ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10889125

MIFAE-Forensics: Masked Image-Frequency AutoEncoder for DeepFake Detection

Hanyi Wang, Zihan Liu, Shilin Wang
School of Electronic Information and Electrical Engineering, *Shanghai Jiao Tong University*

***Abstract*—With continuously evolving generative models and increasingly diverse face forgery products, there is a growing demand for DeepFake detectors with stronger generalization ability and robustness. Previous works mainly capture method-specific forgery artifacts in the training set, thus failing to generalize well to unseen manipulations. In this paper, our key insight is that exploring common characteristics of natural faces is more ideal to alleviate overfitting rather than relying on specific forgery clues, as all sorts of manipulated images have intrinsic distributional differences from those captured by cameras. Hence, we propose a two-stage method, termed MIFAE-Forensics. Specifically, it reconstructs both facial semantics and local details from masked facial regions and high-frequency components, respectively, aiming to capture natural facial consistency in spatial domain and high-frequency details in frequency domain simultaneously. This facilitates the learning of a robust and transferable facial representation specialized for DeepFake detection. Subsequently, the pre-trained model is further fine-tuned to perform binary forgery classification along with reconstructing real faces in spatial domain, which ensures that the detector can maintain the ability to model real faces and encourages it to make decisions based on reconstruction discrepancies. Extensive experiments show superior results over state-of-the-arts on a wide range of DeepFake detection benchmarks. Our code is available at https://github.com/Mark-Dou/Forensics.**

***Index Terms*—Masked Image Modeling, DeepFake Detection.**

I. Introduction

Face manipulation technologies are advancing rapidly with deep generative models, creating increasingly realistic DeepFake content. These technologies pose significant security risks by enabling the creation of fake news, defaming celebrities, and compromising identity authentication systems. Consequently, there is an urgent need for effective and robust DeepFake detection methods in real-world applications.

To address these threats, various detection strategies have been developed. Initial efforts utilized neural networks to identify discriminative features for binary real-vs-fake classification (e.g., XceptionNet [1], capsule networks [2], recurrent neural networks [3], and attention mechanisms [4]). Besides, some approaches [5]–[7] further leveraged frequency cues as supplement information. However, many detectors still struggle with unseen manipulations in cross-dataset evaluations, highlighting the challenge of generalizing to novel forgeries.

One popular line of research focuses on identifying specific forgery traces left by common manipulation techniques to enhance generalization, such as blending boundaries [8]–[11], local feature inconsistencies [12], [13], and anomalous frequency spectrum caused by up-sampling [14], [15]. Despite the encouraging performance improvement under the cross-dataset scenario, their effectiveness is still limited as new breeds of generative models emerge.

In this paper, our key insight is that exploring common characteristics of natural faces is more ideal to alleviate overfitting than relying on specific forgery cues, as all sorts of manipulated images have intrinsic distributional differences from those captured by cameras. By learning a highly concentrated feature distribution of natural faces, we can detect DeepFakes, which deviate from real faces in their distribution, and generalize better to unseen manipulations. Upon such considerations and combining the characteristics of DeepFake products, we hypothesize that a robust facial representation for DeepFake detection should explore both facial region consistency and fine-grained details, particularly those in high-frequency components. DeepFake images, however, disrupt feature consistency and lack high-frequency details, resulting in discrepancies when processed by a genuine facial encoder. These discrepancies can then be used for classification.

To this end, we propse a novel two-stage method termed **M**asked **I**mage-**F**requency **A**uto**E**ncoder Forensics (MIFAE-Forensics) for DeepFake detection. In the first stage, we utilize a self-supervised mask-and-predict paradigm to derive natural facial representations from unlabeled datasets. This involves a facial region guided masking task in the spatial domain and a strategy for masking high-frequency components in the frequency domain, aiming to simultaneously reconstruct spatial consistency and high-frequency details. In the second stage, we fine-tune the encoder by unfreezing only the last Transformer block and incorporate a linear layer for real-vs-fake classification. Additionally, the pre-trained spatial decoder is used to reconstruct real faces only, helping the model maintain its ability to represent genuine faces and base decisions on reconstruction discrepancies.

