Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18751-6

FLAG: frequency-based local and global network for face forgery detection

Kai Zhou[1] · Guanglu Sun[1] · Jun Wang[2] · Jiahui Wang[1] · Linsen Yu[1]

Received: 23 December 2023 / Revised: 4 February 2024 / Accepted: 24 February 2024

Abstract
Deepfake detection aims to mitigate the threat of manipulated content by identifying and exposing forgeries. However, previous methods primarily tend to perform poorly when confronted with cross-dataset scenarios. To address the above issue, we propose an innovative hybrid network called the Frequency-based Local and Global (FLAG) network to explore local and global information with the help of frequency-domain cues for better generalization capability. In consideration of the fact that forged faces often exhibit flaws in the frequency domain, we design a Frequency-based Attention Enhancement Module (FAEM) to enhance the aggregation of CNN and Vision Transformer (ViT). In this design, local features from CNN are attentively enhanced by selected frequency coefficients in FAEM, facilitating generalizable global features learning by the ViT module. The effectiveness of the proposed method is validated via numerous experiments and the generalization performance is improved under cross-dataset scenarios. Especially, the proposed method have obtained an AUC of 99.26% and an ACC of 96.56% using intra-dataset experimental results on FaceForensics++ (C23).

Keywords Multimedia forensics · Deepfake detection · Hybrid network · Vision transformer · Channel attention

1 Introduction

Deepfakes, which refer to manipulated videos by deep neural networks [1, 2], have led to a crisis of social trust and posed a significant threat to social stability [3, 4]. In response to the growing concerns surrounding deepfakes, extensive efforts have been undertaken to distinguish deepfake content from unaltered videos [5–9]. Most existing Deepfake detection methods [10–12] employ Convolutional Neural Networks (CNNs) to extract local

✉ Guanglu Sun
sunguanglu@hrbust.edu.cn

1 School of Computer Science and Technology, Harbin University of Science and Technology, 150080 Harbin, China

2 Department of Information Engineering and Mathematics, University of Siena, 53100 Siena, Italy

information. However, relying solely on local information can make the model more susceptible to being influenced by specific dataset features, resulting in limited generalization performance [13, 14]. Meanwhile, the Vision Transformer (ViT) [15, 16], with its self-attention mechanism, demonstrates strong capabilities in capturing global information. Some researches [7, 17] employ the ViT model for Deepfake detection due to the ability. But ViT always overlook local details that are crucial for Deepfake detection. Due to the highly complementary nature of local and global features, it is essential to integrate the strengths of both CNN and ViT. This enables effective capture of local features while also considering global features, leading to improved overall performance [6, 18]. However, a mere straightforward fusion of model frameworks might result in the algorithm being more tailored to the training data, thus lacking ideal generalization when confronted with unseen datasets. To address this issue, our proposed hybrid network, Frequency-based Local and Global (FLAG) network, which aims to effectively combine CNN and ViT models to achieve improved detection performance not only on the in-dataset but also on unseen datasets.

To enhance the aggregation of CNN and ViT, we propose a Frequency-based Attention Enhancement Module (FAEM), which is specifically designed to improve the model's generalization performance. During the process of face manipulation via deep neural networks, the imperfect generative models introduce visual artifacts in the spatial domain. Several detection methods have been proposed [10, 11] to identify the forged videos based on the visual artifacts in the spatial domain, and significant achievements have been made through the analysis of specific indicators, such as visual color discrepancies [12] and inconsistent head poses [19]. However, these spatial domain algorithms [10, 19, 20] demonstrate fragility when the visual quality of manipulated faces is degraded after common post-image processing attacks (i.e., compression and blur) [21]. To deal with the highly realistic manipulated images, frequency-domain clues have been utilized as generalized features to expose forged videos, by directly utilize frequency domain coefficients as clues or employ them as spatial attention weights for spatial features [22–26]. We argue that simply feeding the network with frequency features makes the network highly reliant on the details, which may disappear during the compression or other process, resulting in a drop in performance. Thus, we propose using channel attention instead of spatial attention to emphasize the discriminability among the channels. Compared to spatial attention, channel attention provides the ability to extract correlations and assess the importance between different channels of feature maps [27, 28].

In this paper, we propose a hybrid network that combines CNN and ViT, connected by FAEM to improve generalization. FAEM is a frequency-based attention enhancement module based on several representative frequency coefficients. More specifically, we select four Discrete Cosine Transform (DCT) coefficients as clues for constructing the channel attention mechanism, which combines Alternating Current (AC) and Direct Current (DC) coefficients. These coefficients provide improved stability and are less susceptible to quality distortion after potential image attacks [29]. Essentially, the amalgamation of AC and DC coefficients is employed to acquire frequency-domain channel attention, subsequently employed to augment the feature weight of mid-level features [10]. To provide a clearer understanding of FAEM and its impact on challenging manipulated faces, we present Gradient-weighted Class Activation Mapping (Grad-CAM) [30] visualizations of the module's effects on Fig. 1 for Deepfakes (DF) [21] and NeuralTextures (NT) [31]. As shown in Fig. 1, our proposed FAEM can effectively focus on manipulated regions, especially in subtle manipulation parts (e.g., face and mouth) of the DF and NT.

The contributions of this paper are summarized as follows.

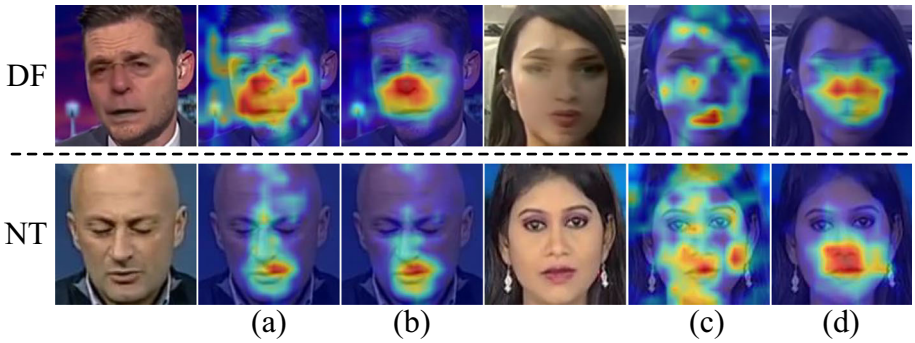

Fig. 1 Grad-CAM [30] visualization of FAEM on two kinds of challenging manipulated faces including DF [21] and NT [31]. (a) and (c) columns reflect features without FAEM, (b) and (d) columns reflect features with FAEM

(1) We propose a novel Deepfake detection network, Frequency-based Local and Global (FLAG) network, which integrates the CNN and ViT through FAEM-enhanced local features to improve generalization performance.
(2) We design a novel Frequency-based Attention Enhancement Module (FAEM) that strengthens the correlations among feature channels and enhances the generalization of method via several frequency coefficients, in particular, three representative AC coefficients and a DC coefficient.
(3) Experimental results demonstrate that the proposed method obtains significant and robustness performance on the in-dataset and effectiveness generalization capability on the cross-datasets. Additionally, the proposed method exhibits strong robustness against various image attacks.

