Where to Focus: Central Attention-based Face Forgery Detection

Jinghui Sun[1,2], Yuhe Ding[1], Jie Cao[3,4], Junxian Duan[3,4], and Aihua Zheng[2,5,6(✉)]

[1] School of Computer Science and Technology, Anhui University, Hefei, 230601, China
[2] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, China
[3] Center for Research on Intelligent Perception and Computing (CRIPAC)
[4] Institute of Automation, Chinese Academy of Sciences
[5] Information Materials and Intelligent Sensing Laboratory of Anhui Province
[6] School of Artificial Intelligence, Anhui University, Hefei, 230601, China
jingh_sun@126.com, madao3c@foxmail.com, jie.cao@cripac.ia.ac.cn, junxian.duan@ia.ac.cn, ahzheng214@foxmail.com

Abstract. Face forgery detection in compressed images is an active area of research. However, previous frequency-based methods are subject to two limitations. One aspect to consider is that they apply the same weight to different frequency bands. Moreover, they exhibit an equal treatment of regions that contain distinct semantic information. To address these limitations above, we propose the Central Attention Network (CAN), a multi-modal architecture comprising two bright components: Adaptive Frequency Embedding (AFE) and Central Attention (CA) block. The AFE module adaptively embeds practical frequency information to enhance forged traces and minimize the impact of redundant interference. Moreover, the CA block can achieve fine-grained trace observation by concentrating on facial regions where indications of forgery frequently manifest. CAN is efficient in extracting forgery traces and robust to noise. It effectively reduces the unnecessary focus of our model on irrelevant factors. Extensive experiments on multiple datasets validate the advantages of CAN over existing state-of-the-art methods.

Keywords: Face Forgery Detection · Multi-level Frequency Fusion · Attention Mechanism.

1 Introduction

Deep learning advancements and the widespread availability of online resources make tools like deepfakes [1] and face2face [2] easily accessible, allowing individuals without professional training to easily manipulate facial expressions, attributes, and identities within images. However, criminals misuse these technologies, resulting in a proliferation of high-quality fake photographs on social media, making it difficult to distinguish between genuine and modified faces.

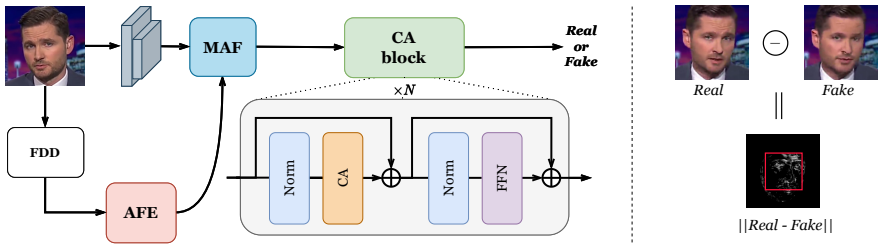

Fig. 1: (a) Overview of our proposed CAN. Combining *FDD* with *AFE* allows for the extraction of fine-grained frequency information as well as the highlighting of the components most useful for forgery detection. *MAF* for collaborative feature interaction. The *CA* block enables the network to focus more on key central areas. (b) Illustration of the differences between *Real* and *Fake*. The forgery traces are clustered in the central region (in the red box), indicating that the center is more important than the other areas.

The above issues prompt the development of face forgery detection based on deep neural networks [3,4,5,6,7,8,9,10,11]. However, they perform poorly in compressed images. Recent works [12,13,14,15] highlight the effectiveness of capturing forgery traces in the frequency domain under high compression. While decent detection results are achieved by combining RGB and frequency information, their method of information processing is coarse-grained, which causes two limitations.

For one thing, previous studies usually obtain frequency domain information through Discrete Cosine Transform and then use hand-crafted filters to extract it into high, middle, and low frequency bands. According to [15], the low and middle frequency preserve rich semantic information, such as human faces and backgrounds, which is highly consistent with RGB input. Meanwhile, the high frequency reveals small-scale details, often related to forging sensitive edges and textures. These show that the role and importance of these three frequency bands are completely different. Previous works show excellent performance by combining frequency information. They apply the same weight for different frequency bands, which may not be optimal for using frequency information. Because this limitation may lead to magnifying irrelevant noise and ignoring the more valuable components.

