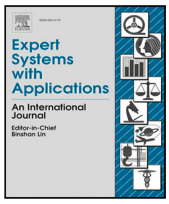
Expert Systems With Applications 276 (2025) 127108

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



# MH-FFNet: Leveraging mid-high frequency information for robust fine-grained face forgery detection

Kai Zhou [a], Guanglu Sun [a,*], Jun Wang [b], Linsen Yu [a], Tianlin Li [a]

[a] School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China
[b] School of Information Science And Technology, Great Bay University, Dongguan 523000, China

ARTICLE INFO

*Keywords:*
Face forgery detection
Fined-grained classification
Frequency cues
Enhancement module

ABSTRACT

The rapid advancement of Deepfake technology has rendered the generation of forged faces highly realistic, while simultaneously introducing significant societal security concerns. The accurate detection of forged facial images has thus emerged as an urgent issue and a formidable challenge. In this paper, we approach face forgery detection as a fine-grained classification problem due to the subtle differences between real and fake faces. We propose a detection framework termed the Mid-High Frequency Based Fine-Grained Network (MH-FFNet), which enhances the detection of forged faces by leveraging mid- and high-frequency information to capture fine-grained forgery cues. To better extract and utilize these cues, we devise two fine-grained feature enhancement modules: the Patch-based Fine-Grained Enhancement Module (P-FGEM) and the Feature-based Fine-Grained Enhancement Module (F-FGEM). The P-FGEM module focuses on extracting mid- and high-frequency information from shallow feature blocks, enhancing forgery representations in shallow features. This design effectively mitigates the loss of mid- and high-frequency cues as the network deepens, thereby improving the algorithm's sensitivity to forgery cues. In contrast, the F-FGEM module captures mid- and high-frequency information from mid-level global features, further enriching forgery representations in these features and significantly enhancing their discriminative power. Experimental results indicate that our proposed method achieves an AUC of 99.44% on the FF++ (C23) dataset and 83.44% on the Celeb-DF (V2) dataset, demonstrating the algorithm's superior detection capability and generalization performance. Additionally, we conduct experiments to comprehensively illustrate the robustness of the algorithm against common image post-processing attacks.

## 1. Introduction

With the continuous emergence of various generative models, Deepfake technology has achieved significant success, leading to the development of numerous facial forgery methods (Liu, Perov, Gao, Chervoniy, Zhou, & Zhang, 2023; Thies, Zollhofer, Stamminger, Theobalt, & Nießner, 2016; Wang, Alamayreh, et al., 2022; Zhang, Yu, Huang, Shen, & Ren, 2024). Videos and images generated through Deepfake have become increasingly indistinguishable from real ones to the human eye. The easy accessibility of Deepfake technology enable anyone to create deceptive videos, resulting in the widespread dissemination of Deepfake content on the internet and social media. This inevitably triggers a crisis of social trust and public panic (Tolosana, Vera-Rodriguez, Fierrez, Morales, & Ortega-Garcia, 2020). To eliminate the hazard caused by Deepfake, extensive research has been conducted focusing on face forgery detection.

Driven by the superior feature learning capabilities, existing deep learning-based algorithms have achieved excellent detection performance within intra-datasets (Afchar, Nozick, Yamagishi, & Echizen, 2018; Rossler, Cozzolino, Verdoliva, Riess, Thies, & Nießner, 2019; Zhao, Zhou, Chen, Wei, Zhang, & Yu, 2021). However, in real-world applications, the detection performance is often unsatisfactory when faced with data generated by unknown face forgery techniques. Therefore, Deepfake detection models need to consider not only the detection performance within intra-datasets but also accurate judgments when facing unseen forgeries under cross-dataset condition, thus enhancing the model's generalization capability. To this end, many studies (Gu, Chen, Yao, Chen, Ding, & Yi, 2022; Liu, Li, et al., 2021; Miao, Tan, Chu, Yu, & Guo, 2022; Qian, Yin, Sheng, Chen, & Shao, 2020) propose using frequency-based information to learn representative generalized features. Jeong, Kim, Min, Joe, Gwon, and Choi (2022) directly apply the

* Corresponding author.
*E-mail addresses:* 2110403072@stu.hrbust.edu.cn (K. Zhou), sunguanglu@hrbust.edu.cn (G. Sun), wangjun@gbu.edu.cn (J. Wang), yulinsen@hrbust.edu.cn (L. Yu), 2010400006@stu.hrbust.edu.cn (T. Li).

https://doi.org/10.1016/j.eswa.2025.127108
Received 30 September 2024; Received in revised form 3 February 2025; Accepted 27 February 2025
Available online 10 March 2025

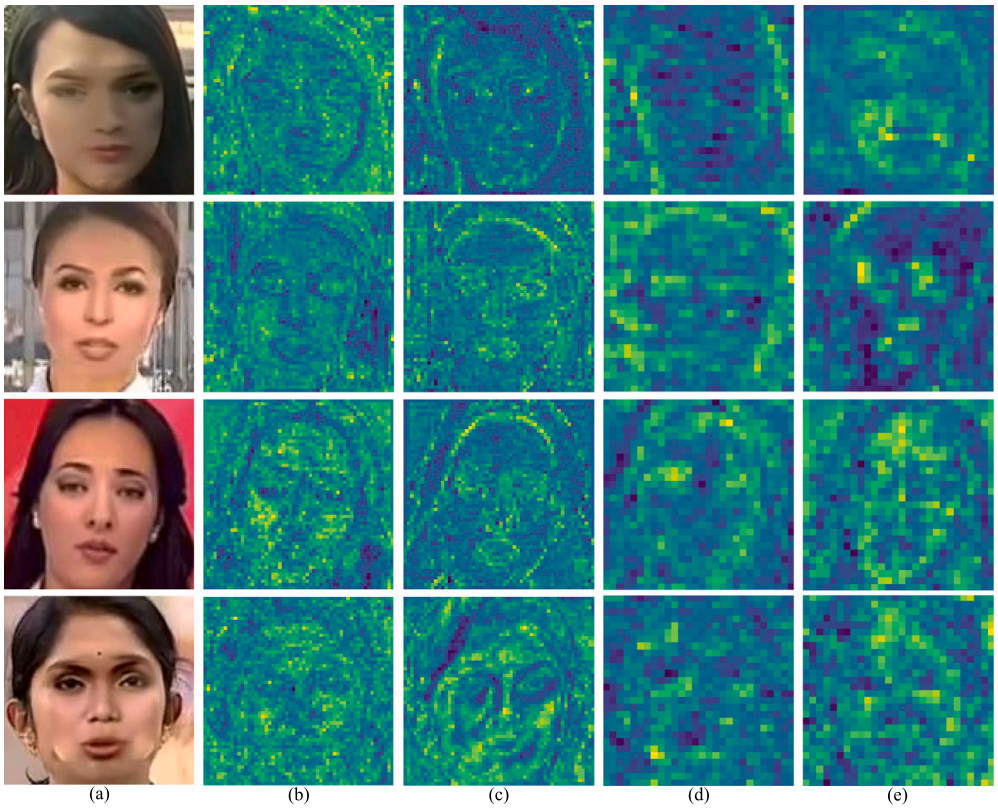



**Fig. 1.** Visualizing features of Deepfakes (DF) at different network stages. Column (a) displays images from the Deepfakes (DF), column (b) represents shallow features, column (c) shows shallow features enhanced by the P-FGEM, column (d) represents mid-level features, and column (e) represents mid-level features enhanced by the F-FGEM.

Discrete Fourier Transform to RGB images and inputs high-frequency components into Convolutional Neural Network (CNN). M2TR (Wang, Wu, et al., 2022) treats frequency features as attentional artifacts to enhance feature expression and improve generalization. Although these approaches have achieved considerable performance, the frequency features extracted by their strategies are relatively coarse, which burdens the discriminative feature patterns of the network. Additionally, directly using frequency domain features or connecting RGB and frequency domain features at the end of the network for face forgery detection distracts networks' attention on subtle and discriminative features.

Inspired by fine-grained classification task that aims to distinguish between highly similar categories such as bird species, car models, and airplane types (Du et al., 2020; Hu, Qi, Huang, & Lu, 2019; Xiao, Xu, Yang, Zhang, Peng, & Zhang, 2015), in this work, we delve deeply into the impact of frequency information on enhancing the generalization capability of the face forgery detector, resulting in a Mid-High[1] Frequency Based Fine-Grained Network (MH-FFNet). Given that the distinguishing features between authentic and manipulated faces can be extremely subtle, our approach focuses on identifying these subtle forgery clues within face classification tasks to improve detection accuracy. We achieve this by selecting representative features based on the influence of frequency information: low frequency typically represents the style and content of an image, while mid-high frequency represents its texture and fine details. On the other hand, based on the fact that the manipulated face videos are blurry and have less texture, we exploit mid- and high-frequency information for fine-grained feature learning, while discarding low-frequency information that presents the same content in individual deepfake videos. To better extract and utilize these fine-grained features, we design two fine-grained enhancement modules at different network stages.