2 Related work

2.1 Deepfake detection

Early methods in the field of forgery detection, such as [32–35], rely on intrinsic statistics or hand-crafted features to model spatial manipulation patterns. Matern et al. [12] propose a method for detecting deepfake videos by leveraging artifacts present in the images, specifically focusing on characteristics such as eye color [8], missing details in the eye and teeth areas. However, with the rapid advancements in deep learning, several studies have focused on developing detectors which based on deep learning that can differentiate manipulated images from real ones by extracting spatial features.

In recent times, there has been a surge in the development of deep learning-based detection methods, which have consistently delivered impressive results. Some Deepfake detection methods based on deep learning with spatial features are proposed. Afchar et al. [10] present a method that utilizes MesoNet for capturing mesoscopic features in the context of deepfake detection. Rössler et al. [11] propose a Deepfake detection method based on XceptionNet, achieving satisfactory results on the FF++ dataset. Li et al. [36] propose a novel spatial image representation called Face-X-ray. Face-X-ray is trained using a self-supervised algorithm on a large dataset consisting of mixed images synthesized from real images. The Face-X-

ray approach can achieve high detection performance in high-quality videos and provide interpretable boundaries for face-swapping. However, it may suffer from a performance drop when encountering low-resolution images. Similarly, Zhao et al. [37] propose a multi-attention detection model to capture subtle forgery traces from spatial features. These spatial-based methods [11, 36–38], however, are fragile when the quality of a manipulated face is degraded by image processing methods. To counter the weakness against quality degradation, our method not only learns spatial artifacts but also builds channel-enhanced attention based on frequency domain coefficients.

In addition, Frank et al. [39] observe that forged images generated by Generative Adversarial Networks (GAN) [2] show particular artifacts in the frequency domain in the essential up-sampling operation, and it has been demonstrated that frequency features possess robust model generalization capabilities for the detection of unseen deepfakes. F3-Net, as described in [24], takes images into the frequency domain and employs two modules to capture global and local frequency cues, respectively. SPSL [40] combines spatial image features and phase spectrum information to effectively capture the up-sampling artifacts commonly found in face forgery images. Kohli et al. [22] convert RGB images into the DCT domain for Deepfake detection. Chen et al. [26] introduce an attention module that is designed for multi-scale feature fusion, aiming to integrate RGB and frequency domain information across various network levels. Luo et al. [25] employ a method that models the correlation and interaction between high-frequency modality and regular modality for detection purposes. In this paper, by considering the benefits of channel attention and the significance of frequency domain information in Deepfake detection tasks, we design a channel attention module that specifically focuses on leveraging frequency domain information.

2.2 Vision transformer

The Transformer has found extensive application in natural language processing (NLP) tasks [41, 42], obtaining impressive performance by effectively modeling long-range dependencies. The ViT [15], as a variant of the Transformer, has been successfully adapted for various computer vision tasks such as object recognition [43, 44], scene classification [45], and face recognition [46]. By dividing an image into a sequence of image patches and leveraging its built-in attention mechanism, ViT excels at capturing global information, thus offering notable advantages in capturing global features.

For Deepfake detection tasks, a number of detection algorithms based on ViT have been proposed, yielding remarkable performance outcomes. In reference [17], high-level convolutional features are extracted using a CNN model. These extracted features are then directly input into ViT for classification purposes. In [18], two CNN models are used to extract feature maps of different sizes. These feature maps are then inputted into a ViT network, which generates two predicted values. The final prediction is obtained by summing these two values. Wang et al. [14] propose a Transformer-based framework that selects more valuable blocks for Deepfake detection by designing the attention module. M2TR [7] designs multi-scale Transformer blocks and frequency-domain features to detect local forgery clues. HFI-Net [6] devises a network structure that combines CNN and ViT, utilizing mid-to-high-frequency information for Deepfake detection. To account for both local and global information in the feature maps, we propose a joint network architecture that incorporates the enhanced local and global features, promoting better model convergence [47].

3 Method

3.1 The proposed frequency-based local and global network

CNN captures local features effectively, leading to superior detection performance in the same manipulation method. However, its limited receptive field hampers its performance on unseen datasets. On the other hand, ViT extracts global features but may overlook subtle clues crucial for Deepfake detection. The combination of local and global features exhibits strong complementarity. We can achieve this by splicing the CNN and ViT networks, significantly improving detection performance on the intra-dataset. However, the simple network splicing approach may be more suitable for the training dataset, limiting its generalization. To address this, we propose a new hybrid network, FLAG network, which utilizes FAEM to aggregate the CNN and ViT models. In this network, local features extracted from CNN are enhanced by FAEM, facilitating generalizable global features learning by the ViT. The FLAG network enables better complementarity between local and global features, resulting in improved detection performance not only on the intra-dataset but also on unseen datasets.