For another thing, the equal treatment of regions with different semantic information prevails in existing methods. However, as shown in Fig. 1(b), most of the differences between real image and fake image are obviously clustered in the central region (in the red box). This means that the central region can provide rich traces of forgery compared to other regions (outside the red box). Treating the regions equally not only results in superfluous noise but also neglects significant evidence.

To address these limitations, we propose a new approach to detect face forgery, termed as **C**entral **A**ttention **N**etwork (CAN), as shown in Fig. 1(a). The CAN consists of four main modules: Frequency Domain Decomposition (FDD), Adaptive Frequency Embedding (AFE), Multi-modal Attention Fusion (MAF), and Central Attention (CA) block. CAN initially uses FDD to extract low, middle, and high frequency information from input images. Then our AFE module concatenates the three frequency bands for richer frequency perception cues. In terms of information extraction granularity and channel allocation, it prioritizes high frequency information. Subsequently, the frequency is fused into the RGB branch by the MAF module. Finally, we add the CA block, which is similar to the Transformer block [16], to prevent the network from focusing over on irrelevant areas. The module uses different scale attention mechanisms for the central and global regions, enabling the network to prioritize the central region more efficiently.

Extensive experiments have demonstrated that our proposed Central Attention Network is effective in capturing forgery traces and significantly improves upon the shortcomings of existing detection methods. Our work makes the following primary contributions:

- We propose the AFE module aiming at mining the more valuable fine-grained frequency components to uncover subtle nuances and hidden artifacts.
- We propose the Central Attention mechanism that provides a refined perspective of forged regions and reduces the attention to irrelevant areas.
- Numerous experiments demonstrate that our proposed Central Attention block is highly versatile and can be seamlessly integrated into various existing networks, resulting in a significant enhancement of their detection capabilities.

2 Related Work

Face forgery detection. With the rise of deep learning, the adverse effects of image forgery techniques on political credibility, social stability, and personal reputation have increasingly received attention from society.

Therefore, various image forgery detection technologies have developed rapidly in recent years. Previous works [7,8,9,10,11] use deep CNN models to predict whether a face region is real or fake. Unfortunately, they are only partially effective in high compression scenarios.

Inspired by [13], recent studies try to improve detection performance in high compression scenes by incorporating frequency domain information into existing detection techniques. Qian et al. [15] proposes a dual-stream network named F^3-Net, where one branch utilizes three filters to perform frequency decomposition on RGB information. Chen et al. [17] uses the Spatial Rich Model to extract residual noise to guide the RGB features. Li et al. [18] and Gu et al. [14] further decompose fine-grained frequency domain information from the perspective of image compression. While previous methods demonstrate significant effects, they either underutilize frequency information or treat all levels of frequency

equally. In contrast, our method involves a decomposition of frequency domain information and adaptive embedding to fully leverage the available frequency.

Vision Transformers. Transformers are known for their powerful remote contextual information modeling capabilities and high performance in natural language processing tasks. While various backbones are proposed to handle computer vision tasks, conventional transformers treat each patch at a single scale. Recent works [19,20,21] introduce multiple scales to focus on objects of different sizes, [22] proposes a multi-modal framework that integrates multi-scale transformer. But these approaches are generic and not tailored to the specific characteristics of forgery image detection. In this paper, we propose a Central Attention block that addresses the fact that fake regions tend to be concentrated in the central area of an image while other areas contain interference information.

3 Proposed Method

3.1 FDD: Frequency Domain Decomposition

For the input $rgb \in \mathbb{R}^{3\times H\times W}$, where H and W are the height and width of the image. First, we apply $\mathcal{DCT}$ as **D**iscrete **C**osine **T**ransform to transform the RGB domain to the frequency domain. Based on [15], we devise $N = 3$ filters that are capable of effectively decomposing the frequency into three distinct frequency bands: high, middle, and low:

$$dct^n = \mathcal{DCT}(rgb) \odot f^n, \qquad n = 1, ..., N. \tag{1}$$

We utilize $\mathcal{ID}$ as **I**nverse **D**iscrete Cosine Transform to transform the frequency domain into RGB domain to obtain the $\tilde{freq} \in \mathbb{R}^{3N\times H\times W}$ which is concatenated by $freq^n$ along the channel dimension. This manipulation helps to preserve the shift invariance and local consistency of natural images.