Specifically, we focus on extracting mid- and high-frequency information from shallow and mid-level features to capture subtle tampering clues, employing these clues to construct two fine-grained enhancement modules for accurate detection of real and fake faces. In the carefully designed Patch-based Fine-Grained Enhancement Module (P-FGEM), we apply patch processing to shallow features and perform frequency domain transformations on shallow feature patches. By fully leveraging the mid- and high-frequency information within these patches, we effectively enhance tampering traces in shallow features, improving the algorithm's sensitivity to tampering information. As shown in Fig. 1(c), the network pays more attention on tampered details of Deepfakes (DF)[2] than column (b). This innovative design mitigates the issue of artifact features diminishing as the network depth increases, preserving key texture information and preventing the negative impact of information loss on detection performance. In addition, compared to shallow features, mid-level features are typically more refined and contain less redundant information. Furthermore, compared to high-level features, they provide more intuitive and valuable information. Therefore, in the Feature-based Fine-Grained Enhancement Module (F-FGEM), we extract and utilize global mid- and high-frequency information from mid-level features, enhancing their richness and significantly improving their discriminative power. By strategically integrating fine-grained enhancement modules at different network level features, our proposed method accurately captures highly discriminative features embedded in subtle clues. This approach not only significantly improves the detection of subtle differences but also enhances the model's generalization capacity in leveraging frequency cues. By deeply exploring and fully utilizing mid- and high-frequency information, our method demonstrates outstanding performance and generalization capabilities in Deepfake detection.

The main contributions of this paper are as follows.

[1] Middle and High-frequency information.

[2] Deepfakes: https://github.com/deepfakes/faceswap.

1. Inspired from fine-grained classification, we propose a fine-grained face forgery detection framework that integrates mid- and high-frequency information, named the Mid-High Frequency Based Fine-Grained Network (MH-FFNet). This framework aims to better capture fine-grained tampering clues for face forgery detection by fusing mid- and high-frequency information.
2. In the face forgery detection task, we introduce mid- and high-frequency information as fine-grained clues. Additionally, we design two attention-based fine-grained feature enhancement modules, P-FGEM and F-FGEM, which respectively enhance the detailed information in shallow features and the semantic information in mid-level features. This approach improves the generalization capability of the model.
3. Extensive experiments demonstrate that the proposed method not only achieves good results within intra-datasets but also performs significantly well in inter-dataset scenarios. Additionally, it exhibits excellent robustness performance when facing various image attacks.

The rest of this paper is organized as follows. In Section 2, we introduce the face forgery detection methods and fine-grained classification. In Section 3, we present a detailed explanation of proposed methods. The experimental results are presented to demonstrate the performance of the proposed method in Section 4. Finally, in Section 5, we summarize the paper.

## 2. Related work

### 2.1. Face forgery detection

Early detection algorithms (Fridrich & Kodovsky, 2012; Li & Lyu, 2018; Matern, Riess, & Stamminger, 2019) traditionally utilize statistical features or detect obvious spatial defects, such as eye blinks (Li & Lyu, 2018), visual artifacts (Matern et al., 2019), and uncoordinated head postures (Yang, Li, & Lyu, 2019). The rapid advancement of deep learning has enabled researchers to employ CNN to automatically learn discriminative forgery features from manipulated inputs. Spatial domain clues, being more easily detectable, have led to the proposal of numerous deep learning detection algorithms that have demonstrated good performance. For example, Rossler et al. (2019) use the Xception network structure to analyze its custom dataset, achieving significant detection performance. Li, Bao, et al. (2020) propose 'Face X-ray', a face forgery detection method that relies on intrinsic image discrepancies at blending boundaries. Similarly, Zhao et al. (2021) employ multi-scale attention mechanisms to capture features in the spatial domain, achieving superior results when tested on the same dataset. Guo, Yang, Zhang, and Xia (2023) reconsider the impact of gradient information in face forgery detection, enhancing authenticity assessment by exploring various filter combinations to extract gradient cues. Wang and Chow (2023) investigate forensic noise traces in Deepfake detection, developing a siamese network to extract noise traces from face and background blocks and introducing a multi-head relative interaction method to evaluate face-background interactions. Zhang, Li, Sangaiah, Li, Deng, and Wu (2024) introduce a two-branch convolutional network incorporating texture suppression. They develop a Texture Suppression Module (TSM) using convolution to attenuate image content textures and extract forgery traces. Gao, Micheletto, et al. (2024) develop the Texture and Artifact Detector (TAD), which improves model generalization by separately analyzing texture inconsistencies and artifact information in spatial domain. However, spatial domain clues are fragile and can be easily lost when images are subjected to adversarial attacks. Additionally, an overreliance on deep learning to capture spatial features can limit the model's adaptability, as it may become overly dependent on the feature distribution of the training dataset. This dependency can result in a significant decline in the detection performance of spatial domain-based models when they encounter unknown datasets.

To address the issue of decreased generalization performance, researchers (Corvi, Cozzolino, Poggi, Nagano, & Verdoliva, 2023; Frank, Eisenhofer, Schönherr, Fischer, Kolossa, & Holz, 2020) study the manipulation process and find that upsampling is an unavoidable step. During the upsampling process, obvious tampering clues are left in the frequency domain. As a result, frequency domain features become discriminative common feature in different generated datasets. Several researchers introduce frequency domain features for face forgery detection, which partially address the generalization problem. For instance, F3-Net (Qian et al., 2020) adopts a two-stream approach, leveraging global and local frequency-aware cues to accurately detect forgery patterns. SPSL (Liu, Li, et al., 2021) employs a combination of RGB features and spectral features to capture artifact information introduced by upsampling. HFI (Miao et al., 2022) utilizes middle and high frequency information and constructs channel attention to detect face forgery content. Additionally, HIFE (Gao, Xia, et al., 2024) utilizes high-frequency information from DCT and DWT to create a frequency domain enhancement module, addressing the issue of reduced detection performance in low-quality images caused by the lack of high-frequency details. Zhang, Chen, Liao, Li, Chen, and Yang (2024) design a dual-feature fusion module to integrate RGB and high-frequency noise features, alongside a local enhancement attention module to improve the model's focus on tampering traces. Wang, Wu, Wang, Zhang, Wei, and Song (2024) propose a spatial-frequency fusion method for Deepfake detection, leveraging knowledge distillation, attention mechanisms, and multi-knowledge transfer to improve forged region localization and enhance generalization on compressed images. However, current utilization of frequency domain features remains rough, either by directly using frequency domain information as detection features or simply concatenating them with RGB features at the end of the network, without fully exploring the intricate use of frequency domain features, thereby overlooking some discriminative subtle features.

### 2.2. Fine-grained classification

Fine-grained classification is a challenging task in computer vision due to it focuses on distinguishing between similar types, such as different bird species that share similar appearances (Hu et al., 2019; Xiao et al., 2015). The key aspect of fine-grained classification lies in identifying and learning discriminative regions within images, as well as encouraging the fusion of these features from various regions. In certain fine-grained classification tasks, attention models are specifically designed to capture local, subtle, and discriminative features, which enhances the accuracy of the classification models (Du et al., 2020; Yang, Luo, Wang, Hu, Gao, & Wang, 2018). Zhu, Gao, Wang, Zhou, and Li (2024) design a novel network architecture called FicNet, which efficiently classifies few-shot fine-grained images by integrating a Multi-Frequency Neighborhood (MFN) module, designed using frequency channel attention, and a Double-Cross Modulation (DCM) module. Consequently, capturing and leveraging these discriminative fine-grained clues is crucial for the efficacy of fine-grained classification.

In recent years, advancements of generative models have produced images and videos that are increasingly indistinguishable from real ones to the human eye. The subtle differences between real and fake content make face forgery detection akin to fine-grained classification tasks (Zhao et al., 2021), where the goal is to classify highly similar entities. However, relying solely on deep learning detectors may not effectively capture more discriminative features to differentiate between real and fake content.

Taking inspiration from Chandrasegaran, Tran, and Cheung (2021); Miao et al. (2022); Corvi et al. (2023), we use mid- and high-frequency information that can effectively represent edge and subtle information as fine-grained features. Building upon the extraction of local features, we further refine these features and extract mid-high frequency information from local feature blocks to obtain more local subtle features. Additionally, we consider the influence of global subtle features by extracting global mid-high frequency information from features and complementing the local subtle features.

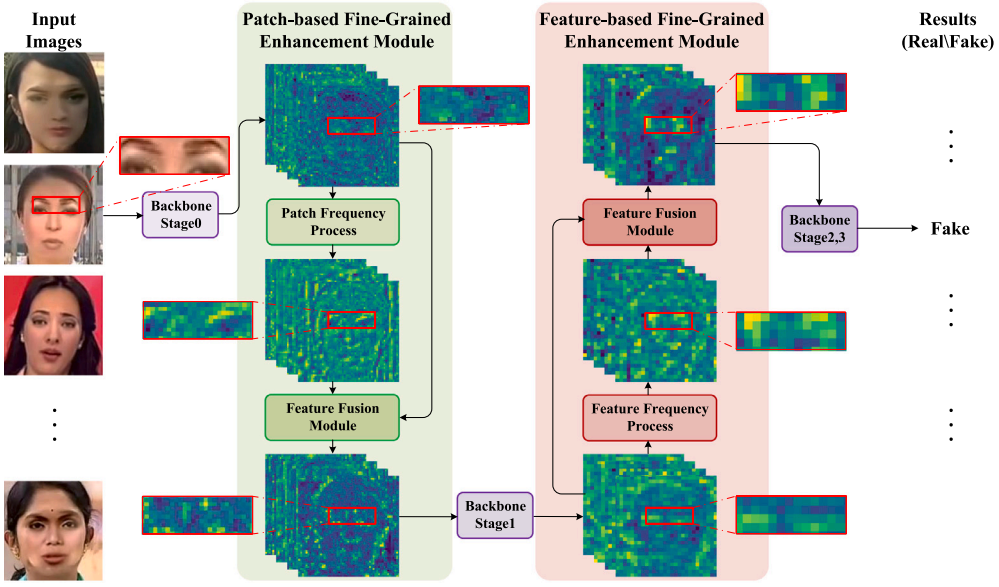



**Fig. 2.** The architecture of our proposed method. The shallow and mid-level features can be effectively enhanced by Patch-based Fine-Grained Enhancement Module (P-FGEM) and Feature-based Fine-Grained Enhancement Module (F-FGEM), respectively. The red boxes mark clearly tampered areas.