The proposed FLAG network is illustrated in Fig. 2. In this hybrid network, we employ convolutional networks to extract local features from the input image $X \in R^{3\times W\times H}$, resulting in middle level network features $X^m \in R^{C\times M\times N}$ of the CNN architecture. FAEM, which utilizes selected robust and generalizable frequency coefficients, is used to enhance the manipulated information within these local features. Afterwards, the enhanced local feature maps $X^m_{fre} \in R^{C\times M\times N}$ are flattened and their channel size is adjusted using 1×1 convolution to meet the requirements of the ViT's transformer module. Class tokens and position embeddings are added, and the number of transformer blocks is modified accordingly. The

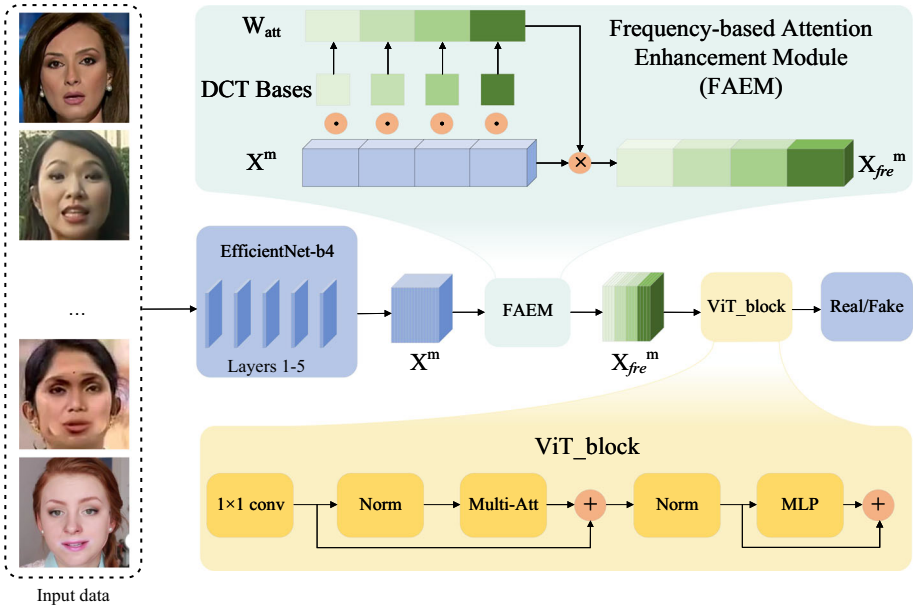

Fig. 2 The overall framework of our proposed method. The extracted middle level features X^m are enhanced with feature weights by the FAEM module, and then enhanced features X^m_{fre} input into the ViT module to obtain global information. $\oplus$ and $\otimes$ denote element-wise sum and channel-wise product

modified features are then fed into the transformer blocks and MLP (Multi-Layer Perceptron) to extract global features for classification. This approach facilitates faster convergence to some extent [47]. As a result, this designed network enables the extraction of both local and global features.

3.2 Frequency-based attention enhancement module

Previous algorithms [25, 26, 40] have shown better generalization by incorporating frequency domain information. However, in these approaches, frequency domain features are often directly extracted and combined with spatial domain features to detect tampering clues. Additionally, they can also be used as spatial attention mechanisms for spatial domain features. In [28], researchers propose a channel attention mechanism based on the frequency domain. They improve the global average pooling method from the perspective of the frequency domain to introduce more channel information and improve the model's performance. However, selecting too many coefficients causes overfitting and decreases generalization performance in Deepfake detection. To address this issue, we consider reducing the number of coefficients and using only relatively robust ones to construct the channel attention module for Deepfake detection. By considering the characteristics of exploring frequency domain coefficients, we select four DCT coefficients, including the DC coefficient and three AC coefficients, to construct FAEM. This aims to improve the generalization and robustness of the detection model.

DCT is a widely used signal processing technique that converts spatial domain information into a frequency domain representation. For a two-dimensional vector $x^{2d} \in R^{M \times N}$, the formula of 2D DCT is as follows:

$$B_{m,n}^{i,j} = \cos\left(\frac{\pi m}{M}\left(i + \frac{1}{2}\right)\right)\cos\left(\frac{\pi n}{N}\left(j + \frac{1}{2}\right)\right), \tag{1}$$

$$F_{m,n}^{2d} = \sum_{i=0}^{M-1}\sum_{j=0}^{N-1} x_{i,j}^{2d} B_{m,n}^{i,j}, \tag{2}$$

where $B_{m,n}^{i,j}$ is the transformation basis function of DCT, M and N represent the length and width of the feature map, i and j represent the position of the feature map ($i = 0, 1..., M$, $j = 0, 1, ...N$), m and n represent the position of the DCT coefficients ($m = 0, 1..., M, n = 0, 1, ..., N$), and $F_{m,n}^{2d}$ is the DCT coefficient.

Specifically, the DC coefficient primarily represents the primary energy of the entire feature map, while the three adjacent AC coefficients represent the horizontal, vertical, and diagonal energy information of the feature map, respectively. These selected 4 DCT coefficients provide stability to the features and are less susceptible to loss in compression. These coefficients are used to construct a frequency-based attention enhancement module, aiming to improve the generalization performance of the detection model. The selected four coefficients represent a significant portion of the energy information in the features. This approach shows better robustness against image attacks like JPEG compression or Gaussian blur compared to using all low-frequency information or other frequency domain information. The effectiveness of this approach is validated through experimental analysis, as presented in Tables 4 and 5.

To provide a clearer understanding, Fig. 3 illustrates the construction of the FAEM using selection coefficients. Meanwhile, Fig. 3 depicts the process of utilizing FAEM to enhance local features. We divide the number of channels of the mid-level network features $X^m \in$

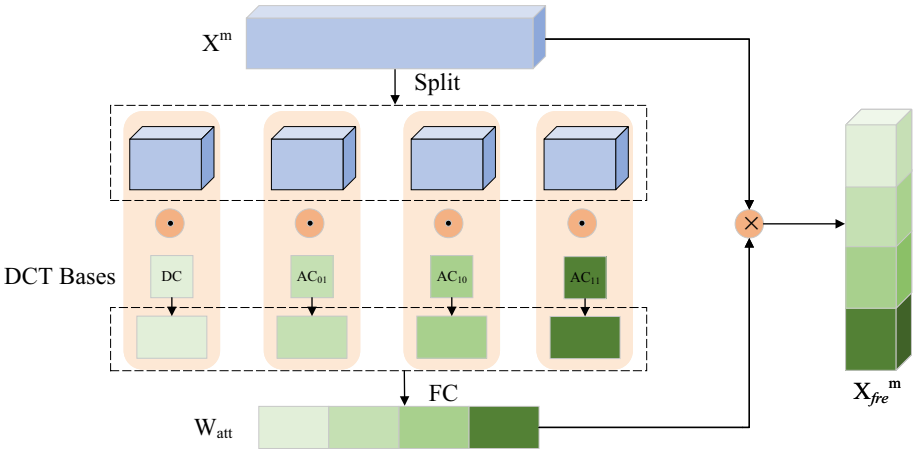

Fig. 3 Proposed Frequency-based Attention Enhance Module (FAEM). $\odot$ and $\otimes$ denote element-wise multiplication and channel-wise product

$R^{C\times M\times N}$ into 4 parts according to the selected 4 frequency domain bases, denoted as $X^k \in R^{C'\times M\times N}$, $k \in \{0, 1, 2, 3\}$, $C' = C/4$. The features of each part correspond to a specific frequency domain base.