$$freq^n = \mathcal{ID}(dct^n), \qquad n = 1, ..., N. \tag{2}$$

To achieve a more refined analysis of the frequency information, we apply $\mathcal{M}$ as the median filter to extract noise information from the input features $\tilde{freq}$:

$$\tilde{freq}_{noise} = \tilde{freq} - \mathcal{M}(\tilde{freq}). \tag{3}$$

To magnify subtle forgery clues, we utilize the following formula:

$$freq = \tilde{freq} + Conv_{1\times1}(Sigmoid(\tilde{freq}_{noise})). \tag{4}$$

Specifically, a 1×1 convolution layer followed by a $Sigmoid$ activation function is used to generate a noise mask, which is then added back to the original feature maps to enhance the frequency input.

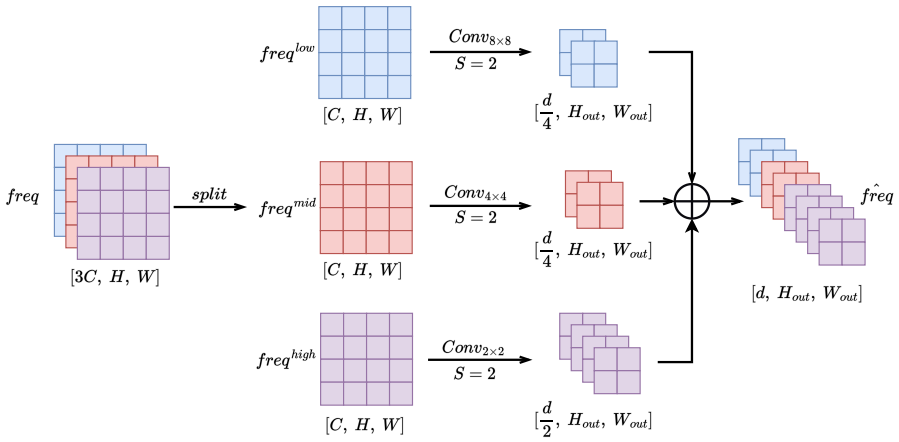

Fig. 2: The illustration of the proposed AFE allocates weight based on the value of frequency levels.

3.2 AFE: Adaptive Frequency Embedding

Previous works show excellent performance by combining frequency information. Applying the same weight to different frequency bands might be the general method in their works. It may not be optimal for using frequency domain information because it may magnify irrelevant noise or misuse the valuable components. To address this point, we propose the AFE module that fully exploits the role of different frequency components, as shown in Fig. 2. The AFE module extracts information from different frequency bands via different convolution kernels. Tampering artifacts mainly reside in the high-frequency spectrum; therefore, we use a 2×2 convolution kernel to extract fine-grained texture information from it. For middle and low frequency that still contain basic information which provides a solid foundation for fusing Frequency and RGB, we adopt 4×4 and 8×8 convolution kernels to extract semantic features respectively. The channel outputs generated by these convolutions are also treated differently based on their importance in different frequency bands. Specifically, $\frac{d}{2}$ channels are allocated for high frequency channels while middle and low frequency each occupy $\frac{d}{4}$ channels. The d represents the number of output feature channels. Ultimately, the three branches are concatenated along the channel to obtain the $\hat{freq}$.

3.3 MAF: Multi-modal Attention Fusion

The complementary relationship between RGB and Freq is acknowledged. The MAF module integrates them by means of an attention mechanism. The RGB feature map is denoted as $\hat{rgb} \in \mathbb{R}^{d \times h \times w}$, while the frequency feature map is represented as $\hat{freq} \in \mathbb{R}^{d \times h \times w}$. We obtain the query vector Q from $\hat{rgb}$ using a 1×1 convolution layer. Similarly, we obtain the key vector K and value vector V from $\hat{freq}$ using 1×1 convolution layers. Then, we flatten them along the

spatial dimension to get 2D embeddings Q_e, K_e, V_e. Using the self-attention mechanism, we generate an attention map that represents relevance between the input features $\hat{rgb}$ and $\hat{freq}$:

$$\hat{W} = softmax(\frac{Q_e K_e}{\sqrt{\boldsymbol{D}}})V_e, \tag{5}$$

where $\boldsymbol{D}$ is the dimensionality of the key vectors. After obtaining attention weights, we compute weighted values via a 3×3 convolution. Additionally, we adopt residual connections to add them to the original input, which alleviates the potential gradient vanishing issue during the training process.