The main contributions of our work are: 1) MIFAE-Forensics, a two-stage method for learning robust facial representations from genuine faces, which effectively transfers to DeepFake detection tasks. 2) The introduction of dual-domain Masked Image Modeling (MIM) in a self-supervised manner for DeepFake detection task, focusing on both spatial and frequency domains. 3) Our method demonstrates superior generalization across multiple benchmarks, including FaceForensics++ [1], Celeb-DF-V2 [16], DFD [17], DFDC [18], FaceShifter [19], DeeperForensics-1.0 [20], and WildDeepfake [21], proving its effectiveness against unseen forgeries.

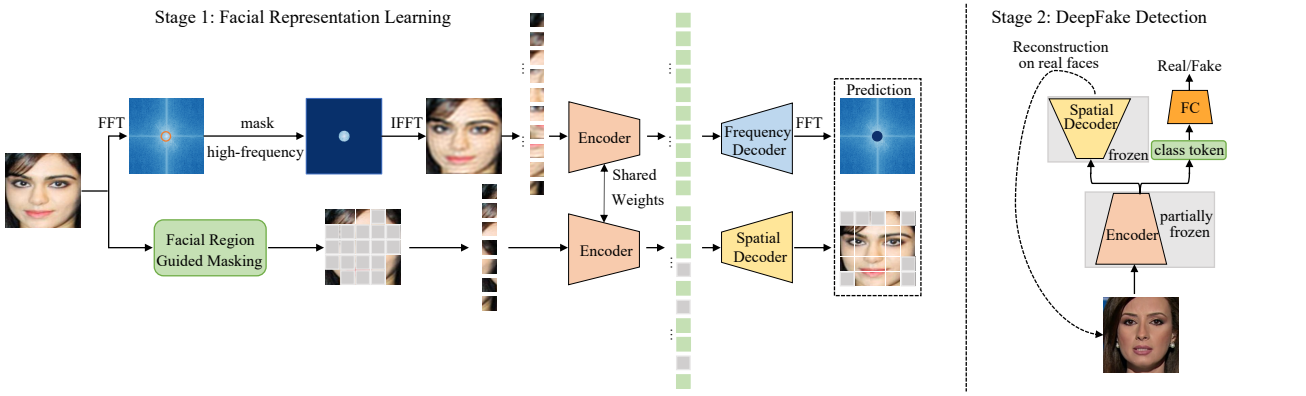

Fig. 1. **MIFAE-Forensics.** In stage 1, a self-supervised pre-training task to reconstruct both facial semantics and local details from masked facial regions and high-frequency components, respectively. In stage 2, the pre-trained encoder is partially fine-tuned with an extra added linear layer to perform the classification task while reconstructing real faces with the pre-trained spatial decoder.

II. Method

To develop a generalize DeepFake detector, a two-stage method is introduced, detailed in Fig.1. The first stage employs Masked Image Modeling (MIM) for learning spatial consistency and Masked Frequency Modeling (MFM) to capture high-frequency details. It also includes a facial region-guided masking strategy [22] that emphasizes facial features over backgrounds. In the second stage, the model undertakes real-vs-fake classification and reconstructs only real faces to maintain pre-trained knowledge and prevent overfitting.

A. Facial Representation Learning

1) MIM for Spatial Consistency Learning:

Facial Region Guided Masking. In contrast to the random patch masking of the vanilla MAE [23], our approach uses a facial region guided masking strategy [22] that focuses on spatial consistencies among crucial facial features. By selectively masking key areas such as the eyes, mouth, and nose, our method prioritizes discrepancies in important facial parts, thereby improving decision-making in detection tasks. This targeted masking challenges the model to reconstruct and learn abundant facial details.

Network Architecture. We utilize a ViT [24] encoder to analyze spatial consistency within prominent facial features, processing only visible patches to capture genuine face correlations. The decoder comprises several Transformer blocks that work on both the encoded visible patches and masked tokens, emphasizing intra-facial consistency.