For each part, the frequency domain-based attention can be expressed as:

$$F^k = \sum_{i=0}^{M-1}\sum_{j=0}^{N-1} X^k_{:,i,j} B^{m_k,n_k}_{i,j}, \tag{3}$$

where (m_k, n_k) are the frequency component 2D indices corresponding to X^k, $F^k \in R^{C'}$ is the $C^{'}$ dimensional vector. Then the frequency-domain attention features of these 4 parts are aggregated together by $cat()$ function,

$$F_{re} = cat\left(\left[F^0, F^1, F^2, F^3\right]\right), \tag{4}$$

where $F_{re} \in R^c$ is the C dimensional vector. Then, we employ a fully connected (FC) and a sigmoid function σ to obtain attentive weights:

$$W_{att} = \sigma\left(FC\left(F_{re}\right)\right), \tag{5}$$

To mitigate the problem of redundant data resulting from directly applying DCT to RGB information, we adopt an alternative approach. Instead of applying DCT to the RGB data directly, we perform it on the network's middle-level features, denoted as X^m. These middle-level features possess better resistance to interference compared to shallow features and contain more detailed information than high-level features [10]. According to the characteristics of middle-level features, by utilizing X^m, we can preserve crucial tampering clues while minimizing the inclusion of redundant information that may affect recognition. Additionally, we enhance the middle-level features by applying frequency-based attentive weights derived from the AC and DC coefficients,

$$X^m_{fre} = W_{att} \otimes X^m, \tag{6}$$

where X^m_{fre} denotes the enhanced features by frequency-based attentive weights, $\otimes$ denotes a channel-wise multiplication. This process highlights the tampering clues within the feature maps, making them more prominent and improving the overall detection capability.

4 Experiment

4.1 Dataset and settings

Datasets To verify the effectiveness and generalization of our proposed method, we conduct experiments on the FaceForensics++ (FF++) dataset [11], the Celeb-DF (V2) dataset [48] and the DeepFake Detection Challenge (DFDC) dataset [49]. The FF++ dataset comprises three versions: the original version (raw), the lightly compressed version (C23), and the heavily compressed version (C40). Each compressed version consists of 1000 real videos and corresponding fake videos generated using four common manipulation methods, including Deepfakes (DF) [21], Face2Face (F2F) [50], FaceSwap (FS) [51], and NeuralTextures (NT) [31]. Based on reference [11], we use 720 training videos, 140 validation videos, and 140 testing videos for every 1000 videos. For training, we select 32 frames for each video, while for validation and testing, we use 100 frames for each video. The Celeb-DF (V2) dataset [48] comprises 890 real videos and 5639 high-quality fake videos. The DFDC dataset [49] includes more than 20000 real videos and more than 100000 fake videos. In this paper, we employ the Celeb-DF (V2) dataset and DFDC dataset for cross-dataset testing.

Implementation details We utilize MTCNN [52] to detect and save the face image size as 224×224. To extract features, we utilize the EfficientNet-b4 model [53], which has been pretrained on the ImageNet dataset [54], specifically up to layer 5. The middle-level features extracted are subsequently input into the FAEM to acquire enhanced features. Then the enhanced features are fed into the ViT [15] network for classification. We employ AdamW with parameters $(0.9, 0.999)$ as the optimizer. The initial learning rate is to 0.0001 and a weight decay of 1e-5. Training is conducted with a batch size of 14, while testing is performed with a batch size of 4. The total training epoch number is 20. As for data augmentation, we only apply random horizontal flip. And we utilize the cross-entropy loss function for the final binary classification.

We implement the framework and conduct experiments using the open-source PyTorch library on a single NVIDIA 2080Ti GPU. The proposed model has a computational complexity of 33.59 GMAC (Giga Multiply-Accumulates) and consists of 103.48 million parameters. During the testing phase, our algorithm achieves a detection speed of approximately 119 images per second with a batch size of 4.

Evaluation metrics We use Accuracy (ACC) and Area Under the Receiver Operating Characteristic Curve (AUC) for validation. Since our method is essentially image-based, we default to evaluating the model with image-level evaluation following [6, 11].

4.2 Comparison with other methods

In this subsection, we conduct comparative experiments with recent state-of-the-art (SOTA) methods to evaluate their performance in various scenarios. More specifically, MseoNet [10], Xception [11], MaDD [37], MTD-Net [20], and CFFs [55] leverage spatial features for deepfake detection. On the other hand, SPSL [40], M2TR [7], and HFI-Net [6] focus on utilizing

Table 1 Intra-dataset evaluation results (AUC (%) and ACC (%)) on FF++ dataset

Methods	C40		C23	
	AUC	ACC	AUC	ACC
MseoNet [10]	–	70.47	–	83.10
Xception [11]	81.76	80.32	94.86	92.39
SPSL [40]	82.82	81.57	95.32	91.50
M2TR [7]	87.15	83.89	96.75	91.86
HFI-Net [6]	88.40	85.69	97.07	91.87
CFFs [55]	**90.35**	–	97.63	–
Proposed	89.94 (-0.41)	**86.59 (+0.90)**	**99.26 (+1.63)**	**96.56 (+4.69)**

Note that the results for comparisons are from [6]

frequency features to enhance the generalization ability of deepfake detection models. We perform intra-dataset performance tests on FF++ (C23) and FF++ (C40). Cross-dataset evaluations are then conducted using Celeb-DF (V2) and DFDC. Cross-manipulation evaluation in constructed on FF++(23). The best results are shown in bold.

Intra-dataset evaluation The FF++ dataset is commonly used for evaluating deepfake detection methods. We conduct training and testing using the FF++ (C23) and FF++ (C40) settings, respectively. The results presented in Table 1 demonstrate that our method achieves competitive performance compared to previous approaches. In detail, the proposed method works better for the C23 subset, where more frequency details are stored compared to the C40 subset. Specifically, in the C23 subset, our method achieves an ACC of 96.56% and an AUC score of 99.26%, surpassing the second-best ACC performer, CFFs [55], by a notable gain of 1.63% in ACC. Additionally, our proposed method outperforms HFI-Net [6] by a substantial margin, achieving a gain of 4.69% in ACC and 2.19% in AUC. As for the highly compressed C40 subset, our proposed method gains an AUC of 89.94%, which lags CFFs [55] by a margin of 0.41% in AUC. However, in comparison to the HFI-Net [6] method, which also utilizes frequency domain features, our proposed method outperforms HFI-Net [6] by a margin of 0.9% in ACC and 1.54% in AUC.