$$f = \hat{rgb} + Conv_{3\times3}(\hat{W}). \tag{6}$$

3.4 CA block: Central Attention block

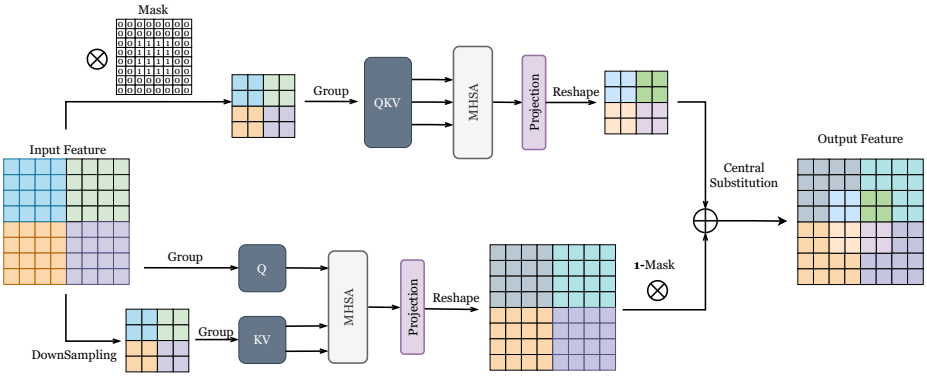

Fig. 3: The proposed Central Attention mechanism when α is 0.5.

The conventional transformer models treat all patches of an image equally, without taking into account the relative significance of distinct areas. Recent studies [20,22] show that incorporating multi-scale information can improve detection accuracy. Yet these models are not optimized for detecting forged face images. Our observation is that forged regions tend to cluster around the centre of input images. Based on this insight, we propose the Central Attention which aids the network in concentrating on key regions.

For the input global feature $f^g \in \mathbb{R}^{c\times h\times w}$, we commence by initializing a $Mask$ of size $h \times w$. Subsequently, we selectively filled the central region, characterized by dimensions of $\alpha h \times \alpha w$, with the value 1. The surrounding area is then filled with the value 0 to complete the mask initialization process. α is the proportion that determines the size of the central region. We then apply this $Mask$ to the input f^g, resulting in a central feature map $f^c = f^g \odot mask$. Fig. 3 illustrates the framework of the Central Attention mechanism, with a value of 0.5 for parameter α.

For the global feature f^g, we downsample it into $\frac{h}{2} \times \frac{w}{2}$ by convolution to obtain f^d. We obtain the embedding Q_g from f^g, the embeddings K_g and V_g from f^d. Inspired by [21], we define the operation of dividing the input into $G \times G$ patches through sliding windows as $SW^G(\cdot)$.

$$Q_g = SW^g(Q_g), \quad K_g,\ V_g = SW^{\frac{g}{2}}(K_g,\ V_g), \tag{7}$$

$$f^g = MHSA(Q_g,\ K_g,\ V_g). \tag{8}$$

Similarly, for the central feature f^c, we embed f^c into Q_c, K_c, V_c.

$$Q_c,\ K_c,\ V_c = SW^c(Q_c,\ K_c,\ V_c), \tag{9}$$

$$f^c = MHSA(Q_c,\ K_c,\ V_c), \tag{10}$$

where $MHSA$ represents Multi-Head Self-Attention.

This allows the network to focus more on the central region while still considering the surrounding areas. In order to maintain spatial coherence, the grouping features are rearranged and subsequently substituted with f^c to replace the corresponding position features. $[\cdot]$ denotes the above operations.

$$f = [f^g, f^c]. \tag{11}$$

The CA block can be described mathematically:

$$f = f^g + CA(Norm(f^g)), \tag{12}$$

$$f = f + FFN(Norm(f)), \tag{13}$$

where $Norm$ and FFN mean BatchNorm, Feed Forward Network separately.