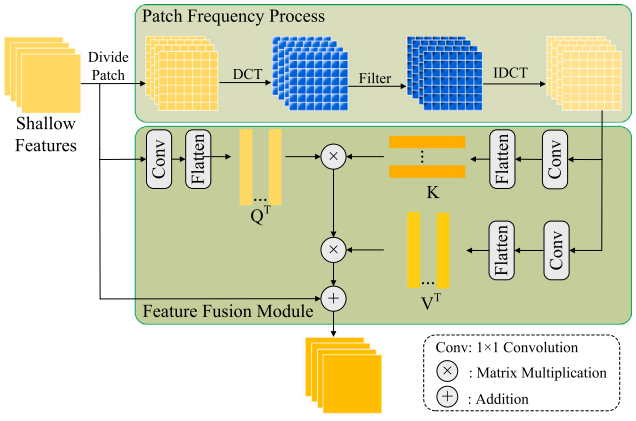



**Fig. 3.** Patch-based Fine-Grained Enhancement Module (P-FGEM). Each image patch undergoes frequency domain processing, where mid-high frequency features are transformed back into the spatial domain to obtain. The original shallow features are then combined with these frequency domain-processed features to construct attention mechanisms, aiming to enhance the shallow features.

## 3. Method

### 3.1. Overview of the proposed framework

The MH-FFNet architecture is depicted in Fig. 2. The red boxes highlight the tampered areas. As indicated by the marked red region on the feature map, our proposed feature enhancement modules excel at focusing the model's attention on the tampered areas. This improvement surpasses the performance of the original shallow and mid-level features, thereby enhancing the model's detection capability.

Specifically, for an input RGB image $X \in R^{3\times H\times W}$, the shallow features $F \in R^{C\times M\times N}$ are first extracted via the stage0 of the Convnext network. Then, the proposed Patch-based Fine-Grained Enhancement Module (P-FGEM) is employed, integrating local fine-grained clues $F' \in R^{C\times M\times N}$ into $F$. The enhanced shallow features $ELF \in R^{C\times M\times N}$, output by P-FGEM, are then fed back to stage1 of the Convnext network to obtain mid-level features $MF \in R^{C'\times M'\times N'}$. The mid- and high-frequency information, which potentially lost after stage1, can be further enhanced by the proposed Feature-based Fine-Grained Enhancement Module (F-FGEM). Finally, the enhanced mid-level features $EMF \in R^{C'\times M'\times N'}$ passe through the remaining layers of the Convnext network for detection.

### 3.2. Fine-grained enhancement module

Texture features are complex visual characteristics that can describe the roughness and regularity of images. In general, manipulated videos appear visually smoother than real ones, leading to a loss of facial texture details. The detection of these texture features has been widely acknowledged as a factor in improving the generalization of Deepfake detectors (Gao, Micheletto, et al., 2024; Liu, Xie, Wang, & Zha, 2024; Zhao, Jin, Gao, Wu, Yao, & Jiang, 2023). However, as the depth of neural networks increases, these crucial texture details tend to fade, resulting in performance drops when encountering unknown manipulated videos. Ensuring the preservation of texture differences throughout the network can significantly enhance generalization performance.

Inspired by fine-grained classification tasks, we regard these detailed texture information as fine-grained features and present the Fine-Grained Enhancement Module (FGEM). Unlike a simple enhancement module, FGEM preserves texture information in the frequency domain both locally and globally at various stages of the network, ensuring the retention of these essential details. In this paper, the FGEM comprises two key components: the Patch-based Fine-Grained Enhancement Module (P-FGEM) and the Feature-based Fine-Grained Enhancement Module (F-FGEM).

#### 3.2.1. Patch-based fine-grained enhancement module

Patch-based Fine-Grained Enhancement Module (P-FGEM) is utilized to enhance the shallow features in the first stage of the network, preserving most of the texture and color information. Fig. 3 illustrates the architecture of the proposed module P-FGEM. The mid- and high-frequency information, representing subtle clues in the features and reflecting fine differences, is filtered from the shallow features and then reassigned to the features using a self-attention mechanism. More specifically, the shallow features $F \in R^{C\times M\times N}$ undergo non-overlapping patch processing, followed by the application of the Discrete Cosine Transform (DCT) to the feature patches, capturing the differences in local regions of the feature maps.

For each feature patch $x^p \in R^{C\times m\times n}$, where $m$ and $n$ represent the height and width of the feature patch, the input features can be redefined as $F = \{x^p\}_{p=1}^{(M/m)\times(N/n)}$, $(M/m)\times(N/n)$ representing the number of feature patches. We utilize the DCT to obtain the frequency domain coefficients of the feature patches. For the $p$th feature patch $x_{i,j}^p$, $(i = 0, \dots, m,\ j = 0, \dots, n)$, the 2-D DCT is calculated by

$$DF_{u,v}^p = \sum_{i=0}^{m-1}\sum_{j=0}^{n-1} x_{i,j}^p \times \cos\left(\frac{\pi u}{m}\left(i+\frac{1}{2}\right)\right)\times\cos\left(\frac{\pi v}{n}\left(j+\frac{1}{2}\right)\right), \tag{1}$$

where $u$ and $v$ are the vertical and horizontal frequency domain indices, respectively, and $DF_{u,v}^p$ represents the DCT coefficients of the $p$th feature patch. We discard the low-frequency information and retain the mid-high frequency information $DF_{u,v}^{p'}$ by filter. Then, in order to maintain the translational invariance and local consistency of natural images, we transform the DCT feature patch back to the feature domain. The inverse DCT of the $p$th feature patch is calculated by

$$x\_f_{i,j}^{p'} = \sum_{i=0}^{m-1}\sum_{j=0}^{n-1} DF_{u,v}^{p'} \times \cos\left(\frac{\pi u}{m}\left(i+\frac{1}{2}\right)\right)\times\cos\left(\frac{\pi v}{n}\left(j+\frac{1}{2}\right)\right). \tag{2}$$

The mid- and high-frequency information is extracted exclusively from the feature patches using this approach. These patches are aggregated to derive shallow features $F'$ enriched with fine-grained clues and then use $F'$ to enhance the original shallow features. To fuse the fine-grained clues $F'$ with shallow features $F$, we draw inspiration from the Vision Transformer (ViT), which uses self-attention (Dosovitskiy et al., 2021) to facilitate interactions between different patches. Our feature fusion module not only considers the interaction between the corresponding channels of the two features but also facilitates interactions between different channels of the two features. This approach enhances the fusion process, leading to a more effective integration of the fine-grained clues and shallow features. First, we apply three $1\times1$ convolutions $C_1$ $(C_1^1, C_1^2, C_1^3)$ to $F$ and $F'$ separately to obtain their corresponding features:

$$F_Q = C_1^1(F), F_K' = C_1^2(F'), F_V' = C_1^3(F'), \tag{3}$$

and we have these corresponding features $F_Q$, $F_K'$ and $F_V' \in R^{C\times M\times N}$. Then, we compute the fused feature of $F_Q$, $F_K'$ and $F_V'$ by

$$F_u = \mathrm{softmax}\left(\frac{\mathcal{F}\left(F_Q'\right)^T\mathcal{F}\left(F_K'\right)}{\sqrt{M\times N\times C}}\right)\times\mathcal{F}\left(F_V'\right)^T, \tag{4}$$

where $\mathcal{F}(\cdot)$ represents the flatten function that flattens the features along the channel dimension. At the end, we utilize residual connection and $1\times1$ convolution $C_1^4$ to further fuse the shallow features and the shallow features based on fine-grained clues, obtaining the fine-grained enhanced shallow features $ELF$ by

$$ELF = C_1^4(F + F_u). \tag{5}$$

#### 3.2.2. Feature-based fine-grained enhancement module

Merely applying P-FGEM to the shallow features is not sufficient, as feature loss may occur due to the deepening of the model structure. Considering that mid-level features contain stronger semantic information than shallow features and more visual details than deep features, we address this issue by enhancing tampering clues at the mid-level. Instead of applying patch-wise processing, we directly perform DCT on the mid-level features to leverage the global mid-high frequency information in the proposed Feature Fine-Grained Enhancement Module (F-FGEM). Fig. 4 illustrates the architecture of the proposed module F-FGEM. Unlike P-FGEM, we eliminate patch processing and perform DCT on each channel of the mid-level network features, denoted as $MF \in R^{C'\times M'\times N'}$. Then, we discard the low-frequency information of the features, retaining only the mid-high frequency information of the mid-level global features. Afterward, we apply the inverse DCT to the mid-level features with the retained mid-high frequency information, resulting in mid-level features $MF' \in R^{C'\times M'\times N'}$ that contain only fine-grained clues

$$MF' = D^{-1}\left(Fi\left(D\left(MF\right)\right)\right), \tag{6}$$

where $D$ represents the DCT operation, $Fi$ indicates the filter, $D^{-1}$ indicates the inverse DCT operation.

