LDFnet: Lightweight Dynamic Fusion Network for Face Forgery Detection by Integrating Local Artifacts and Global Texture Information

Zhiqing Guo, Liejun Wang, Wenzhong Yang, Gaobo Yang, and Keqin Li, *Fellow, IEEE*

***Abstract*— Face forgery detection has become a new research hotspot. Though existing detection works have achieved impressive performance, they are difficult to achieve a proper trade-off between detection accuracy and model complexity. To solve this problem, we design some low-complexity modules and construct a lightweight dynamic fusion network (LDFnet) to achieve high accuracy and lightweight face forgery detection. Firstly, we regard significant local visual artifacts as a correct semantic feature needed for detection. A spatial group-wise enhance (SGE) module is introduced as a supervision to suppress possible noise and capture local artifacts. Secondly, we design a manipulation trace extraction block (TraceBlock), which can replace vanilla convolution to achieve global inference, thus capturing the texture information in the global scope. Based on TraceBlock, we construct a global texture representation (GTR) network to extract global manipulation features hierarchically. Finally, we design a dynamic fusion mechanism (DFM) to fully fuse local and global clues, and dynamically generate a more discriminating feature representation. Extensive experimental results show that the proposed LDFnet is significantly superior to the previous detection works on some popular face forgery datasets, such as FF++, DFDC, CelebDF and HFF. In particular, LDFnet only uses 963k model parameters and 801M FLOPs, which is far lower than the calculation cost of face forgery detection based on large model, and achieves better detection results.**

***Index Terms*— Face forgery detection, lightweight model, local artifacts, global texture information, dynamic fusion.**

I. Introduction

WITH the continuous development of computer graphics and generative methods, face forgery techniques have made considerable progress in manipulating multimedia content. While promoting the film industry, these forgery techniques may also be used to seek illegal benefits,[1] such as fabricating fake news, creating political crises or blackmailing someone. Thus, there is an urgent need to develop some countermeasures to neutralize the negative effects of face forgery techniques.

In recent years, with the escalating concerns over face forgeries, researchers in the multimedia forensics community have developed a lot of detection works. In previous studies, some lightweight detection networks have been proposed [1], [2], [3]. These methods focus on some obvious forgery clues, such as visual artifacts [2] and head posture inconsistency [3], and use lightweight convolutional neural networks (CNNs) or machine learning methods to detect face forgery. However, with the rapid development of face forgery technique, it is difficult for lightweight methods to capture the subtle traces left by more advanced forgery technique. Especially in the compressed dataset, the manipulation traces are laundered, which limits the detection of lightweight models. In addition, due to the limited feature representation ability, these lightweight models usually have poor generalization ability in cross-dataset evaluation. Thus, more and more studies have begun to build large models based face forgery detection networks with strong feature representation ability around advanced backbone networks such as ResNet [4], Xception [5], Vision Transformer [6], etc. On the one hand, some studies promote face forgery detection by forcing the backbone network to learn discriminant features directly from local artifact regions [7], [8]. On the other hand, some studies capture the manipulation traces in the global scope by expanding the attention range of the backbone network [9], [10]. Recently, some studies show that both local artifacts and global information are important manipulation clues, which can be used to stimulate the backbone network to achieve high detection accuracy [11]. Although these methods can achieve good detection results, they also bring high model parameters and computational complexity, as shown in Fig 1. This is not conducive to deploying detection algorithms in portable devices. Thus, the above defects motivate us to develop a lightweight detection network, which can achieve high accuracy and generalization while maintaining low model parameters and computational complexity.