Reconstruction. The model learns spatial consistencies among facial regions by predicting masked facial patches. We use mean squared error $\mathcal{L}_{spa}$ to assess the pixel-level accuracy of the reconstruction, calculated solely on the masked patches:

$$\mathcal{L}_{spa} = \frac{1}{K}\sum_{k=1}^{K}(\hat{p_k} - p_k)^2, \tag{1}$$

where K denotes the number of masked patches, $\hat{p_k}$, p_k represents the predicted and reference image patches, respectively.

2) MFM for High-frequency Details Learning:

Preliminary. For a single channel image $x \in \mathbb{R}^{H\times W}$, its frequency representation via the 2D Discrete Fourier Transform (DFT) is denoted as $\mathcal{F}(x)$. The 2D Inverse DFT (IDFT), denoted by $\mathcal{F}(x)^{-1}$. For RGB images, each channel is transformed independently.

High-frequency Masking. Inspired by the masking strategy from MFM [25], we propose to simply mask the high-frequency components with low-pass filters to encourage the model to learn the fine-grained details of genuine faces. We first define a binary mask $M \in (0,1)^{H\times W}$, which separates the low and high frequency components of $\mathcal{F}(x)$ according to a hyper-parameter, i.e., radius r:

$$M(u,v) = \begin{cases} 1, & d((u,v),(c_h,c_w)) < r \\ 0, & else \end{cases}, \tag{2}$$

where (u,v) denotes the coordinate on the frequency spectrum and (c_h, c_w) denotes the center of frequency spectrum $\mathcal{F}(x)$, $d(\cdot,\cdot)$ denotes a distance criterion. Here we use Euclidean distance and a circle masking shape following [25]. With mask M, we then obtain the decomposed low-frequency component x_l and high-frequency component x_h, respectively.

$$x_l = \mathcal{F}^{-1}(\mathcal{F}(x) \odot M), \tag{3}$$

$$x_h = \mathcal{F}^{-1}(\mathcal{F}(x) \odot (I - M)), \tag{4}$$

where I denotes the identity matrix.

Then the low-frequency image components are fed into an encoder, followed by a frequency decoder to predict the corresponding high-frequency counterparts. This cross-frequency prediction can serve as a challenging task for the model to learn fine-grained high-frequency details more effectively. Note that after pre-training, the encoder is applied to uncorrupted images for subsequent detection, so we take the converted spatial images as input rather than frequency spectrum to avoid such domain gap. Moreover, the encoder for embedding low-frequency preserved image and unmasked patches can share weights, which reduces the computational complexity and promote the model to mine complementary information from both modalities.

Network Architecture. The encoder processes low-frequency components and shares weights with the unmasked patch encoder to minimize computational load and enhance information extraction from both domains. A linear frequency decoder is adopted to predict the masked high-frequency components.
Reconstruction. The reconstruction of high-frequency components involves predicting missing frequency values from the low-pass filtered spectrum. To better capture the important frequency information in high-frequency components, We use focal frequency loss [26] to optimize this reconstruction:

$$\mathcal{L}_{freq} = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} \gamma(\hat{\mathcal{F}}(x)(u,v), \mathcal{F}(x)(u,v)), \quad (5)$$

where $\hat{\mathcal{F}}(x)$, $\mathcal{F}(x)$ denotes the reconstructed frequency spectrum and ground truth, respectively.

$$\gamma(\hat{\mathcal{F}}, \mathcal{F}) = \sqrt{(\hat{\mathcal{R}} - \mathcal{R})^2 + (\hat{\mathcal{I}} - \mathcal{I})^2}, \quad (6)$$

where $\mathcal{R}$ and $\mathcal{I}$ represents the real and imaginary part of the frequency spectrum $\mathcal{F}$. Loss is computed solely on the masked high-frequency components.

3) Pre-training Objective.:

Overall loss. During pre-training, our MIFAE-Forensics learns a robust facial representation by reconstructing facial semantics and local details from both spatial and frequency domains:

$$\mathcal{L} = \mathcal{L}_{spa} + \lambda \mathcal{L}_{freq}, \quad (7)$$

where $\mathcal{L}$ is the total loss and λ is the weight adjusting the impact of frequency domain reconstruction.

