotemporal layout. This layout method stitches consecutive single-frame images into a plane image, effectively transforming the video's temporal sequence information into a static spatial structure, which can reduce the computational complexity of frame-by-frame processing.

For the global branch, we utilize the Multi-Selective State-Space (MS3) module to scan the spatiotemporal layout in rows and columns, expanding it into multiple sequences according to the four orders shown in Fig. 1 (a), allowing each pixel to integrate information from different directions. This method combines the linear computational complexity with the global receptive field while maintaining high computational efficiency.

To enhance the temporal relationship modeling of local features, the spatiotemporal layout is also input into the local branch. Unlike the global branch which processes the entire layout, the local branch inputs each frame individually into the MS3 module, as shown in Fig. 1 (b). We then design a Sequential Tri-frame Local (STL) module to explicitly incorporate temporal relations into the local feature extraction process. Each frame first undergoes a masking operation while excluding it from the 2 × 2 layout, then the remaining three frames are spliced together in chronological order to generate local features. By repeating this process for each frame, we

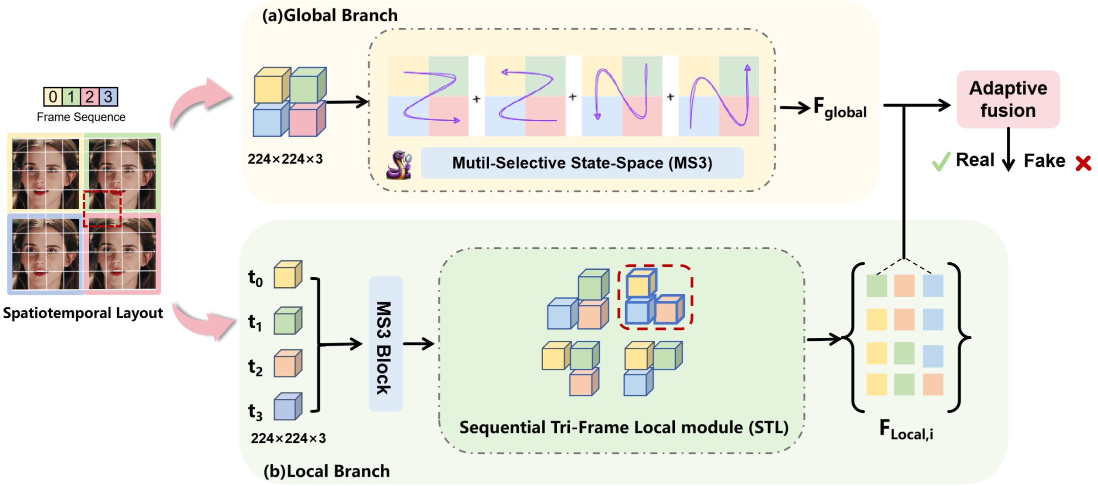

Fig. 1: Workflow of the Dual-PST architecture: a) The global branch uses MS3 to process the downsampled spatiotemporal layout to extract multi-scale features; b) The local branch processes each frame independently, applying the STL to combine adjacent frames, preserving temporal order and extracting local details and temporal dynamics. Finally, an adaptive weighted mechanism fuses the features for the final detection result.

obtain four local features. These local features are stitched together in strict chronological order, capturing not only the details of each frame but also the dynamic relationships between video frames. To effectively fuse the features from the global and local branches, we design an Adaptive Fusion module and develop an adaptive weighted multi-task loss function, which dynamically adjusts the weights of each loss component to balance their impact on the final decision.

We conducted extensive experiments and validated the contribution of each key component through ablation studies. Experiments demonstrate that these improvements boost Dual-PST's spatiotemporal feature modeling while maintaining efficiency, making it a strong solution for face video forgery detection.

II. Methodology

A. *Mutil-Selective State-Space (MS3)*

The Mamba [9] introduces a **S**elective **S**tate-**S**pace **M**odel, ensuring that its computational complexity scales linearly with the length of the sequence, as opposed to the quadratic scaling typical of transformers, thereby performing exceptionally well with long sequence data. Building on Mamba's Selective State-Space Model, we propose the **M**ulti-**S**elective **S**tate-Space (**MS3**) module and apply it to the visual domain.