3.5 Overall Loss

After passing through several CA blocks, the feature is sent into the remaining backbone network to extract richer features f. Then a fully connected layer and a *sigmoid* function are used to obtain the final prediction probability y. So the Binary cross-entropy loss is defined as:

$$\mathcal{L}_{Bce}(y) = y \log \hat{y} + (1 - y) \log(1 - \hat{y}), \tag{14}$$

where y is set to 1 if the face image has been manipulated, otherwise it is set to 0. To ensure feature consistency, we use the Consistency loss function $\mathcal{L}_{Cos}$ in [23] to constrain the feature distribution. f_1 and f_2 are the final features obtained from the same input image after through distinct data augmentation and being passed through the network. Mathematically:

$$\mathcal{L}_{Cos}\left(f_1, f_2\right) = \left(1 - \tilde{f}_1 \cdot \tilde{f}_2\right)^2, \tag{15}$$

where $\tilde{f} = \frac{f}{\|f\|_2}$ denotes the normalized vector of the representation vector f.

So we combine the Binary cross-entropy loss and the Consistency loss function linearly with $\beta = 2$.

$$\mathcal{L}_{all} = \mathcal{L}_{Bce}(y_1) + \mathcal{L}_{Bce}(y_2) + \beta \mathcal{L}_{Cos}\left(f_1, f_2\right). \tag{16}$$

Table 1: Quantitative results on Celeb-DF dataset and FF++ dataset.

Methods	FF++(HQ)		FF++(LQ)		Celeb-DF	
	Acc	**AUC**	**Acc**	**AUC**	**Acc**	**AUC**
MesoNet [6]	83.10	-	70.47	-	-	-
Xception [24]	95.73	96.30	86.86	89.30	97.90	99.73
Face X-ray [7]	-	87.40	-	61.60	-	-
Two-branch [25]	96.43	98.70	86.34	86.59	-	-
RFM [11]	95.69	98.79	87.06	89.83	97.96	99.94
Add-Net [9]	96.78	97.74	87.5	91.01	96.93	99.55
F3-Net [15]	97.52	98.10	90.43	93.30	95.95	98.93
FDFL [18]	96.69	99.30	89.00	92.40	-	-
Multi-Att [8]	97.60	99.29	88.69	90.40	97.92	99.94
SIA [26]	97.64	99.35	90.23	93.45	-	-
PEL [14]	97.63	99.32	**90.52**	94.28	-	-
Ours	**97.65**	**99.44**	90.40	**95.09**	**99.36**	**99.98**

4 Experiments

4.1 Experimental Setup

Datasets. We adopt two widely-used public datasets in our experiments, *i.e.*, FaceForensics++ [27], Celeb-DF [28].
1) **FaceForensics++** (FF++) [27] is a large forensics dataset containing 1000 original video sequences and 4000 manipulated video sequences produced by four automated face manipulation methods: *i.e.*, Deepfakes [1], Face2Face [2], FaceSwap [29], NeuralTextures [30]. Raw videos are compressed, resulting in two versions: high quality (HQ) and low quality (LQ). Following the official splits, we utilized 720 videos for training, 140 for validation, and 140 for testing.
2) **Celeb-DF** [28] dataset comprises 590 authentic videos sourced from YouTube, featuring individuals of varying ages, ethnicities, and genders. Additionally, the dataset includes 5639 corresponding DeepFake videos.

Implementation detail. The EfficientNet-B4 [31] pre-trained on ImageNet is adopted as the backbone of our network. We insert several CA blocks respectively after the second and third convolutional blocks with $\alpha = 0.5$. The input images are resized to 320×320. The whole network is trained with Adam optimizer with the learning rate of 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The batch size is 48 split on 4 $\times$ RTX 3090 GPUs.

Evaluation Metrics. Following the convention [27,15,10,14,22], we apply Accuracy score (Acc), Area Under the Receiver Operating Characteristic Curve (AUC) as our evaluation metrics.

Comparing Methods. We compare our methods with several advanced methods: MesoNet [6], Xception [24], Face X-ray [7], Two-branch [25], RFM [11], Add-Net [9], F^3-Net [15], FDFL [18], Multi-Att [8], SIA [26], PEL [14].

Table 2: The effect of each component. CAB represents CA blocks.

RGB	Freq	AFE	CAB	Acc	AUC
✓				88.70	92.87
	✓			88.49	92.63
✓	✓			88.89	92.89
✓	✓	✓		89.94	93.57
✓	✓		✓	90.36	94.15
✓	✓	✓	✓	**90.40**	**95.09**

Table 3: Ablation study of other backbones with our CA blocks.