For the feature map obtained by vanilla CNNs, due to the lack of supervision of local artifact regions and possible noises in the face image, the spatial distribution of manipulation

Manuscript received 17 April 2023; revised 6 June 2023; accepted 21 June 2023. Date of publication 26 June 2023; date of current version 6 February 2024. This work was supported in part by the Scientific and Technological Innovation 2030 Major Project under Grant 2022ZD0115802, in part by the National Natural Science Foundation of China under Grant 62262065 and Grant U1903213, in part by the Autonomous Region Key Research and Development Task Special under Grant 2022B01008, and in part by the Scientific and Technological Innovation Leading Talent Project under Grant 2022TSYCLJ0037. This article was recommended by Associate Editor Z. Qian. *(Corresponding authors: Liejun Wang; Gaobo Yang.)*

Zhiqing Guo, Liejun Wang, and Wenzhong Yang are with the College of Information Science and Engineering, Xinjiang University, Urumqi 830017, China (e-mail: guozhiqing@xju.edu.cn; wljxju@xju.edu.cn; ywz_xy@163.com).

Gaobo Yang is with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China (e-mail: yanggaobo@hnu.edu.cn).

Keqin Li is with the Department of Computer Science, The State University of New York, New Paltz, New York, NY 12561 USA (e-mail: lik@newpaltz.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3289147.

Digital Object Identifier 10.1109/TCSVT.2023.3289147

[1] https://www.ted.com/talks/danielle_citron_how_deepfakes_undermine_truth_and_threaten_democracy

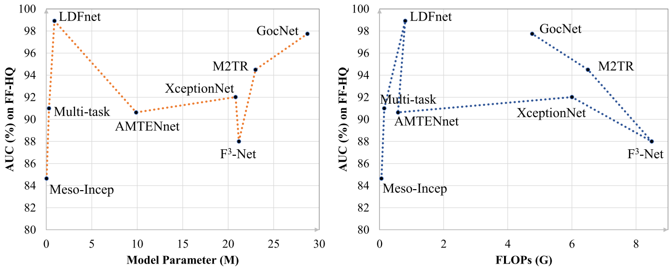


Fig. 1. Relationship between model complexity and detection performance. From this, we can observe that the LDFnet constructed by designing the lightweight mechanism achieves a better trade-off between accuracy and complexity. Compared with the existing detection methods, the proposed LDFnet has obvious advantages in AUC score, model parameters and FLOPs. Details are in Table I.

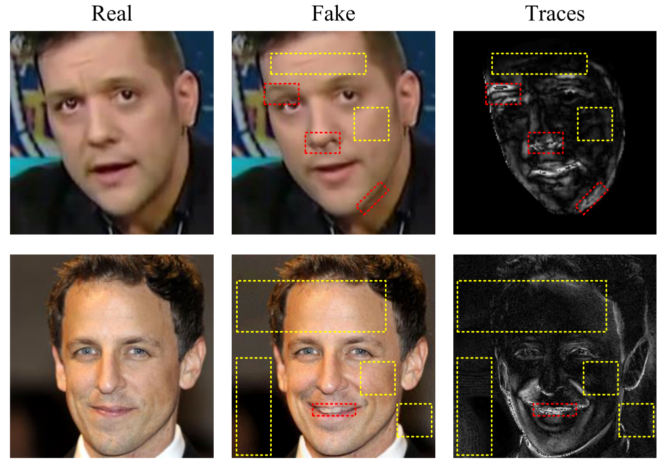


Fig. 2. Distribution of manipulation clues. The first and second columns are real faces and corresponding fake faces, respectively. To clearly show the manipulation traces in the face image, we calculate the absolute difference (closer to white color means greater pixel difference) between the real and fake images in the third column. The red and yellow boxes mark some significant local artifacts and subtle texture changes in the global scope, respectively.

clues will suffer some chaos, which weakens the representation of learned discriminant features. To effectively capture the local manipulation artifacts in the feature map, we introduce a Spatial Group-wise Enhance (SGE) module which hardly needs additional parameters and calculation cost when designing the detection network [12]. Different from some popular attention modules, the SGE module uses the similarity between global statistical features and local statistical features of each location as the generation source of attention masks, and scales feature vectors at all locations to suppress possible noise and highlight correct semantic feature regions. In the face forgery detection, local visual artifacts (see the red box in Fig. 2) are the most significant differences between real and fake face images, which can be regarded as the correct semantic features required for detection. Thus, SGE module can be used to enhance the feature representation of local manipulation artifacts to promote face forgery detection.