To better process image data, we expand it into a sequence by scanning the image pixel tokens. Unlike traditional single-directional traversal methods for processing images [10], the MS3 module scans in four directions (from top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right) as shown in Fig. 1 (a), converting image information into four sequences. These sequences integrate pixel information from different directions, ensuring that each pixel captures its own spatiotemporal relationships while sharing feature information with other frames. Finally, these sequences are merged into a unified global feature representation that effectively reflects the spatiotemporal dynamics and global image features within the image.

B. *Sequential Tri-frame Local (STL)*

In the local branch, we process each frame in the spatiotemporal layout independently. Each frame is sequentially input into the MS3 for four-directional scanning, moving from each corner to its diagonal counterpart. This approach ensures that each frame integrates global information from different directions, capturing spatiotemporal inconsistencies between frames while maintaining spatial continuity.

To explicitly incorporate temporal relationships into the process of local feature extraction, we propose an innovative **S**equential **T**ri-frame **L**ocal (**STL**) module, as illustrated in Fig. 1 (b). Specifically, after the MS3 module extracts features from each frame, we traverse the features of the spatiotemporal layout. For each frame, a masking operation is applied to exclude it from the 2×2 layout, ensuring that feature extraction relies only on the spatiotemporal information from the remaining three frames. Next, the remaining three frame features are arranged according to the time order of the original spatiotemporal layout to form a local feature. By repeating this operation for each frame in the Four-Frame Layout, we eventually generate four local features. These local features are strictly arranged in chronological order, allowing the model to not only capture the static information of individual frames but also model the dynamic relationships between frames, significantly enhancing the model's spatiotemporal feature modeling capabilities.

C. *Dual-branch Spatiotemporal-Planar Network (Dual-PST)*

1) Dual-Branch Feature Extraction: For a video $\mathbf{V} \in \mathbb{R}^{T\times C\times H\times W}$, where T is the frame length, C is the number of channels, and $H\times W$ is the resolution of the frames, we divide the video into N equal clips, each with a length of T/N. From each clip, we randomly sample t consecutive frames (default is 4) to form a spatiotemporal layout. These four frames are arranged according to a specific layout and used as input for both the global and local branches to process independently. The four frames are arranged in a specified layout and then used as input for both the global and local branches to process independently.

In the global branch, we first adjust the spatiotemporal layout to match the original image size, then input it into the MS3 for global feature extraction. Through this layout and scanning method, we can extract not only spatial features but also temporal features by capturing the relationships between frames. This multi-directional scanning combines MS3 with the spatiotemporal layout, enabling effective integration of pixel information from different directions across the four-frame images, ensuring that each pixel captures contextual information from multiple perspectives. This approach achieves global receptive field coverage without increasing the computational complexity, making feature extraction both efficient and comprehensive. The extracted global features $\mathbf{F}_{\text{global}}$ are used to generate the global prediction $\mathbf{p}_{y|\mathbf{F}_{\text{global}}}$ through a linear layer. During the embedding process, the global features are compressed to improve feature fusion efficiency and reduce computational complexity, ultimately producing the compressed global prediction $\mathbf{p}_{y|\mathbf{z}}$. We introduce Global Information Loss $\mathcal{L}_{GIL}$ to ensure that the compressed features retain key information and distribution consistency, enabling efficient global feature modeling while minimizing complexity.

In the local branch, MS3 is used to independently process each frame in the spatiotemporal layout. To effectively integrate temporal relationships, we use the STL module to process each frame, ultimately generating four local features $\mathbf{F}_{\text{local},i}$. These local features are then mapped into the embedding space through a linear layer, producing the corresponding local predictions $\mathbf{P}_{y|\mathbf{f}_i}$. To ensure the independence and integrity of these local features and to minimize the impact of redundant information, we introduce Local Information Loss $\mathcal{L}_{LIL}$.