Model		Acc	AUC
PF	+None	66.79	69.28
	+CAB	78.79	80.31
CNX	+None	76.45	77.92
	+CAB	80.43	80.64
PF*	+None	86.93	90.09
	+CAB	87.22	90.34
CNX*	+None	87.57	90.77
	+CAB	87.93	91.07

4.2 Comparison to the State-of-the-Arts

Following [27,15], we compare our method with various advanced techniques on the FF++ dataset with different quality settings (*i.e.*, HQ and LQ), and further evaluate the performance of our approach on the Celeb-DF dataset. In Tab. 1 the best, second, third results are shown in Red, Blue, Green. The performance of our proposed method, especially under high compression, is comparable or superior to existing methods, as evidenced by the Acc and AUC metrics. It is worth noting that the method PEL [14] is a two-stream network with twice as many parameters as ours. We achieve competitive results using only half the parameters. These gains mainly come from the CAN's ability to fully utilize frequency information and reduce interference from irrelevant information.

4.3 Ablation study and architecture analysis

Components. As shown in Tab. 2, we develop several variants and conduct a series of experiments on the FF++ (LQ) dataset to explore the impact of different components in our proposed method. Using only RGB or frequency as input in the single-stream setting leads to similar results. Combining both original streams can slightly improve performance, which demonstrates that frequency and RGB are unique and complementary. Adding an AEF module or CA blocks can significantly improve performance, achieving optimal results using the overall CAN framework. It shows that each module is effective: the AFE module fully mines frequency domain information and filters noise, and the CA blocks strengthen the network to focus on forged regions.

Validity of the CA block. We insert the CA block into Transformer and CNN to further examine its validity and universality. PoolFormer-S (PF) [32] and ConvNeXt-S (CNX) [33] are chosen as backbone. The results on FF++ (LQ) are displayed in Tab. 3, where * means loading pre-trained weight. Embedding CA blocks significantly improves the performance of both baseline networks due to their critical attention to central regions.

Convolution kernel size. In the AFE module, we conduct experiments with

Table 4: Quantitative results of different convolution kernel sizes in AFE.

Kernel	Acc	AUC
[2, 4, 8]	**90.40**	**95.09**
[2, 8, 16]	90.09	94.04
[4, 8, 16]	89.79	94.10

Table 5: The results on FF++ (LQ) with different α.

α	Acc	AUC
0.5	**90.40**	**95.09**
0.6	90.11	94.59
0.7	90.13	94.26

several convolution kernel combinations under the same settings. The specific results are shown in Tab. 4. The combination of [2, 4, 8] achieves the best performance.

Hyperparameter α. The hyperparameter α has a significant impact on the CA block's performance by restricting the size of the central area. In Tab. 5, we conduct experiments with different value of α and find that the optimal performance is achieved when the α is 0.5. It means that the inclusion of too much irrelevant information would weaken the performance, and the center area can supply adequate forgery traces.

4.4 Visualizations

To further understand how our method makes decisions, we use Grad-CAM [34] to show the attention maps of input samples for both the baseline and CAN. Fig. 4 demonstrates that all four forgery methods have their faked areas centered in the center. The baseline network is significantly disturbed as a result of a considerable increase in noise information after compression. However, with the AFE module filtering out noise information and Central Attention emphasis focused on central areas, the CAN can more reliably capture forgery traces.

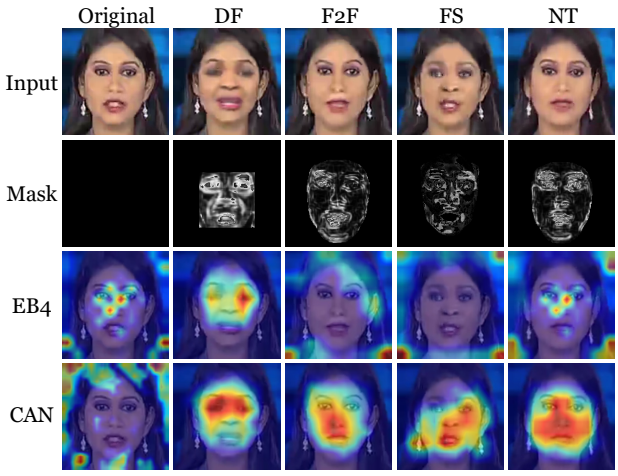

Fig. 4: The attention maps for different kinds of faces

4.5 Limitations

When applying improper masks, the performance drops significantly, suggesting that a more meticulous attention mechanism is required. Focusing on specific facial components may lead to better results, which we will explore in the future.