B. DeepFake Detection

To apply pre-trained knowledge to DeepFake detection, we initially freeze the well-trained encoder, which has learned robust facial representations emphasizing spatial consistency and high-frequency details. We then add a linear layer upon the encoder for real-vs-fake classification using the class token of ViT. However, this absence of trainable non-linear features limits the model's ability to distinguish between genuine and manipulated faces. To address this, we partially fine-tune the encoder by unfreezing only the last Transformer block, preserving the integrity of the learned genuine face distributions. This approach, supported by findings in [23], has proven effective in our experiments.

Additionally, we employ the pre-trained spatial decoder for reconstruction tasks. This ensures the fine-tuned representations can detect forgeries without forgetting initial training, maintaining the capability to model real faces and aiming to establish a more robust decision boundary to minimize overfitting. In pre-training, representations focus on the consistencies and detailed features of facial parts, whereas DeepFakes typically disrupt these consistencies and lack high-frequency details. By leveraging these discrepancies in reconstructions, our model can more accurately identify manipulated content.

For classification, we use cross-entropy loss $\mathcal{L}cls$:

$$\mathcal{L}_{cls} = \frac{1}{N} \sum_{k=1}^{N} -[y_k log(\hat{y_k}) + (1 - y_k) log(1 - \hat{y_k})], \quad (8)$$

where N is the number of samples, and $\hat{y_k}$, y_k represent the predicted and actual labels, respectively.

We also evaluate the reconstruction quality in the spatial domain using mean squared error $\mathcal{L}_{recon}$:

$$\mathcal{L}_{recon} = \frac{1}{N} \sum_{k=1}^{N} (\hat{p_k} - p_k)^2, \quad (9)$$

where N is the number of pixels, and $\hat{p_k}$, p_k are the predicted and actual pixels, respectively.

Thus, the overall loss L for fine-tuning in DeepFake detection is:

$$\mathcal{L} = \mathcal{L}_{cls} + \beta \mathcal{L}_{recon}, \quad (10)$$

where β adjusts the importance of reconstruction.

III. Experiments

A. Experimental Settings.

Datasets and Metrics. We pre-train using the VGGFace2 dataset [33], which contains over 3.3 million images across 9131 identities, without any labels. For fine-tuning, we use the FaceForensics++ (FF++) [1] dataset. To evaluate generalization, we test on multiple benchmarks including Celeb-DF-V2 [16], DFD [17], FaceShifter [19], DeeperForensics-1.0 [20], DFDC [18], and WildDeepfake [21]. We assess DeepFake detection using Area Under the Receiver Operating Characteristic Curve (AUC) and Equal Error Rate (EER).
Implementation Details. Input images are preprocessed to 224 x 224 using RetinaFace [34]. Our model employs a 75% masking ratio for spatial learning and a radius of 16 for learning high-frequency details. The loss weight λ is set at 0.1. We pre-train from scratch over 400 epochs. For fine-tuning, we unfreeze the last Transformer block and add a linear layer, training for 10 epochs with a 5-epoch warm-up. The reconstruction loss weight β is set at 0.5.

B. Comparison with State-of-the-arts

Generalization to unseen datasets. Table I demonstrates that our method significantly outperforms baseline models, validating its effectiveness in identifying unseen forgeries through the exploration of genuine face feature distributions. Unlike RECCE [30], which may produce blurry results by reconstructing entire images without facial semantic guidance, our approach focuses on facial consistencies and fine-grained details. We employ a facial region-guided masking component and reconstruct high-frequency components, enabling precise detection based on discrepancies in facial reconstructions. Our method, treating DeepFake detection as out-of-distribution (OOD) detection, effectively learns from common real face characteristics and implements challenging self-supervisory tasks to capture facial semantic consistency and detailed frequency patterns. This strategy shows potential in generalizing to newer generative models, offering a promising perspective on addressing generalization challenges in future research.

TABLE I
CROSS-DATASET EVALUATION IN TERMS OF AUC, EER. UNDERLINES DENOTES THE SECOND-BEST RESULTS.