Then the processed features $MF'$ and $MF$ are fed into the self-attention block, and the output features are concatenated with the $MF$ as Eqs. (3)–(5). Finally, the enhanced features $EMF$ can be obtained through F-FGEM. The overall learning algorithm of MH-FFNet is provided in Algorithm 1.

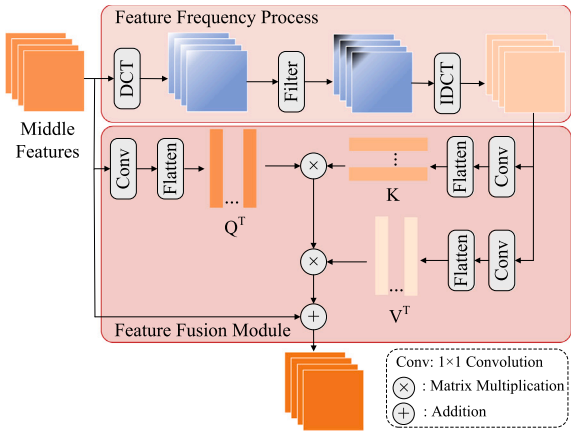



**Fig. 4.** Feature-based Fine-Grained Enhancement Module (F-FGEM). Middle features undergoes frequency domain processing, where mid-high frequency features are transformed back into the spatial domain. The original middle features are then combined with these frequency features to construct attention mechanisms, aiming to enhance the middle features.

## 4. Experiments

### 4.1. Experiment setup

**Datasets** To validate the effectiveness of the proposed method, we mainly conduct experiments on FaceForensics++ (FF++) dataset (Rossler et al., 2019), Celeb-DF (V2) dataset (Li, Yang, Sun, Qi, & Lyu, 2020), and Deepfake Detection Challenge (DFDC) dataset (Dolhansky et al., 2020). The FF++ dataset (Rossler et al., 2019) contains three versions: original (Raw), lightly compressed (C23), and highly compressed (C40). Each version consists of 1000 real videos and 4000 corresponding fake videos. These 4000 fake videos are generated using four manipulation methods: DeepFakes (DF), Face2Face (F2F) (Thies et al., 2016), FaceSwap (FS),[3] and NeuralTextures (NT) (Thies, Zollhöfer, & Nießner, 2019). Since these detection models achieve nearly perfect detection performance on the Raw version, we use the C23 and C40 versions of FF++ for our experiments. When training and validating the FF++ dataset, we follow Rossler et al. (2019) and select 720 training videos, 140 validation videos, and 140 testing videos for every 1000 videos. We only use 32 frames of each video for training, while 100 frames are selected for validation and testing. Celeb-DF (V2) dataset (Li, Yang, et al., 2020) contains 890 real videos and 5639 high-quality fake videos. DFDC (Dolhansky et al., 2020) dataset includes over 2000 real videos and more than 100 000 fake videos. In this paper, we only use the testing set of Celeb-DF (V2) and DFDC datasets for testing purposes. We use MTCNN (Zhang, Zhang, Li, & Qiao, 2016) to extract faces from frames and resize them to $224\times224$.

[3] FaceSwap: http://www.github.com/MarekKowalski.

**Table 1**
Intra-dataset evaluation results (AUC (%) and ACC (%) ) on FF++ dataset. The bolded results are the best, while the underlined ones are the second-best. Note that the results for comparisons are from Miao et al. (2022). *Indicates the model is trained by us using the official code.

| Methods | Reference | Backbone | C40 | | C23 | |
|---|---|---|---|---|---|---|
| | | | AUC | ACC | AUC | ACC |
| MseoNet (Afchar et al., 2018) | WIFS 2018 | MseoNet | – | 70.47 | – | 83.10 |
| Xception (Rossler et al., 2019) | ICCV 2019 | Xception | 81.76 | 80.32 | 94.86 | 92.39 |
| F3-Net (Qian et al., 2020) | ECCV 2020 | Xception | 87.22 | 84.53 | 97.80 | 93.12 |
| M2TR (Wang, Wu, et al., 2022) | ICMR 2022 | Efficient-B4 | 87.15 | 83.89 | 96.75 | 91.86 |
| SPSL* (Liu, Li, et al., 2021) | CVPR 2021 | Xception | 84.60 | 84.03 | 98.68 | 95.51 |
| MAT* (Zhao et al., 2021) | CVPR 2021 | Efficient-B4 | **88.68** | **86.15** | 98.87 | 96.61 |
| HFI-Net (Miao et al., 2022) | TIFS 2022 | ViT | 88.40 | 85.69 | 97.07 | 91.87 |
| GocNet* (Guo et al., 2023) | ESWA 2023 | ResNet50 | 81.96 | 83.68 | 97.47 | 93.48 |
| HIFE* (Gao, Xia, et al., 2024) | ESWA 2024 | Xception | 83.62 | 84.99 | 98.83 | 95.66 |
| Proposed | | Convnext | 87.44 | 85.90 | **99.44** | **97.37** |

**Algorithm 1 Main algorithm of training MH-FFNet**

**Input:** Training dataset $T = X, Y$ with image data $X$ and label $Y$;
Validation dataset $V$ ;
Hyperparameters: Batch size $B$, training epoch $E$;
**Output:** The proposed Method for face forgery detection.
**Initialize:** Initialize parameters of Convnext with Imagenet pretrained weights;

**for** epoch in $E$ **do**
  **for** Image $X$ in Batch $B$ **do**
    Shallow features $F \leftarrow$ Convnext stage 0 $(X)$;
    Divide the shallow features into non-overlapping patches to obtain feature patches;
    **for** all feature patches **do**
      Apply DCT to each feature patch and discard low-frequency information by filter, then apply inverse DCT;
    **end for**
    $F\_Q \leftarrow C_1^1(F)$, $F'\_K \leftarrow C_1^2(F')$, $F'\_V \leftarrow C_1^3(F')$;
    Obtain the fine-grained enhanced shallow features $ELF$ by Eqs. (4) and Eqs. (5);
    Middle level features $MF \leftarrow$ Convnext stage 1 $(ELF)$;
    Apply DCT to $MF$ and discard low-frequency information, then apply inverse DCT to obtain $MF'$;
    Sent $MF'$ and $MF$ to self-attention and output features are concatenated with $MF$ as Eqs. (3), Eqs. (4), and Eqs. (5). Obtain the enhanced features $EMF$ by F-FGEM;
    preds = Convnext stage 2, 3 $(EMF)$;
    loss = CE_loss(preds, $Y$);
  **end for**
  Update all parameters of MH-FFNet to minimize loss by AdamW;
**end for**
**return** The model with weights $\leftarrow$ argmax(ACC(MH-FFNet(V))).

**Implementation Details** Our backbone network is pre-trained Convnext-Base (Liu, Mao, Wu, Feichtenhofer, Darrell, & Xie, 2022) on ImageNet dataset (Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009). We employ the AdamW optimizer with parameters (0.9, 0.999) to train the proposed model in the open-source PyTorch framework. The initial learning rate of the model is set to 4e−5. The batch size is set to 16, and a total of 20 epochs are trained. During training, some data augmentation, such as Gaussian Blur, Gaussian Noise, and random horizontal flip. Our experimental environment, is performed on a Nvidia GeForce RTX 2080Ti GPU with 11 GB RAM.

**Metrics** For evaluation metrics, in line with most experimental setups, we employ frame-level Accuracy (ACC) and Area Under the Curve of ROC (AUC) following Rossler et al. (2019); Miao et al. (2022) as performance measures.

### 4.2. Evaluation and comparison on FF++

#### 4.2.1. Different video compression

We conduct training and validation on the FF++ dataset under two compression settings, C23 and C40, and Table 1 displays the comparative results with other methods. The experimental outcomes, shown in terms of AUC and ACC in Table 1, clearly demonstrate that our method achieves superior performance under the C23 compression setting. Specifically, our method attains an AUC score of 99.44% and an ACC of 97.37%, outperforming the second-best AUC scorer, MAT (Zhao et al., 2021), with gains of 0.57% in AUC and 0.76% in ACC. MAT (Zhao et al., 2021) utilizes spatial information to construct multiple spatial attention mechanisms to focus on tampered areas. SPSL (Liu, Li, et al., 2021) leverages the characteristics of the phase spectrum and the spectral features in the frequency domain to capture clues for detection. Under the higher compression setting of C40, our method records an AUC score of 87.44% and an ACC of 85.90%, slightly trailing behind the top performer, MAT (Zhao et al., 2021), by a margin of 1.24% in AUC and 0.25% in ACC. However, compared to the SPSL (Liu, Li, et al., 2021) algorithm, our method exhibits higher AUC and ACC scores. The primary reason is that the frequency domain information, which represents details, inevitably gets lost under high compression, impacting the detection performance of our method and SPSL (Liu, Li, et al., 2021) significantly. Subsequent robustness tests consistently validate our findings, as we conduct experiments across various compression ratios to continually substantiate this phenomenon. Methods marked with an asterisk * are those we replicate on our dataset, using the same data augmentation techniques.