5 Conclusion

The paper proposes a Central Attention Network (CAN) framework for detecting forged images. We conduct a comprehensive analysis of the frequency amplification forgery traces, which has laid a strong foundation for the network's optimal performance. The Central Attention block filters out irrelevant background noise effectively, ensuring the network concentrates primarily on capturing forgery traces. Visualizing class activation mapping explains the internal mechanism and demonstrates the effectiveness of our methodology.

Acknowledgment

This research is supported by National Natural Science Foundation of China (Grant No. 62206277) and the University Synergy Innovation Program of Anhui Province (No. GXXT-2022-036). The authors would like to thank Ran He (Professor at CASIA) and Jiaxiang Wang (Ph.D. at AHU) for their valubale suggestions.

References

1. Tora: Deepfakes (2018), `https://github.com/deepfakes/faceswap/tree/v2.0.0`
2. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proc. CVPR (2016)
3. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: IEEE International Conference on Acoustics, Speech and Signal Processing (2019)
4. Matern, F., Riess, C., Stamminger, M.: Exploiting visual artifacts to expose deepfakes and face manipulations. In: IEEE Winter Applications of Computer Vision Workshops (2019)
5. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: Proc. CVPR (2021)
6. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: IEEE International Workshop on Information Forensics and Security (2018)
7. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: Proc. CVPR (2020)
8. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. In: Proc. CVPR (2021)
9. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: Wilddeepfake: A challenging real-world dataset for deepfake detection. In: Proc. ACM-MM (2020)
10. Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K.: On the detection of digital face manipulation. In: Proc. CVPR (2020)
11. Wang, C., Deng, W.: Representative forgery mining for fake face detection. In: Proc. CVPR (2021)

12. Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., Ji, R.: Local relation learning for face forgery detection. In: Proc. AAAI (2021)
13. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: Proc. ICML (2020)
14. Gu, Q., Chen, S., Yao, T., Chen, Y., Ding, S., Yi, R.: Exploiting fine-grained face forgery clues via progressive enhancement learning. In: Proc. AAAI (2022)
15. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: Proc. ECCV (2020)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Proc. NeurIPS (2017)
17. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: Proc. CVPR (2021)
18. Li, J., Xie, H., Li, J., Wang, Z., Zhang, Y.: Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In: Proc. CVPR (2021)
19. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proc. ICCV (2021)
20. Ren, S., Zhou, D., He, S., Feng, J., Wang, X.: Shunted self-attention via multi-scale token aggregation. In: Proc. CVPR (2022)
21. Wang, W., Yao, L., Chen, L., Lin, B., Cai, D., He, X., Liu, W.: Crossformer: A versatile vision transformer hinging on cross-scale attention. In: Proc. ICLR (2022)
22. Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.G., Li, S.N.: M2tr: Multi-modal multi-scale transformers for deepfake detection. In: Proc. ICMR (2022)
23. Ni, Y., Meng, D., Yu, C., Quan, C., Ren, D., Zhao, Y.: Core: Consistent representation learning for face forgery detection. In: Proc. CVPR Workshops (2022)
24. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proc. CVPR (2017)
25. Masi, I., Killekar, A., Mascarenhas, R.M., Gurudatt, S.P., AbdAlmageed, W.: Two-branch recurrent network for isolating deepfakes in videos. In: Proc. ECCV (2020)
26. Sun, K., Liu, H., Yao, T., Sun, X., Chen, S., Ding, S., Ji, R.: An information theoretic approach for attention-driven face forgery detection. In: Proc. ECCV (2022)
27. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proc. ICCV (2019)
28. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proc. CVPR (2020)
29. Kowalski, M.: Faceswap (2018), `https://github.com/marekkowalski/faceswap`
30. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. Acm Transactions on Graphics (2019)
31. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: Proc. ICML (2019)
32. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: Proc. CVPR (2022)
33. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proc. CVPR (2022)
34. Li, H., Huang, J.: Localization of deep inpainting using high-pass fully convolutional network. In: Proc. ICCV (2019)