Cross-dataset evaluation In the cross-dataset evaluation, our model is trained on the FF++ (C23) dataset and tested on the Celeb-DF (V2) and DFDC datasets using the AUC metric. The experimental results, compared with SOTA methods, are presented in Table 2. Notably, our

Table 2 Cross-dataset evaluation (AUC (%)) from FF++ (C23) to Celeb-DF (V2) and DFDC datasets

Methods	Train set	Test set	
		Celeb-DF (V2)	DFDC
Xception [11]	FF++ (C23)	65.30	72.20
M2TR [7]	FF++ (C23)	65.17	–
SPSL [40]	FF++ (C23)	72.39	–
MaDD [37]	FF++ (C23)	67.44	–
MTD-Net [20]	FF++ (C23)	70.12	–
HFI-Net [6]	FF++ (C23)	**83.28**	73.65
CFFs [55]	FF++ (C23)	74.20	72.09
Proposed	FF++ (C23)	78.84 (-4.44)	**75.64 (+1.99)**

proposed approach demonstrates superior generalization on the DFDC dataset, achieving a 1.99% improvement in AUC compared to HFI-Net [6]. For Celeb-DF (V2), our method attains an AUC of 78.84%, surpassing CFFs [55] by a margin of 4.64%. It is worth mentioning that HFI-Net [6] attains the highest performance in Celeb-DF (V2) evaluation by incorporating a global-local interaction module at each stage, effectively suppressing certain features in the training dataset and enhancing generalization across diverse datasets. However, our method outperforms HFI in terms of intra-dataset performance, as indicated in Table 1.

Cross-manipulation evaluation To demonstrate the generalization of our method across different manipulation methods, we conducted this experiment on the FF++ (C23) dataset. Following the standard protocols in [5], we train a model on three manipulation methods from the FF++ dataset and test it on the remaining manipulation method. During training, we use datasets that include three manipulation methods as the training and validation sets, while the remaining manipulation method is exclusively included in the test set. For instance, GID-DF (23) means training on the other three manipulated methods of FF++ (C23) and testing on Deepfakes class, as well as GID-F2F (23). The evaluation metrics used in this study are video level AUC and ACC. Based on references [5, 24], we utilize the average score of a sequence of frames to generate the video-level prediction. The comparative experimental results presented in Table 3 are obtained from [5].

First of all, our method excels in the GID-F2F scenario, achieving a remarkable performance with an ACC of 66.31% and an AUC of 86.50%. This surpasses the second-best ACC performer, LTW [5], by a notable margin, exhibiting a gain of 0.71% in accuracy and 5.3% in AUC. In the case of GID-DF, LTW [5] has a slightly higher AUC by 0.24% compared to our method, although our method achieves a higher ACC by 2.34% than LTW [5]. The advantage of GID-DF is still comparable to SOTA on average.

4.3 Ablation study

Effectiveness of different components To verify the effectiveness of the proposed modules, we conduct several ablation studies. Starting with the pure EfficientNet-b4 model as the baseline, we gradually add the proposed modules. 'EfficientNet-b4+Vit_block' refers to the utilization of the EfficientNet-b4 model with ViT block, excluding the attention module. 'EfficientNet-b4+FAEM' represents the use of the EfficientNet-b4 model with FAEM. These models are trained on the FF++ (C23) and tested on the FF++ (C40) and Celeb-DF (V2) datasets. As the number of proposed modules increases, the proposed model gradually gains

Table 3 Cross-manipulation evaluation results (AUC (%) and ACC (%)) on FF++ (C23) dataset

Methods	GID-DF (C23)		GID-F2F (C23)	
	AUC	ACC	AUC	ACC
EfficientNet [53]	91.11	82.40	80.10	63.32
Forensic [56]	–	72.01	–	64.50
Multi-task [57]	–	70.30	–	58.74
MLDG [58]	91.82	84.21	77.10	63.46
LTW [5]	92.70	**85.60**	80.20	65.60
Proposed	**95.04 (+2.34)**	85.36 (-0.24)	**86.50 (+5.3)**	**66.31 (+0.71)**

The evaluation metrics are video-level AUC and ACC. GID-DF (C23) means training on three manipulation methods of FF++ (C23) and testing on Deepfakes class. GID-F2F (C23) means test on Face2Face class

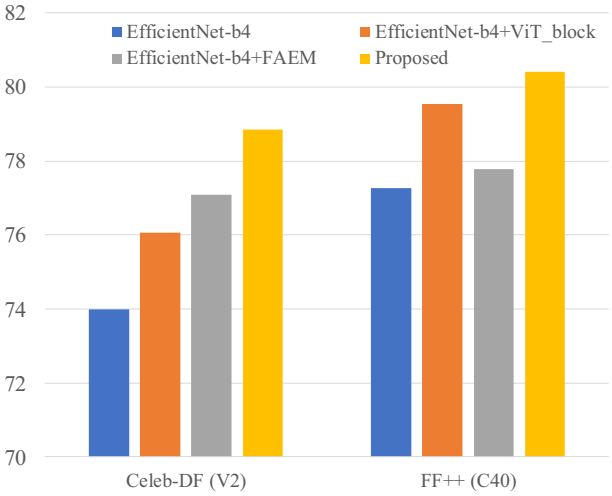

Fig. 4 Ablation results for different components

enhanced expressive capabilities, leading to improved discriminative performance. Figure 4 demonstrates that the hybrid model structure combining CNN and ViT provides more confident information for decision-making compared to the pure EfficientNet-b4 model. The proposed enhanced module, utilizing frequency domain information, also improves the generalization capability of the model. Simultaneously, the proposed network structure, which takes into account both local and global features, has significantly improved the performance compared to the original pure EfficientNet-b4 model. The AUC scores indicate that the proposed algorithm improves intra-dataset performance and generalization across diverse datasets.