The calculation is defined as follows:

$$\mathcal{L}_{GIL} = \mathrm{KL}\left(\mathrm{Softmax}\left(\frac{\mathbf{p}_{y|\mathbf{F}_{\text{global}}}}{T}\right), \mathrm{Softmax}\left(\frac{\mathbf{p}_{y|\mathbf{z}}}{T}\right)\right), \quad (1)$$

$$\mathcal{L}_{LIL} = \sum_{i\neq j} \mathrm{KL}\left(\mathrm{Softmax}\left(\frac{\mathbf{P}_{y|\mathbf{f}_i}}{T}\right), \mathrm{Softmax}\left(\frac{\mathbf{P}_{y|\mathbf{f}_j}}{T}\right)\right), \quad (2)$$

where **KL** represents the Kullback-Leibler Divergence, and T is a temperature parameter that controls the smoothness of the softmax output.

2) Adaptive weighted fusion: To effectively fuse the features extracted by the global and local branches in the Dual-PST architecture, we designed an adaptive weighted multi-task loss function. The total loss function $\mathcal{L}_{\text{total}}$ consists of three main components:

$$\mathcal{L}_{Total} = \lambda_1 \cdot \mathcal{L}_{CE}(\mathbf{p}_{y|\mathbf{z}}, y) + \lambda_2 \cdot \mathcal{L}_{GIL} + \lambda_3 \cdot \mathcal{L}_{LIL}. \quad (3)$$

In Eq. (3), $\mathcal{L}_{CE}(\mathbf{p}_{y|\mathbf{z}}, y)$ represents the cross-entropy loss between the compressed global prediction and the ground truth labels. λ_1, λ_2, and λ_3 are the weighting coefficients of the loss function. To balance the impact of different loss components on the final decision, this loss function incorporates an adaptive weighting mechanism, dynamically adjusting the weight of each loss term λ_i through learnable parameters σ_i:

$$\lambda_i = \frac{0.5}{\sigma_i^2} + \log(1 + \sigma_i^2). \quad (4)$$

This mechanism allows the model to automatically adjust the importance of global and local features during training, thereby improving classification performance.

III. Experiments

We evaluated our model on three commonly used datasets: FaceForensics++(FF++) [11], Celeb-DF [12] and DFDC [13]. FF++ contains 1,000 original videos and 4,000 manipulated videos generated by four typical forgery methods, with two quality levels (C23 and C40). Celeb-DF includes 590 real videos and 5,639 high-quality fake videos which are crafted by the improved DeepFake algorithm. DFDC is a large-scale dataset that contains 128,154 facial videos of 960 subjects. We used MTCNN [14] to detect faces in the video frames. The backbone model employed is the Visual State Space Model [15] pretrained on ImageNet-1K. We used the Adam optimizer with a learning rate of 1.5e-4 and a batch size of 4, along with a cosine annealing scheduler with 10 linear warm-up epochs. The evaluation metrics are Accuracy (Acc) and the Area Under the Receiver Operating Characteristic Curve (AUC), with comparison results sourced from their respective papers. The best performance is highlighted in bold.

A. Intra-dataset Performance

In this section, We use 720 training videos, 140 validation videos, and 140 test videos out of every 1000 videos following the methodology [11]. Because of the data distribution problem in the FF++ dataset (1:4 ratio of true to false), we usually consider the AUC to be more informative than the Acc, and we achieve the highest value as shown in Table. I. Specifically, our method achieves an AUC of 100% on the FF++ (C23) dataset, outperforming the TALL-Swin method. Under the FF++ (C40) quality setting, our approach shows a performance improvement of 1.81% over TALL-Swin. We also test our method on the Celeb-DF dataset, where it demonstrates outstanding performance, achieving 100% in both Acc and AUC metrics.

B. Cross-dataset Performance

In this section, we evaluated the generalization ability of the model. We trained the model on FF++ (C40) and then tested it on the Celeb-DF and DFDC datasets. As shown in Table. II, our method demonstrates significant improvement on unseen datasets. On the Celeb-DF dataset, the AUC increased from 76.70% (Two-branch) to 82.05%; on the DFDC dataset, the AUC improved from 69.06% (RECCE) to 74.70%. The

TABLE I: **Intra-dataset evaluations.** We report the Acc(%) and AUC (%) on the FaceForensics++ dataset.