Method	Celeb-DF-V2		DFD		DFDC		FSh		DF-1.0		WildDeepfake	
	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
Xception [1]	65.27	38.77	87.86	21.04	69.90	35.41	72.00	-	84.50	-	66.23	40.16
EN-b4 [27]	68.52	35.61	87.37	21.99	70.12	34.54	-	-	85.31	22.80	64.27	37.60
Face X-ray [9]	74.76	31.16	93.47	12.72	71.15	32.62	92.80	-	86.80	-	-	-
F^3-Net [5]	71.21	34.03	86.10	26.17	72.88	33.38	79.14	-	82.27	24.68	67.71	40.17
SRM [6]	75.31	32.48	85.51	25.64	71.58	34.77	80.26	-	82.54	-	66.51	41.52
MAT (EN-b4) [4]	76.65	32.83	87.58	21.73	67.34	38.31	-	-	89.34	17.38	70.15	36.53
LTW [28]	77.14	29.34	88.56	20.57	74.58	33.81	-	-	-	-	67.12	39.22
Local-relation [13]	78.26	29.67	89.24	20.32	76.53	32.41	-	-	-	-	68.76	37.50
DCL [29]	82.30	26.53	91.66	16.63	76.71	31.97	-	-	-	-	71.14	36.17
RECCE [30]	82.08	27.01	92.09	16.98	75.66	33.10	-	-	95.45	11.01	70.71	37.78
SFDG [31]	83.30	24.20	93.80	13.70	76.64	32.07	-	-	97.10	4.10	72.27	34.60
IID [32]	83.39	23.85	95.87	10.01	77.03	31.89	-	-	-	-	-	-
Ours	**83.39**	**23.51**	**95.90**	**9.93**	**77.19**	**31.81**	**97.84**	**6.55**	**97.71**	**3.35**	**72.39**	**35.80**

TABLE II
THE EFFECT OF DIFFERENT RECONSTRUCTION DOMAIN DURING PRE-TRAINING IN TERMS OF CROSS-DATASET EVALUATION.

Method	Recon.		Celeb-DF-V2		DFDC		WildDeepfake	
	Spa.	Freq.	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
w/o MFM	✓	✗	82.11	25.76	73.77	32.88	70.64	37.51
w/o MIM	✗	✓	80.38	27.81	74.61	35.55	69.77	36.18
MIFAE-Forensics	✓	✓	**83.39**	**23.51**	**77.19**	**31.81**	**72.39**	**35.80**

TABLE III
THE EFFECT OF DIFFERENT RECONSTRUCTION CONSTRAINTS DURING FINE-TUNING IN TERMS OF CROSS-DATASET EVALUATION.

Method	$\mathcal{L}_r$		Celeb-DF-V2		DFDC		WildDeepfake	
	real	fake	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
w/o reconstruction	✗	✗	81.26	26.88	75.01	32.32	71.37	36.36
full reconstruction	✓	✓	82.12	26.37	75.72	31.94	71.63	36.16
MIFAE-Forensics	✓	✗	**83.39**	**23.51**	**77.19**	**31.81**	**72.39**	**35.80**

C. Ablation Study

Effect of reconstruction domains during pre-training. We evaluated our representation learning strategies through cross-dataset experiments as shown in Table II. In MIAE-Forensics, which reverts to the baseline MAE method focusing solely on pixel reconstruction, the model primarily captures global semantic information, neglecting fine-grained high-frequency details. In contrast, MFAE-Forensics highlights high-frequency details, showing improved performance and underscoring the importance of high-frequency reconstruction for effective DeepFake detection. Our hybrid approach combines spatial consistency with high-frequency details, leading to promising results across multiple datasets.

Effect of reconstruction during fine-tuning. To assess the impact of applying reconstruction constraints only to real faces during fine-tuning, we conducted experiments detailed in Table III. We compared three scenarios: fine-tuning without any reconstruction loss, applying reconstruction loss to both real and fake faces, and our MIFAE-Forensics method, which applies reconstruction loss only to real faces. This strategy enhances the model's ability to detect discrepancies, as the pre-trained encoder is tailored to represent genuine faces only. Consequently, focusing reconstruction loss on real faces amplifies the differences when encountering manipulated images.