Effectiveness of different frequency components To verify the effectiveness of the selected frequency coefficients in our work, ablation studies are made by comparing different choices of frequency coefficients. The results are listed in Table 4, in which 'FLAG_mh' indicates that the FAEM module is constructed using mid-frequency and high-frequency coefficients and 'FLAG_low' uses low-frequency coefficients to construct the FAEM module. Overall, the proposed method obtains the best performance in both scenarios: the in-domain test (train on FF++ 40 and test on FF++ C40), compression robustness test (train on FF++ C23 and test on FF++ C40) and the cross-dataset test (train on FF++ C23 and test on Celeb-DF). More specifically, we see that using low-frequency coefficients shows better performance than using high and middle frequencies, with at least a 0.81% improvement. This improvement is further raised by focusing on four low-frequency coefficients considered in our work.

Table 4 Ablation study on different frequency components via training on FF++ (C23) and C40 respectively

Methods	Train set	Test set		Train set	Test set	
		Celeb-DF (AUC)	C40 (AUC)		C40 (AUC)	C40 (ACC)
FLAG_mh	C23	73.54	78.43	C40	87.73	85.63
FLAG_low	C23	76.31	79.33	C40	89.42	86.44
Proposed	C23	**78.84**	**80.40**	C40	**89.94**	**86.59**

In the real-world situations, images are often affected by image attack, which can result in a reduction in image quality. All these processing will cause a distortion that decreases the generalization performance. In other words, high-frequency features tend to be lost during image processing, further exacerbating the issue. To verify the effectiveness of coefficient selection in the face of image attack, we conduct a robustness test on the channel attention enhancement model constructed using different coefficients. First, we train the model on the C23 training set without image attack, and then we test it on the C23 test set after applying different image attacks [59]. We use various types of image attacks, including: (1) JPEG compression with JPEG quality factors with JPEG quality factors of 50, 30, and 20; (2) filter windows of sizes 7×7, 5×5, and 3×3 for Gaussian Blur; (3) Color Saturation with saturation levels of 0.1, 0.2, and 0.3; (4) Block-wise with a size of 8×8 and varying numbers of occluded blocks: 80, 64, and 48; (5) Color contrast with contrast ratios of 0.6, 0.725, and 0.85. Table 5 presents the different image attacks, their corresponding parameters, and the tested AUC results. Additionally, Fig. 5 provides visual examples illustrating the effects of different image processing techniques and their corresponding levels. Each column of images, from top to bottom, corresponds to different parameter level attacks as listed in Table 5. When facing a JPEG compression attack with a compression factor of 20, the proposed enhancement module, built based on four frequency domain coefficients, improves the AUC by 2% compared to FLAG_low. In the case of Gaussian blur with a filter window size of 7×7, the proposed enhancement module achieves an AUC of 88.05%. Furthermore, when confronted with a color contrast attack featuring a contrast ratio of 0.6, the proposed module shows an AUC improvement of 0.56% compared to the FLAG_low.

Furthermore, we restrict the selection to four coefficients at different positions, as shown in Fig. 6. Figure 6 illustrates the specific selection of locations for the coefficients in the experiment, where (a) represents the four coefficients selected by the proposed method, (b) represents the four middle-frequency coefficients selected, (c) represents another four middle-frequency coefficients selected, and (d) represents the selection of the four high-frequency

Table 5 AUC results of C23 after image processing operation of different types and degrees

Processing	Parameters	FLAG_mh	FLAG_low	Proposed
JPEG	50	91.03	91.54	**91.78**
	30	82.10	84.57	**85.21**
	20	73.47	77.26	**79.26**
Gaussian blur	7	87.59	87.04	**88.05**
	5	95.58	95.28	**96.30**
	3	98.96	98.91	**99.08**
Color saturation	0.1	97.77	98.14	**98.46**
	0.2	97.93	98.28	**98.50**
	0.3	98.19	98.50	**98.65**
Block-wise	80	77.03	85.77	**86.31**
	64	83.72	89.14	**89.53**
	48	90.41	**92.75**	92.67
Color contrast	0.6	98.39	98.29	**98.85**
	0.725	98.95	98.95	**99.18**
	0.85	99.18	99.17	**99.31**

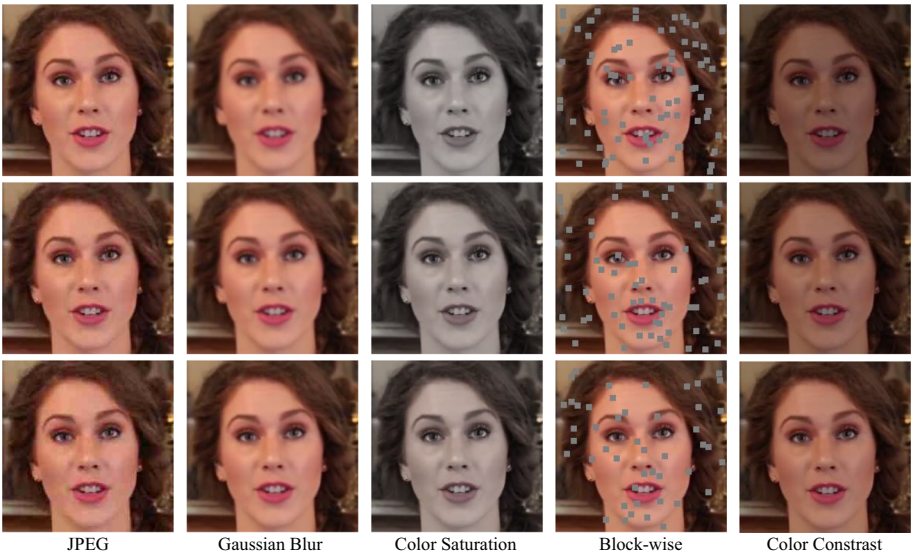

Fig. 5 Image visualization on the levels of severity for five image processing operations. We utilize three severity levels for five distortion types in the robust testing

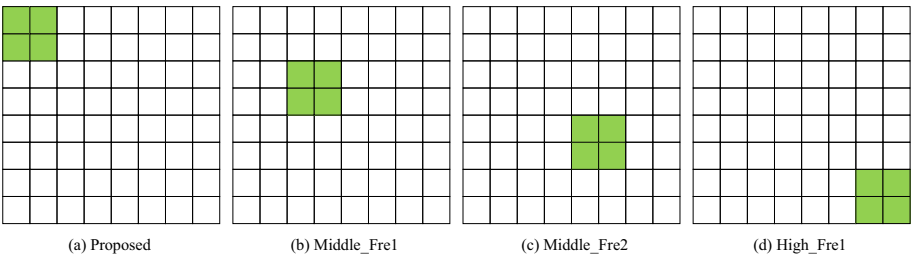

Fig. 6 Four coefficients are chosen at various positions. The picked locations are highlighted in green