Method	FF++ (C23)		FF++ (C40)	
	Acc ↑	AUC ↑	Acc ↑	AUC ↑
MesoNet[16]	83.10	–	70.47	
Xception[17]	95.73	96.30	86.86	89.30
Face X-Ray[18]	–	87.40	–	61.60
Two-branch[19]	96.43	98.70	86.34	86.59
RFM[20]	95.69	98.79	87.06	89.83
Add-Net[21]	96.78	97.74	87.50	91.01
F3-Net[22]	97.52	98.10	90.43	93.30
Multi-Att[23]	97.60	99.29	88.69	90.40
FDFL[24]	96.69	99.30	89.00	92.40
DIANet[25]	96.37	98.80	89.77	88.20
UIA-ViT[26]	96.06	98.97	86.71	89.62
RECCE[27]	97.06	99.32	91.03	95.02
ITA-SIA[28]	97.64	99.35	90.23	93.45
DisGRL[29]	97.69	99.48	91.27	95.19
TALL-Swin[10]	**98.65**	99.87	**92.82**	94.57
Dual-PST (ours)	98.43	**100.00**	92.29	**96.38**

TABLE II: **Cross-dataset comparison results.** We report the AUC (%) on two unseen datasets: Celeb-DF and DFDC.

Method	Training dataset	Celeb-DF	DFDC
Xception[17]	FF++(c40)	61.80	63.61
F3-Net[22]		61.51	64.60
Add-Net[21]		65.29	64.78
Multi-Att[23]		67.02	68.01
RECCE[27]		68.71	69.06
Two-branch[19]		76.70	–
M2TR[30]		72.05	66.02
Dual-PST (ours)	FF++(c40)	**82.05**	**74.70**

results demonstrate that our method performs exceptionally well on unseen datasets, with better generalization ability than previous methods.

C. Ablation Study

We designed a series of ablation study in FF++(C23) and Celeb-DF to validate the effectiveness of the modules in Dual-PST. We train the VSSM model on the FF++ (C23) dataset and test it on the Celeb-DF dataset. To further validate the effectiveness of VSSM, we compare it against the Swin Transformer (Swin-T) [31].

We first scan the spatiotemporal layout using Swin-T, then replace Swin-T with MS3, the model's performance significantly improved both within and across datasets. Next, we introduced the STL module while keeping Swin-T unchanged and observed further performance improvements. Particularly for the Celeb-DF dataset, the AUC increased by 2.42%, which is a larger impact than that of MS3. This is due to the fact that the distribution of image features varies between datasets, but STL can better extract the inherent temporal relationships within the video. As shown in Table. III, the experimental results demonstrate that both the MS3 and STL modules make significant contributions to spatiotemporal feature extraction in video analysis.

TABLE III: **Ablation study of Dual-PST.** We show Acc (%) and AUC (%) training on FF++ (C23) and testing on Celeb-DF

Methods	Variant	FF++ (C23)		Celeb-DF	
		Acc ↑	AUC ↑	Acc ↑	AUC ↑
1	Swin-T	93.53	96.43	76.96	81.60
2	MS3	98.43	99.67	79.04	82.70
3	Swin-T+STL	96.86	99.09	78.79	84.02
4	**Dual-PST**	**98.43**	**100.00**	**79.84**	**84.28**

In addition, we also explore the effect of the number of local frames in the STL module. As shown in Fig. 2, compared to single-frame extraction, two-frame extraction provides some temporal information to the model, thus improving its capability when the number of frames rises to three, the model capability gains a large boost. This is because three-frame splicing not only adds information from an additional frame but also includes information about the "masked" frame, avoiding the information loss and incomplete features seen with single-frame splicing.

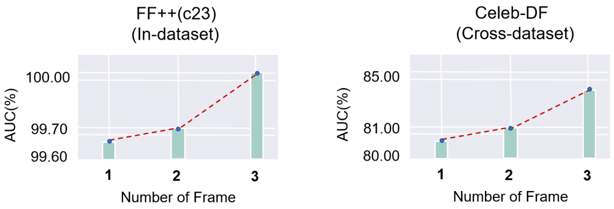

Fig. 2: Ablation Study on Local Frame Numbers in the STL Module.