D. Visualization of anomalous regions.

Visualization of Anomalous Regions. To demonstrate how our model identifies forgeries by analyzing genuine face features, we visualized the reconstructed faces and pixel-level differences from the original images, as shown in Fig. 2. Real faces are reconstructed with minimal blur, while forged areas in fake faces show noticeable distortions due to spatial inconsistencies. This occurs because our pre-trained encoder, which models spatial consistency and high-frequency details in real faces, fails to accurately reconstruct fake faces that lack these attributes. Additionally, we used Grad-CAM [35] to highlight spatial anomalies used by the model for decision-making, aligning with the reconstruction discrepancies. This visualization confirms that the model prioritizes facial features over background, validating the effectiveness of our facial region-guided masking strategy.

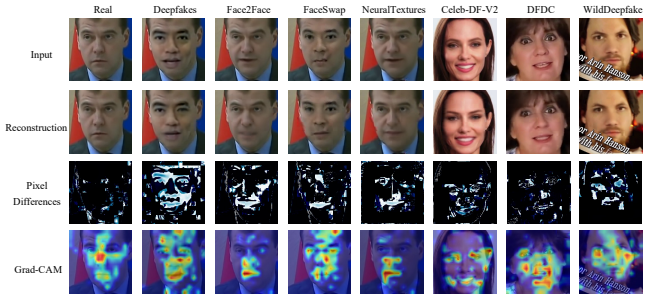

Fig. 2. Visualization results of MIFAE-Forensics on FF++, Celeb-DF-V2, DFDC and WildDeepfake, respectively. The first row displays input images, the second row shows the reconstruction results, the third row shows the pixel-level differences between input and reconstructed faces. The fourth row demonstrates the Gard-CAM [35] visualization.

IV. CONCLUSION

Most existing DeepFake detection methods focus on identifying method-specific artifacts, which limit their generalization capabilities due to an over-reliance on specific forgery patterns. Our approach, MIFAE-Forensics, emphasizes the importance of recognizing common characteristics of natural faces to mitigate overfitting. This method reconstructs facial semantics and local details from masked regions and high-frequency components, promoting the learning of spatial consistencies and subtle details within these regions. By employing flexible mask-and-predict strategies, MIFAE-Forensics develops a robust and adaptable facial representation that enhances DeepFake detection across various benchmarks, as demonstrated by our extensive experiments.

V. ACKNOWLEDGEMENTS

The work described in this paper was supported by the National Science Foundation of China (62271307).

REFERENCES

[1] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[2] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2307–2311.

[3] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, no. 1, pp. 80–87, 2019.

[4] 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*, 2021, pp. 2185–2194.

[5] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*. Springer, 2020, pp. 86–103.

[6] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 317–16 326.

[7] 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*, 2021, pp. 6458–6467.

[8] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656*, 2018.

[9] 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*, 2020, pp. 5001–5010.

[10] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 710–18 719.

[11] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–18 729.

[12] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 023–15 033.

[13] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, "Local relation learning for face forgery detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1081–1088.

[14] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7890–7899.

[15] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 772–781.

[16] 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*, 2020, pp. 3207–3216.

[17] J. Nick Dufour, Andrew Gully, "Deepfake detection challenge." July 12 2023. [Online]. Available: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

[18] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[19] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5074–5083.

[20] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2889–2898.

[21] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2382–2390.

[22] 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," *arXiv preprint arXiv:2211.06627*, 2022.

[23] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[25] J. Xie, W. Li, X. Zhan, Z. Liu, Y. S. Ong, and C. C. Loy, "Masked frequency modeling for self-supervised visual pre-training," *arXiv preprint arXiv:2206.07706*, 2022.

[26] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 919–13 929.

[27] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[28] K. Sun, H. Liu, Q. Ye, Y. Gao, J. Liu, L. Shao, and R. Ji, "Domain general face forgery detection by learning to weight," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 3, 2021, pp. 2638–2646.

[29] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji, "Dual contrastive learning for general face forgery detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2316–2324.

[30] 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*, 2022, pp. 4113–4122.

[31] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7278–7287.

[32] B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang, and D. Ye, "Implicit identity driven deepfake face swapping detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4490–4499.

[33] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[34] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212.

[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.