#### 4.2.2. Different manipulation methods

To evaluate the generalizability of the proposed model across different deepfake forgery methods, we evaluate the proposed framework within the FF++ (C23) dataset against various manipulation methods. The model is trained on the C23 version and tested across different manipulated subsets of C23. As shown in Figs. 5 and 6, using ACC and AUC as performance metrics, our method achieves state-of-the-art results across all four manipulated sub-datasets. On the DF manipulated subset, our proposed algorithm achieves an accuracy of 97.2% and an AUC of 99.84%, surpassing the second-best performing algorithm by 1.63% and 0.61%, respectively. Particularly noteworthy is the performance on the NT (Thies et al., 2019) subset, where only the lip region is altered. Here, our model achieves an ACC of 94.65% and an AUC of 98.69%, demonstrating its effectiveness in capturing subtle, localized tampering details.

### 4.3. Generalization testing on unknown datasets

The advancement of deep forgery technology introduces sophisticated and previously unseen methods, challenging the generalization capacity of detection models. To verify the generalization of the proposed model, we conduct cross dataset validation. Our model is trained

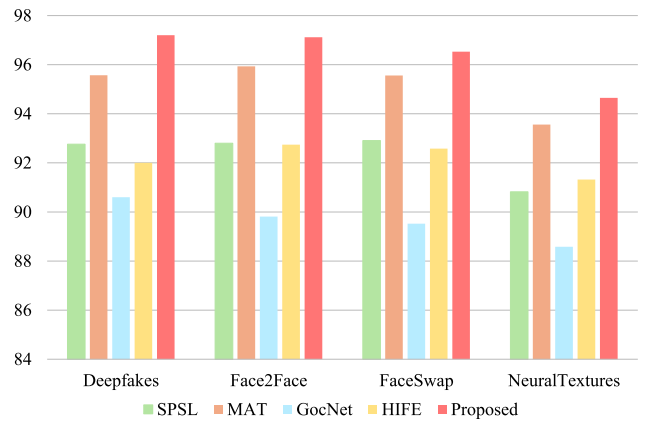



**Fig. 5.** Quantitative detection results in terms of ACC (%) on FF++ (C23) dataset with four manipulation methods: DeepFakes (DF), Face2Face (F2F) (Thies et al., 2016), FaceSwap (FS), NeuralTextures (NT) (Thies et al., 2019).

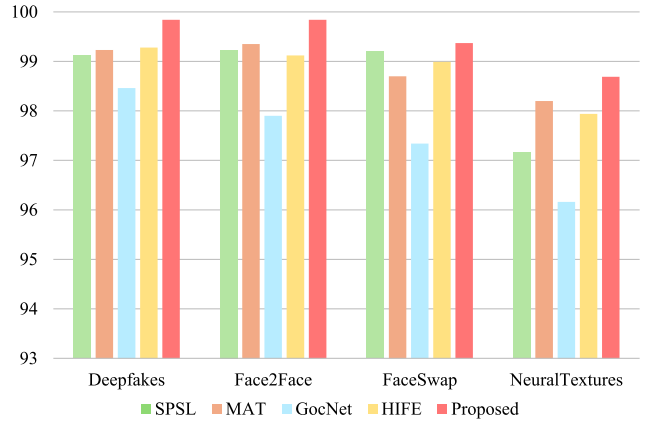



**Fig. 6.** Quantitative detection results in terms of AUC (%) on FF++ (C23) dataset with four manipulation methods: DeepFakes (DF), Face2Face (F2F) (Thies et al., 2016), FaceSwap (FS), NeuralTextures (NT) (Thies et al., 2019).

on the C23 dataset of FF++ and tested on the Celeb-DF (V2) and DFDC datasets, respectively. Due to the significant differences between different datasets, generalization testing is a challenging task for most detection algorithms. For example, compared to FF++, the Celeb-DF (V2) dataset reduces many visual artifacts during generation and post-processing, resulting in higher video quality. As expected, all methods have varying degrees of degradation in cross dataset tasks. According to the AUC results in Table 2, our proposed algorithm achieves an AUC of 83.44% on the Celeb-DF (V2) dataset and 73.83% on the DFDC dataset. The second-best performer HFI-Net (Miao et al., 2022) obtains an AUC of 83.28% on Celeb-DF (V2), which falls behind the proposed method by a margin of 0.16%. These experimental results demonstrate that our proposed algorithm not only achieves good results within the same dataset but also further enhances the generalization performance of the model by introducing frequency domain information representing fine-grained clues.

### 4.4. Robustness testing on unknown image attacks

Since most videos and images originate from the internet, they are likely subjected to various image attacks during transmission and reception, which can cause image distortion and affect the performance of Deepfake detection systems. To simulate this potential real-world scenario, we design relevant robustness testing experiments. We first train our model on the original C23 training set and then process the test set with different types of image attacks at varying intensities. As detailed in Table 3, we employ eight different types of image attacks: color saturation (CS), color contrast (CC), block-wise (BW), Gaussian noise (GN), Gaussian blur (GB), JPEG compression (JPEG), Rotate (RO), and Affine Transformation (AF). The severity of each image attack is evaluated at five levels, and examples of various attack levels are visualized in Fig. 7. For affine transformation (AF), we apply comprehensive processing to the images: first, the images are rotated by 10 degrees and translated by 10 pixels, followed by scaling the images to varying degrees between 0.8 and 1.2. Detailed attack parameters are provided in Table 3, where severity 0 indicates the absence of any attacks.

**Table 2**
Cross-dataset evaluation results on the unseen dataset. The AUC (%) results of Celeb-DF (V2) and DFDC are shown. The bolded results are the best, while the underlined ones are the second-best. Note that the results for comparisons are from Miao et al. (2022). *Indicates the model is trained by us using the official code.

| Methods | Train set | Test set | |
|---|---|---|---|
| | | Celeb-DF (V2) | DFDC |
| Xception (Rossler et al., 2019) | FF++ (C23) | 65.30 | 72.20 |
| F3-Net (Qian et al., 2020) | FF++ (C23) | 68.69 | 67.45 |
| M2TR (Wang, Wu, et al., 2022) | FF++ (C23) | 69.94 | – |
| SPSL* (Liu, Li, et al., 2021) | FF++ (C23) | 68.50 | 68.08 |
| MAT* (Zhao et al., 2021) | FF++ (C23) | 74.17 | 70.80 |
| HFI-Net (Miao et al., 2022) | FF++ (C23) | <u>83.28</u> | <u>73.65</u> |
| GocNet* (Guo et al., 2023) | FF++ (C23) | 65.56 | 66.73 |
| HIFE* (Gao, Xia, et al., 2024) | FF++ (C23) | 68.41 | 65.46 |
| Proposed | FF++ (C23) | **83.44** | **73.83** |

To demonstrate the anti-attack capabilities of our model, we present the frame-level AUC performance bar graph for six levels of interference in Fig. 8. Furthermore, we compare the average performance across severity levels 1–5 for all interference categories with previous studies in Table 4. Fig. 8 clearly demonstrates that the proposed method excels in four categories of image attacks: color contrast, block-wise, Gaussian noise, and Gaussian blur. For Affine Transformation attacks, our proposed method demonstrates superior performance across five attack intensity levels, validating its effectiveness in resisting Affine Transformation attacks. In the Rotation attacks, while our method achieves suboptimal results under large-angle rotations of 60°, 120°, and 150°, this is likely due to the extreme rotation angles causing significant distortion of image texture information, thereby hindering the effective extraction of frequency-domain clues from feature blocks. Under JPEG compression attacks, our algorithm maintains optimal effectiveness at lower compression intensities. However, as the compression level increases, the performance of the proposed method and SPSL (Liu, Li, et al., 2021), which are frequency domain-based algorithms, declines. This deterioration occurs due to the loss of detailed information at higher compression levels, resulting in a corresponding lack of frequency domain information and subsequent performance degradation in both models.