Conclusion

In this paper, we propose the Dual-PST network architecture, which extracts both image features and temporal relations by stitching consecutive frames of video into pictures with low computational consumption. Specifically, the MS3 module efficiently captures global spatiotemporal features through multi-directional scanning while maintaining linear computational complexity. The STL module incorporates inter-frame temporal information, enhancing the spatiotemporal consistency of local features. Together, these components effectively capture both global and local features in video data. Experimental results demonstrate that Dual-PST performs exceptionally well across various datasets, particularly showing strong generalization on unseen datasets, indicating its potential for real-world deepfake detection.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (Grant No. 62206277), the University Synergy Innovation Program of Anhui Province (Grant GXXT-2022-036).

References

[1] Y. Zang, Y. Zhang, M. Heydari, and Z. Duan, "Singfake: Singing voice deepfake detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 156–12 160.

[2] C. Fu, Y. Hu, X. Wu, G. Wang, Q. Zhang, and R. He, "High-fidelity face manipulation with extreme poses and expressions," vol. 16, p. 2218–2231, 2021.

[3] J. Cao, Y. Hu, B. Yu, R. He, and Z. Sun, "3d aided duet gans for multi-view face image synthesis," *IEEE/CVF Transactions on Information Forensics and Security*, pp. 2028–2042, 2019.

[4] Z. Ba, Q. Liu, Z. Liu, S. Wu, F. Lin, L. Lu, and K. Ren, "Exposing the deception: Uncovering more forgery clues for deepfake detection," in *AAAI Conference on Artificial Intelligence*, 2024, pp. 719–728.

[5] Z. Lei, S. Liao, R. He, M. Pietikainen, and S. Li, "Gabor volume based local binary pattern for face representation and recognition," in *IEEE/CVF International Conference on Automatic Face and Gesture Recognition*, 2008, pp. 1–6.

[6] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, and L. Ma, "Delving into the local: Dynamic inconsistency learning for deepfake video detection," in *AAAI Conference on Artificial Intelligence*, 2022, pp. 744–752.

[7] Y. Lai, G. Yang, Y. He, Z. Luo, and S. Li, "Selective domain-invariant feature for generalizable deepfake detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 2335–2339.

[8] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, F. Huang, and L. Ma, "Spatiotemporal inconsistency learning for deepfake video detection," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 3473–3481.

[9] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[10] Y. Xu, J. Liang, G. Jia, Z. Yang, Y. Zhang, and R. He, "Tall: Thumbnail layout for deepfake video detection," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22 658–22 668.

[11] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "Faceforensics++: Learning to detect manipulated facial images," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11.

[12] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3204–3213.

[13] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Cristian Ferrer, "The deepfake detection challenge dataset," *ArXiv:2006.07397*, 2020.

[14] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," pp. 1499–1503, 2016.

[15] Y. Liu, Y. Tian, Y. Zhao, H. Yu, X. L.X., Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[16] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1–7.

[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800–1807.

[18] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5000–5009.

[19] I. Masi, A. Killekar, R. Mascarenhas, S. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 667–684.

[20] C. Wang and W. Deng, "Representative forgery mining for fake face detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 923–14 932.

[21] B. Zi, M. Chang, J. Chen, X. Ma, and Y. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 2382–2390.

[22] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 86–103.

[23] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2185–2194.

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6458–6467.

[25] Z. Hu, H. Xie, Y. Wang, J. Li, Z. Wang, and Y. Zhang, "Dynamic inconsistency-aware deepfake video detection," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021, pp. 736–742.

[26] W. Zhuang, Q. Chu, Z. Tan, Q. Liu, H. Yuan, C. Miao, Z. Luo, and N. Yu, "Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 391–407.

[27] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4113–4122.

[28] K. Sun, H. Liu, T. Yao, X. Sun, S. Chen, S. Ding, and R. Ji, "An information theoretic approach for attention-driven face forgery detection," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 111–127.

[29] Z. Shi, H. Chen, L. Chen, and D. Zhang, "Discrepancy-guided reconstruction learning for image forgery detection," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.

[30] J. Wang, Z. Wu, J. Chen, and Y. Jiang, "M2tr: Multi-modal multi-scale transformers for deepfake detection," in *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*, 2022, pp. 615–623.

[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 012–10 022.