Table 6 Ablation study of four coefficients are chosen at various positions

Methods	Train set	Test set	
		Celeb-DF (V2)	DFDC
Middle_Fre1	FF++ (C23)	77.52	72.68
Middle_Fre2	FF++ (C23)	75.27	**76.45**
High_Fre	FF++ (C23)	75.86	74.91
Proposed	FF++ (C23)	**78.84**	75.64

The AUC (%) results of Celeb-DF (V2) and DFDC datasets are shown

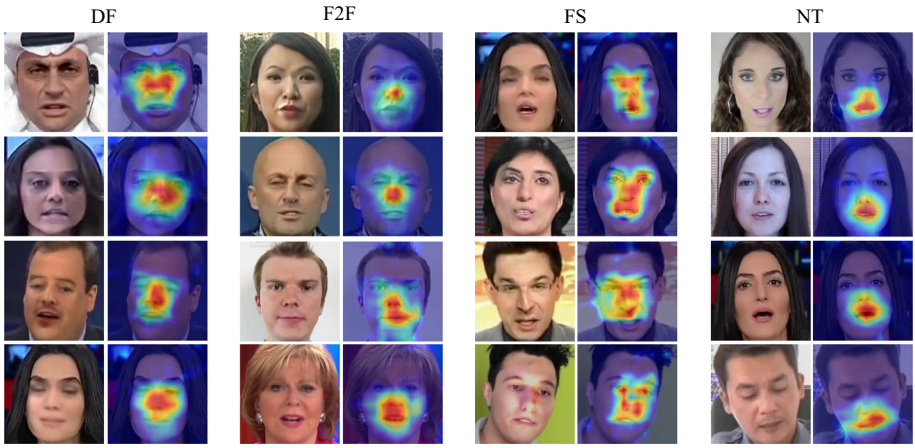

Fig. 7 The visualization experiment of proposed methond through Grad-CAM [30]. The shown images includes DF [21], F2F [50], FS [51] and NT [31], corresponding to each column. And each column includes RGB images and corresponding Grad-CAM [30]

coefficients in the lower right corner. The results are shown in Table 6 for the cross-dataset test. The proposed selection shows the best performance in Celeb-DF (V2) dataset. Regarding the suboptimal performance on the DFDC dataset, it is likely due to the dataset primarily comprising excellent quality forged videos. In such cases, the four frequency domain coefficients represented by Middle_Fre2 may be more effective in detecting and capturing tampering information. More importantly, this study supports our motivation to use four frequency coefficients to design channel attention instead of a random sampling strategy.

4.4 Visualization experiments

To further understand the effectiveness of our proposed method, we provide visualizations of our method through Grad-CAM [30] on different tampering methods in Fig. 7. Two tampering methods, DF [21] and FS [51], are employed for face swapping by replacing the target face with the source face. F2F [50] and NT [31] are facial reenactment technologies that specifically manipulate facial expressions and lip movements. In Fig. 7, we observe that our method focuses on the face regions in the DF [21] and FS [51] columns, while in the F2F [50] and NT [31] columns, our method focuses on manipulation regions such as the nose and mouth. These visualizations demonstrate that our proposed method captures discriminative and reasonable features, especially for NT where only the mouth part is manipulated.

5 Conclusion

This paper introduces the Frequency-based Local and Global (FLAG) network architecture, which effectively explores both local and global information by leveraging frequency-domain cues. By combining the strengths of CNN and ViT, the framework effectively captures tampering information at both local and global scales. Additionally, we propose a frequency-based attention enhancement module that carefully considers the characteristics of frequency domain coefficients. This module effectively integrates the CNN and ViT,

resulting in improved generalization performance of the model. Experimental results on public datasets demonstrate the satisfactory performance of our proposed method. Furthermore, we hope that the FLAG framework can serve as inspiration for researchers to further explore the potential of frequency domain coefficients in the field of Deepfake detection.

Acknowledgements This study is in part supported by the Key Research and Development Project of Heilongjiang Province (2022ZX01A34), the 2020 Heilongjiang Province Higher Education Teaching Reform Project (SJGY 20200320).

Author Contributions Kai Zhou, Guanglu Sun and Jun Wang made substantial contributions to the conception of the work; Kai Zhou and Jiahui Wang drafted the work and made significant contributions to the acquisition, analysis or interpretation of the data; Guanglu Sun, Jun Wang and Linsen Yu revised it critically for important intellectual content.

Funding This study is funded by the Key Research and Development Project of Heilongjiang Province (2022ZX01A34), the 2020 Heilongjiang Province Higher Education Teaching Reform Project (SJGY 20200320).

Availability of data and materials Not applicable.

Code availability Not applicable.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication The Author confirms: that the work described has not been published before; that it is not under consideration for publication elsewhere; that its publication has been approved by all co-authors; that its publication has been approved by the responsible authorities at the institution where the work is carried out.

Competing Interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Pu Y, Gan Z, Henao R, Yuan X, Li C, Stevens A, Carin L (2016) Variational autoencoder for deep learning of images, labels and captions. Advan Neural Inform Process Syst 29
2. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. Advan Neural Inform Process Syst 27
3. Citron DK (2019) How deepfakes undermine truth and threaten democracy. https://www.ted.com
4. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. Inform Fusion 64:131–148
5. Sun K, Liu H, Ye Q, Gao Y, Liu J, Shao L, Ji R (2021) Domain general face forgery detection by learning to weight. Proc AAAI Conf Artif Intell 35:2638–2646
6. Miao C, Tan Z, Chu Q, Yu N, Guo G (2022) Hierarchical frequency-assisted interactive networks for face manipulation detection. IEEE Trans Inf Forensics Secur 17:3008–3021
7. Wang J, Wu Z, Ouyang W, Han X, Chen J, Jiang Y-G, Li S-N (2022) M2TR: multi-modal multi-scale transformers for deepfake detection. In: Proceedings of the 2022 international conference on multimedia retrieval, pp 615–623
8. Wang J, Tondi B, Barni M (2022) An eyes-based Siamese neural network for the detection of GAN-generated face images. Front Signal Process 2:918725
9. 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 Trans Signal Inform Process 11(2)
10. 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 (WIFS), pp 1–7. IEEE