### 4.5. Ablation study

#### 4.5.1. Effectiveness on the proposed components

In this subsection, we conduct ablation experiments to demonstrate the generalization and effectiveness of our proposed module. The experiments are trained on the C23 dataset from FF++ and verified on the Celeb-DF (V2) and DFDC datasets. Table 5 presents the comparison results of the ablation experiment. To validate the effectiveness and generalization of our proposed network, we use the Convnext network as the baseline model and incrementally integrate our proposed modules into it. Specifically, 'Baseline + P-FGEM' refers to adding a P-FGEM to the Convnext baseline, while 'Baseline + F-FGEM' involves integrating a F-FGEM that incorporates global features into the Baseline. P-FGEM-S refers to a patch-based fine-grained enhancement module

**Table 3**
The eight types of image attacks in our experiment and their corresponding detailed parameters for five severity levels.

| Image Attacks | Details | Severities | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 |
| Color Saturation (CS) | Transform images by varying saturation | No | 0.4 | 0.3 | 0.2 | 0.1 | 0.0 |
| Color Contrast (CC) | Modify luminance in the image | No | 0.85 | 0.725 | 0.6 | 0.475 | 0.35 |
| Block-wise (BW) | Add random blocks on the image | No | 16 | 32 | 48 | 64 | 80 |
| Gaussian Noise (GN) | Add white Gaussian noises to the image | No | 0.001 | 0.002 | 0.005 | 0.01 | 0.05 |
| Gaussian Blur (GB) | Conduct filtering on the image | No | 3 × 3 | 5 × 5 | 7 × 7 | 9 × 9 | 13 × 13 |
| JPEG Compression (JPEG) | Conduct JPEG compression on the image | No | 90 | 70 | 50 | 30 | 20 |
| Rotate (RO) | Conduct Rotate on the image | No | 30 | 60 | 90 | 120 | 150 |
| Affine Transformation (AF) | Translate, rotate, and scale the image | No | 0.8 | 0.9 | 1.0 | 1.1 | 1.2 |

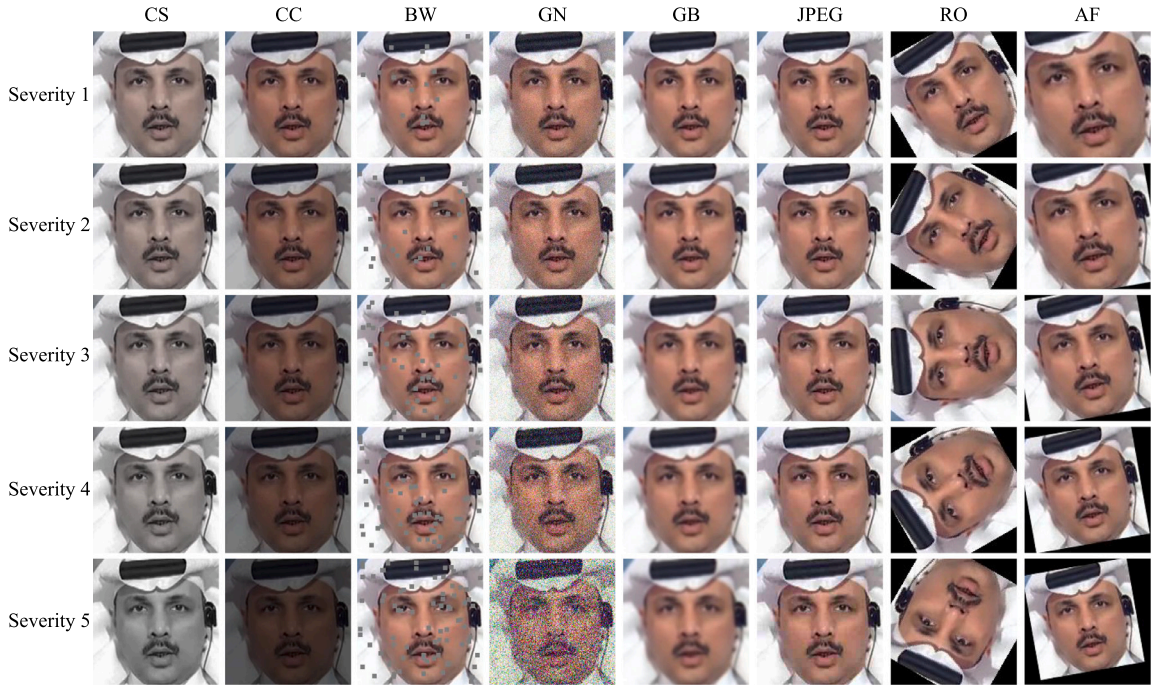



**Fig. 7.** Image visualization on the levels of severity for eight image attacks types. We utilize five severity levels for eight types of image attacks in robustness experiments.

**Table 4**
Quantitative evaluation of eight types of image attacks. The results presented in Table 5 represent the mean of AUC (%) results for 1-5 severity levels. The bolded results are the best, while the underlined ones are the second-best. *Indicates the model is trained by us using the official code.

| Methods | Image attacks | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | CS | CC | BW | GN | GB | JPEG | RO | AF |
| SPSL* (Liu, Li, et al., 2021) | 98.41 | 96.86 | 84.65 | 55.63 | 91.03 | 89.01 | 65.04 | 96.10 |
| MAT* (Zhao et al., 2021) | 98.28 | 97.97 | 91.97 | 67.66 | 96.63 | **91.57** | **85.19** | 97.86 |
| GocNet* (Guo et al., 2023) | 94.30 | 94.36 | 68.94 | 47.10 | 90.27 | 84.53 | 64.22 | 94.73 |
| HIFE* (Gao, Xia, et al., 2024) | 97.07 | 96.19 | 88.74 | 56.95 | 94.07 | 88.48 | 76.36 | 97.74 |
| Proposed | **98.42** | **98.59** | **94.69** | **77.84** | **97.66** | 91.04 | 84.10 | **98.73** |

**Table 5**
Ablation study of proposed modules. The AUC (%) results of Celeb-DF (V2), DFDC and FF++ (C23) datasets are shown.

| Methods | Train set | Test set | | |
|---|---|---|---|---|
| | | Celeb-DF (V2) | DFDC | FF++ (C23) |
| Baseline | FF++ (C23) | 79.45 | 71.71 | 99.42 |
| Baseline + P-FGEM | FF++ (C23) | 80.56 | **74.45** | 99.46 |
| Baseline + F-FGEM | FF++ (C23) | 81.19 | 73.42 | **99.50** |
| Baseline + P-FGEM-S + P-FGEM-M | FF++ (C23) | 81.10 | 74.05 | 99.45 |
| Baseline + F-FGEM-S + F-FGEM-M | FF++ (C23) | 81.48 | 72.40 | 99.41 |
| Baseline + F-FGEM-S + P-FGEM-M | FF++ (C23) | 80.77 | 74.05 | 99.42 |
| Baseline + P-FGEM + F-FGEM (Proposed) | FF++ (C23) | **83.44** | 73.83 | 99.44 |

for shallow features, while P-FGEM-M applies a feature patch-based fine-grained enhancement module to middle-level features. F-FGEM-S is a feature-based fine-grained enhancement module that utilizes global mid- and high-frequency information from shallow features, and F-FGEM-M processes the global mid- and high-frequency information from middle-level features.

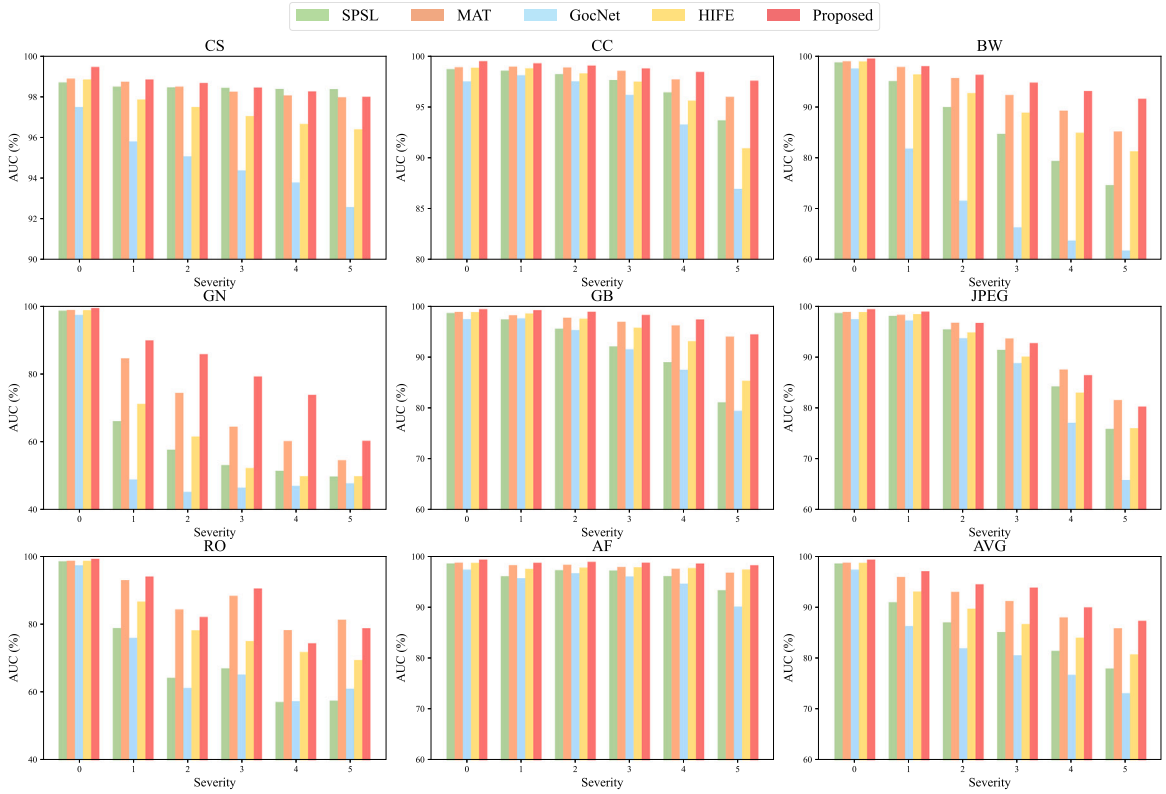



**Fig. 8.** Robustness evaluation for image attacks. We present the AUC scores for various image attacks, including CS, CC, BW, GN, GB, JPEG, RO, and AF across six severity levels. 'AVG' refers to the mean AUC computed across all image attacks at each specified intensity level.

The results, as shown in Table 5, indicate that employing all proposed modules collectively leads to superior performance compared to using them individually. For instance, by adding the P-FGEM to the Convnext model, we achieve an AUC of 80.56% on Celeb-DF (V2) and 74.45% on DFDC, which are improvements of 1.11% and 2.74% AUC points over the baseline Convnext model, respectively. The effectiveness of P-FGEM in utilizing mid- and high-frequency information from shallow local features is demonstrated, resulting in improved detection of tampering clues and enhanced generalization capabilities of the model.

The configuration "Baseline + F-FGEM-S + P-FGEM-M" involves first applying F-FGEM to the global mid- and high-frequency information of shallow features, followed by a P-FGEM for middle-level features. When this construction order is used, the model achieves an AUC of 74.05% on the DFDC dataset, representing a 0.22% improvement over the proposed method. This improvement may result from the extraction of mid- and high-frequency information from feature blocks in middle-level features, enabling the model to better fit the training data and capture features similar to those in the training set. Given the relatively low image quality and high similarity between the DFDC and C23 datasets, this improved fit likely enhances performance. However, on the high image quality Celeb-DF (V2) dataset, "Baseline + F-FGEM-S + P-FGEM-M" achieves an AUC of 80.77%, which is 2.67% lower than our proposed method. Thus, the proposed method demonstrates a superior generalization to capture common forgery features across forgery types, ultimately resulting in better overall detection performance.

### 4.5.2. Effectiveness of different frequency components

To validate the effectiveness of utilizing middle and high frequency coefficients to represent fine-grained information in Deepfake detection, we design feature enhancement modules based on different frequency domain components and incorporate them into the Convnext network. We train the models on the C23 dataset and test them on Celeb-DF (V2) and DFDC. The results are illustrated in the Table 6. The module labeled 'Low' utilizes only low-frequency coefficients to construct the fine-grained enhancement module, 'Middle' employs only middle frequency coefficients, and 'High' uses exclusively high-frequency coefficients for this purpose. Among these configurations, the proposed method, which incorporates middle and high frequency coefficients, achieves an AUC of 87.12% on Celeb-DF (V2), demonstrating superior performance compared to the enhancement modules built with other frequency coefficients. This result underscores that using medium and high frequency coefficients as fine-grained clues to construct feature enhancement modules not only better captures the common characteristics of tampering but also significantly enhances the generalization capability of the model.

**Table 6**
Ablation study of Different Frequency components. The AUC (%) results of Celeb-DF (V2), DFDC and FF++ (C23) datasets are shown.

| Methods | Train set | Test set | | |
|---|---|---|---|---|
| | | Celeb-DF (V2) | DFDC | FF++ (C23) |
| Low | FF++ (C23) | 81.68 | 73.33 | 99.36 |
| Middle | FF++ (C23) | 78.87 | 74.19 | **99.46** |
| High | FF++ (C23) | 80.11 | **74.68** | 99.44 |
| Proposed | FF++ (C23) | **83.44** | 73.83 | 99.44 |

### 4.5.3. Effectiveness of different approaches for feature fusion

To further validate the effectiveness of the proposed feature fusion in enhancement modules, we conduct additional ablation experiments with three additional fusion strategies, namely 'Convnext + Add', 'Convnext + Cat' and 'Convnext + Mu'. In 'Convnext + Add' experiment, we directly add the features together as the input of the next layer. 'Convnext + Cat' means the two features are simply concatenated along the channel dimension and we use a $1 \times 1$ convolution to reduce the

**Table 7**
Ablation study of backbone with proposed modules. The AUC (%) results of Celeb-DF (V2), DFDC and FF++ (C23) datasets are shown.

| Architecture | Train set | Test set | | |
|---|---|---|---|---|
| | | Celeb-DF (V2) | DFDC | FF++ (C23) |
| Resnet50 | FF++ (C23) | 67.14 | 66.08 | **98.83** |
| Resnet50 + FGEM | FF++ (C23) | **69.12** | **67.04** | 98.82 |
| Efficient-B4 | FF++ (C23) | 73.84 | 70.14 | 98.22 |
| Efficient-B4 + FGEM | FF++ (C23) | **77.96** | **72.35** | **99.12** |
| Swin Transform-Base | FF++ (C23) | 78.52 | 72.78 | **99.36** |
| Swin Transform-Base + FGEM | FF++ (C23) | **80.41** | **74.54** | 99.31 |
| Convnext | FF++ (C23) | 79.45 | 71.71 | 99.42 |
| Convnext + FGEM | FF++ (C23) | **83.44** | **73.82** | **99.44** |

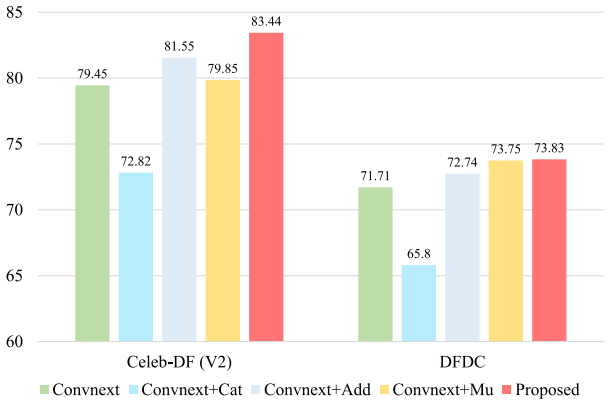



**Fig. 9.** The AUC (%) of using different approaches for feature fusion in enhancement modules.

channel dimension. While for 'Convnext + Mu', we consider Mutual-Enhancement Module in Gu et al. (2022) to enhance and fuse the features. We train on the C23 dataset of FF++ and respectively test on the Celeb-DF (V2) and DFDC datasets.

The Fig. 9 presents the AUC results of different algorithms on these two datasets. From Fig. 9, it is evident that the detection performance decreases when the two features are simply concatenated along the channel dimension. This decline may result from redundancy introduced by the simple channel concatenation, which adversely impacts the model's detection capability. In contrast, a more simple addition fusion ('Convnext + Add') results in 81.55% AUC on Celeb-DF (V2) and 72.74% AUC on DFDC respectively, showing improvement than a more complicated fusion strategy, Mutual-Enhancement Module (Gu et al., 2022) on Celeb-DF (V2) and decrease on the DFDC dataset. However, the designed Fine-Grained Enhancement Module achieves the best AUC of 83.44% and 73.83% on Celeb-DF (V2) and DFDC, respectively. Compared to other approaches for feature fusion, this feature fusion method more effectively mines common detection clues and enhances the model's generalization performance.

#### *4.5.4. Effectiveness of module with different backbone*

To further verify the effectiveness of the proposed module, we conduct comparative experiments using different network backbones, including Resnet50 (He, Zhang, Ren, & Sun, 2016), Efficient-b4 (Tan & Le, 2019), Swin Transformer-Base (Liu, Lin, et al., 2021), and Convnext-Base (Liu et al., 2022) models. 'Resnet50 + FGEM', 'Efficient-b4 + FGEM', 'Swin Transformer-Base + FGEM', and 'Convnext + FGEM' respectively represent the integration of our proposed P-FGEM and F-FGEM into different basic backbones. The models all train on the FF++ (C23) dataset and test on the FF++ (C23), Celeb-DF (V2), and DFDC datasets respectively. Table 7 displays the comparison results. The results clearly show that, the proposed modules significantly enhance the generalization of the models compared to the backbone networks alone.

For example, 'Resnet50 + FGEM' achieves scores of 69.12% on Celeb-DF (V2) and 67.04% on DFDC, which are improvements of 1.98% and 0.96% respectively over the original Resnet50 network on these datasets. Although the proposed models perform slightly below the original baseline model on the C23 dataset, they demonstrate better generalization across different datasets. 'Swin Transformer-Base + FGEM' achieves AUC scores of 80.41% and 74.54% on the Celeb-DF (V2) and DFDC datasets, respectively. Compared to using Swin Transformer-Base alone, this combination improves AUC scores by 1.89% and 1.76% on the two datasets. These results demonstrate that the proposed module not only enhances the generalization performance of models within CNN-based frameworks but also significantly improves the detection generalization capability of models within Transformer-based frameworks. The results in Table 7 confirm that the proposed module effectively prevents the model from overfitting to the training data, enables it to learn more generalized forgery cues, and enhances the overall generalization capability of the model.

### *4.6. Visualization*

To evaluate the performance of our proposed method, we use Grad-CAM (Selvaraju et al., 2017) visualizations to illustrate its effectiveness across various tampering techniques, as shown in Fig. 10. In these heat maps, warm colors indicate the regions most influential in the prediction. Columns (a) and (c) exhibit Grad-CAM images produced by Convnext, whereas columns (b) and (d) display Grad-CAM images generated through the proposed method. For NT (Thies et al., 2019), compared to Convnext, our method shows a better focus on manipulated regions, such as the nose and mouth. These visualizations highlight that the proposed method effectively captures discriminative and relevant features, particularly in NT (Thies et al., 2019), where only the mouth part is manipulated.