11. 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
12. 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 (WACVW), pp 83–92. IEEE
13. Ni Y, Meng D, Yu C, Quan C, Ren D, Zhao Y (2022) CORE: consistent representation learning for face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12–21
14. Wang P, Liu K, Zhou W, Zhou H, Liu H, Zhang W, Yu N (2022) ADT: anti-deepfake transformer. In: ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2899–1903
15. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929
16. Arkin E, Yadikar N, Xu X, Aysa A, Ubul K (2023) A survey: object detection methods from CNN to transformer. Multimed Tool Appl 82(14):21353–21383
17. Wodajo D, Atnafu S (2021) Deepfake video detection using convolutional vision transformer. arXiv:2102.11126
18. Coccomini DA, Messina N, Gennaro C, Falchi F (2022) Combining efficientnet and vision transformers for video deepfake detection. In: International conference on image analysis and processing, pp 219–229. Springer
19. Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 8261–8265. IEEE
20. Yang J, Li A, Xiao S, Lu W, Gao X (2021) MTD-Net: learning to detect deepfakes images by multi-scale texture difference. IEEE Trans Inf Forensics Secur 16:4234–4245
21. Deepfakes (2022) GitHub. https://github.com/deepfakes/faceswap
22. Kohli A, Gupta A (2021) Detecting deepfake, faceswap and face2face facial forgeries using frequency CNN. Multimed Tool Appl 80:18461–18478
23. Yu Y, Ni R, Li W, Zhao Y (2022) Detection of AI-manipulated fake faces via mining generalized features. ACM Trans Multimed Comput Commun Appl 18(4):1–23
24. Qian Y, Yin G, Sheng L, Chen Z, Shao J (2020) Thinking in frequency: face forgery detection by mining frequency-aware clues. In: European conference on computer vision, pp 86–103. Springer
25. Luo Y, Zhang Y, Yan J, Liu W (2021) Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16317–16326
26. Chen S, Yao T, Chen Y, Ding S, Li J, Ji R (2021) Local relation learning for face forgery detection. Proc AAAI Conf Artif Intell 35:1081–1088
27. Hou Q, Zhou D, Feng J (2021) Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13713–13722
28. Qin Z, Zhang P, Wu F, Li X (2021) FcaNet: frequency channel attention networks. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 783–792
29. Wan W, Wang J, Li J, Meng L, Sun J, Zhang H, Liu J (2020) Pattern complexity-based JND estimation for quantization watermarking. Pattern Recogn Lett 130:157–164
30. Selvaraju RR, 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 international conference on computer vision, pp 618–626
31. Thies J, Zollhöfer M, Nießner M (2019) Deferred neural rendering: image synthesis using neural textures. Acm Trans Graphics (TOG) 38(4):1–12
32. Fridrich J, Kodovsky J (2012) Rich models for steganalysis of digital images. IEEE Trans Inf Forensics Secur 7(3):868–882
33. Carvalho T, Faria FA, Pedrini H, Torres RdS, Rocha A (2015) Illuminant-based transformed spaces for image forensics. IEEE Trans Inform Forensics Secur 11(4):720–733
34. Peng B, Wang W, Dong J, Tan T (2016) Optimized 3D lighting environment estimation for image forgery detection. IEEE Trans Inf Forensics Secur 12(2):479–494
35. Cozzolino D, Poggi G, Verdoliva L (2017) Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: Proceedings of the 5th ACM workshop on information hiding and multimedia security, pp 159–164
36. Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, Guo B (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

37. 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
38. Dong S, Wang J, Liang J, Fan H, Ji R (2022) Explaining deepfake detection by analysing image matching. In: European conference on computer vision, pp 18–35. Springer
39. Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolossa D, Holz T (2020) Leveraging frequency analysis for deep fake image recognition. In: International conference on machine learning, pp 3247–3258. PMLR
40. Liu H, Li X, Zhou W, Chen Y, He Y, Xue H, Zhang W, Yu N (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
41. Tetko IV, Karpov P, Van Deursen R, Godin G (2020) State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. Nat Commun 11(1):5575
42. Khurana D, Koli A, Khatter K, Singh S (2023) Natural language processing: state of the art, current trends and challenges. Multimed Tool Appl 82(3):3713–3744
43. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision, pp 213–229. Springer
44. Li Y, Mao H, Girshick R, He K (2022) Exploring plain vision transformer backbones for object detection. In: European conference on computer vision, pp 280–296. Springer
45. Xu K, Deng P, Huang H (2022) Vision transformer: an excellent teacher for guiding small networks in remote sensing image scene classification. IEEE Trans Geosci Remote Sens 60:1–15
46. Dan J, Liu Y, Xie H, Deng J, Xie H, Xie X, Sun B (2023) TransFace: calibrating transformer training for face recognition from a data-centric perspective. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 20642–20653
47. Xiao T, Singh M, Mintun E, Darrell T, Dollár P, Girshick R (2021) Early convolutions help transformers see better. Adv Neural Inf Process Syst 34:30392–30400
48. Li Y, Yang X, Sun P, Qi H, Lyu S (2020) Celeb-DF: a large-scale challenging dataset for deepfake forensics. In: CVPR, pp 3207–3216
49. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, Ferrer CC (2020) The deepfake detection challenge (DFDC) dataset. arXiv:2006.07397
50. 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 conference on computer vision and pattern recognition, pp 2387–2395
51. Faceswap (2019) GitHub. http://www.github.com/MarekKowalski
52. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503
53. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning, pp 6105–6114. PMLR
54. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on computer vision and pattern recognition, pp 248–255. IEEE
55. Yu P, Fei J, Xia Z, Zhou Z, Weng J (2022) Improving generalization by commonality learning in face forgery detection. IEEE Trans Inf Forensics Secur 17:547–558
56. Cozzolino D, Thies J, Rössler A, Riess C, Nießner M, Verdoliva L (2018) Forensictransfer: weakly-supervised domain adaptation for forgery detection. arXiv:1812.02510
57. Nguyen HH, Fang F, Yamagishi J, Echizen I (2019) Multi-task learning for detecting and segmenting manipulated facial images and videos. In: 2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS), pp 1–8. IEEE
58. Li D, Yang Y, Song Y-Z, Hospedales T (2018) Learning to generalize: meta-learning for domain generalization. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
59. Dong X, Bao J, Chen D, Zhang T, Zhang W, Yu N, Chen D, Wen F, Guo B (2022) Protecting celebrities from deepfake with identity consistency transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9468–9478