## 5. Conclusion

This paper proposes a facial forgery detection framework called MH-FFNet, inspired by fine-grained classification techniques. The framework captures subtle manipulation traces by integrating mid- and high-frequency information. To enhance the accuracy of forgery detection, we design two fine-grained enhancement modules: P-FGEM and F-FGEM. P-FGEM captures mid- and high-frequency information from feature blocks, enhancing forged cues in shallow features and preventing the loss of these details as the network deepens. F-FGEM, on the other hand, captures global mid- and high-frequency information, further enhancing forged cues in intermediate features. Experimental results demonstrate that our method achieves superior detection performance and strong generalization in both in-dataset and cross-dataset. Additionally, the proposed method shows robust resistance to various image-based attacks, ensuring reliable detection under diverse attack scenarios.

Furthermore, with the growing diversity of forged samples, the interplay between different modalities and their features plays a critical role in detection performance. In future work, we aim to investigate

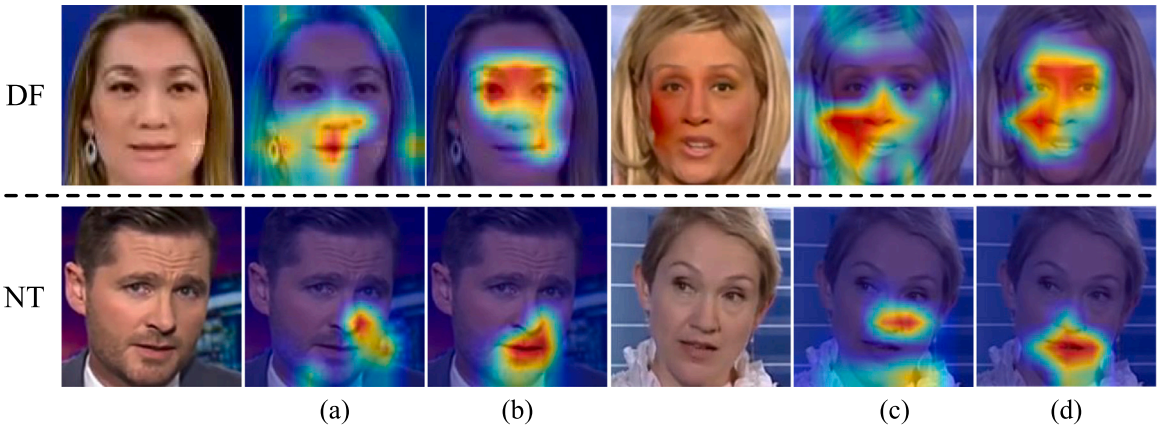



**Fig. 10.** The visualizations of proposed method via Grad-CAM (Selvaraju, Cogswell, Das, Vedantam, Parikh, & Batra, 2017). The images are randomly selected from DF and NT (Thies et al., 2019). And each row includes RGB images and corresponding Grad-CAM.

additional modal features and develop a more systematic and effective feature fusion method. By leveraging the intrinsic connections between different modalities, we hope to further enhance the effectiveness and accuracy of Deepfake detection tasks.

## CRediT authorship contribution statement

**Kai Zhou:** Conceptualization, Data prepare, Visualization, Methodology, Software, Writing – original draft. **Guanglu Sun:** Resources, Supervision, Funding acquisition, Writing – review & editing. **Jun Wang:** Validation, Writing – review & editing. **Linsen Yu:** Writing – review & editing. **Tianlin Li:** Visualization, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This research is partly supported by the Key Research and Development Project of Heilongjiang Province, China (2022ZX01A34), the 2020 Heilongjiang Province Higher Education Teaching Reform Project, China (SJGY 20200320).

## Data availability

The authors do not have permission to share data.

## References

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security* (pp. 1–7). IEEE.

Chandrasegaran, K., Tran, N.-T., & Cheung, N.-M. (2021). A closer look at fourier spectrum discrepancies for cnn-generated images detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7200–7209). IEEE.

Corvi, R., Cozzolino, D., Poggi, G., Nagano, K., & Verdoliva, L. (2023). Intriguing properties of synthetic images: From generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 973–982). IEEE.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., et al. (2020). The deepfake detection challenge (DFDC) dataset. arXiv preprint arXiv:2006.07397.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 9th international conference on learning representations*.

Du, R., Chang, D., Bhunia, A. K., Xie, J., Ma, Z., Song, Y.-Z., et al. (2020). Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *Proceedings of the European conference on computer vision* (pp. 153–168). Springer.

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th international conference on machine learning* (pp. 3247–3258). PMLR.

Fridrich, J., & Kodovsky, J. (2012). Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, *7*, 868–882.

Gao, J., Micheletto, M., Orrù, S., Feng, X., Marcialis, G. L., & Roli, F. (2024). Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection. *Engineering Applications of Artificial Intelligence*, *133*, Article 108450.

Gao, J., Xia, Z., Marcialis, G. L., Dang, C., Dai, J., & Feng, X. (2024). Deepfake detection based on high-frequency enhancement network for highly compressed content. *Expert Systems with Applications*, *249*, Article 123732.

Gu, Q., Chen, S., Yao, T., Chen, Y., Ding, S., & Yi, R. (2022). Exploiting fine-grained face forgery clues via progressive enhancement learning. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 735–743).

Guo, Z., Yang, G., Zhang, D., & Xia, M. (2023). Rethinking gradient operator for exposing ai-enabled face forgeries. *Expert Systems with Applications*, *215*, Article 119361.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 770–778). IEEE.

Hu, T., Qi, H., Huang, Q., & Lu, Y. (2019). See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. arXiv preprint arXiv:1901.09891.

Jeong, Y., Kim, D., Min, S., Joe, S., Gwon, Y., & Choi, J. (2022). Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 48–57). IEEE.

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., et al. (2020). Face X-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5001–5010). IEEE.

Li, Y., & Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656.

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3207–3216). IEEE.

Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., et al. (2021). Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 772–781). IEEE.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022). IEEE.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11976–11986). IEEE.

Liu, K., Perov, I., Gao, D., Chervoniy, N., Zhou, W., & Zhang, W. (2023). Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognition*, *141*, Article 109628.

Liu, J., Xie, J., Wang, Y., & Zha, Z.-J. (2024). Adaptive texture and spectrum clue mining for generalizable face forgery detection. *IEEE Transactions on Information Forensics and Security*, *19*, 1922–1934. http://dx.doi.org/10.1109/TIFS.2023.3344293.

Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE winter applications of computer vision workshops* (pp. 83–92). IEEE.

Miao, C., Tan, Z., Chu, Q., Yu, N., & Guo, G. (2022). Hierarchical frequency-assisted interactive networks for face manipulation detection. *IEEE Transactions on Information Forensics and Security*, *17*, 3008–3021.

Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of the European conference on computer vision* (pp. 86–103). Springer.

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

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 618–626). IEEE.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th international conference on machine learning* (pp. 6105–6114). PMLR.

Thies, J., Zollhöfer, M., & Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, *38*, 1–12.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2387–2395). IEEE.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131–148.

Wang, J., Alamayreh, O., Tondi, B., Costanzo, A., Barni, M., et al. (2022). Detecting deepfake videos in data scarcity conditions by means of video coding features. *APSIPA Transactions on Signal and Information Processing*, *11*.

Wang, T., & Chow, K. P. (2023). Noise based deepfake detection via multi-head relative-interaction. In *Proceedings of the AAAI conference on artificial intelligence: vol. 37*, (pp. 14548–14556).

Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.-G., et al. (2022). M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 international conference on multimedia retrieval* (pp. 615–623).

Wang, B., Wu, X., Wang, F., Zhang, Y., Wei, F., & Song, Z. (2024). Spatial-frequency feature fusion based deepfake detection through knowledge distillation. *Engineering Applications of Artificial Intelligence*, *133*, Article 108341.

Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., & Zhang, Z. (2015). The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 842–850). IEEE.

Yang, X., Li, Y., & Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. In *2019 IEEE international conference on acoustics, speech and signal processing* (pp. 8261–8265). IEEE.

Yang, Z., Luo, T., Wang, D., Hu, Z., Gao, J., & Wang, L. (2018). Learning to navigate for fine-grained classification. In *Proceedings of the European conference on computer vision* (pp. 420–435). Springer.

Zhang, D., Chen, J., Liao, X., Li, F., Chen, J., & Yang, G. (2024). Face forgery detection via multi-feature fusion and local enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*, 8972–8977. http://dx.doi.org/10.1109/TCSVT.2024.3390945.

Zhang, D., Li, D., Sangaiah, A. K., Li, F., Deng, Z., & Wu, C. (2024). Generalizing face forgery detection by suppressed texture network with two-branch convolution. *IEEE Transactions on Computational Social Systems*, 1–9. http://dx.doi.org/10.1109/TCSS.2024.3441251.

Zhang, Y., Yu, Z., Huang, X., Shen, L., & Ren, J. (2024). Genface: A large-scale fine-grained face forgery benchmark and cross appearance-edge learning. arXiv preprint arXiv:2402.02003.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*, 1499–1503.

Zhao, Y., Jin, X., Gao, S., Wu, L., Yao, S., & Jiang, Q. (2023). Tan-gfd: generalizing face forgery detection based on texture information and adaptive noise mining. *Applied Intelligence*, *53*, 19007–19027.

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2185–2194). IEEE.

Zhu, H., Gao, Z., Wang, J., Zhou, Y., & Li, C. (2024). Few-shot fine-grained image classification via multi-frequency neighborhood and double-cross modulation. *IEEE Transactions on Multimedia*, *26*, 10264–10278. http://dx.doi.org/10.1109/TMM.2024.3405713.