The convolution operation in CNNs only extracts information from local neighborhood pixels, which makes it difficult to realize global inference [13]. However, the manipulation traces left by face forgery are usually distributed in the global image (see Fig. 2), especially for the fake face images generated by GANs. To capture global subtle texture changes, it is necessary to improve the locality problem in CNNs. Thus, we design a manipulation trace extraction block (TraceBlock), which can replace the vanilla convolution to capture subtle manipulation traces distributed in the global scope. In addition, we also build a global texture representation (GTR) network. Specifically, we first use the well-known constrained convolution layer to extract the primary manipulation traces [14]. Then, the hierarchical feature extraction module is constructed by stacking TraceBlocks to further learn the global manipulation trace features.

To generate a more discriminating feature representation, we propose a dynamic fusion mechanism (DFM), which includes pre-fusion and post-fusion. Specifically, the local and global features are concatenated by pre-fusion. By adaptively updating the weights, local and global features are convolved dynamically to generate a new feature representation. Then, the attention matrix generated from the global features is projected into the dynamically learned feature space to complete post-fusion. By realizing high-dimensional feature interaction between the extracted local artifacts and global texture information, we enhance the representation of discriminant features. In the existing feature fusion mechanism, one method is to fuse local and global features directly through element-wise sum (see Fig. 3 (c)). Another method is to preserve all feature information for recognition through simple feature concatenation. These methods are only simple addition or concatenation in the feature space, and cannot completely capture the local artifacts and global texture changes. Thus, the superiority of DFM is that it not only considers retaining all forgery information and generating joint feature representation, but also uses the global attention matrix to promote the detection network to pay attention to forgery clues distributed in the global scope.

To achieve efficient face forgery detection, we adopt SGE module, TraceBlock and DFM to build a lightweight dynamic fusion network (LDFnet). Extensive experiments show that LDFnet can achieve better detection results than existing methods on the premise of using only a few parameters and low computational complexity. The main contributions of this paper are summarized as follows:

- We propose a unified end-to-end framework, called LDFnet, which can realize efficient face forgery detection.
- We regard local visual artifacts as a semantic feature, and use SGE module as a supervision to promote the detection network to capture local artifacts.
- We design a novel TraceBlock, which can be used to capture the global texture changes in face forgery images.
- We design a DFM that can fully integrate local and global information to further refine the discriminant features.

The rest of this paper is organized as follows. Section II briefly describes the related works. Section III presents the details of the proposed method. Section IV reports the experimental results and analysis. Conclusion is made in Section V.

II. Related Works

A. Small Model Based Face Forgery Detection

With the development of AI-enabled face forgery technology, face forgery detection has attracted the attention of

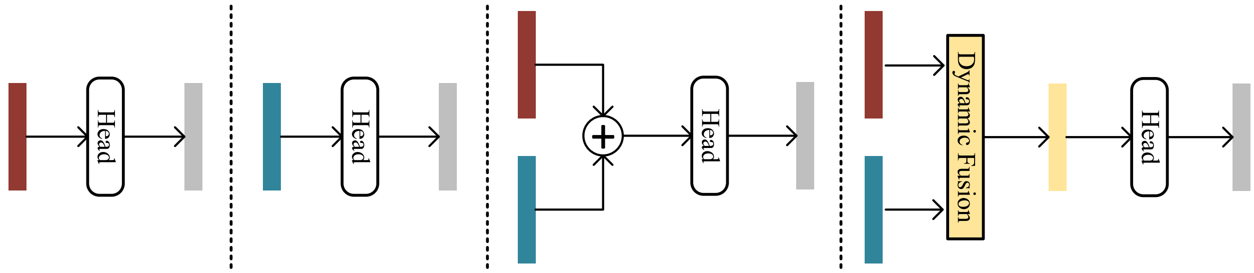


Fig. 3. Compare the proposed work with the existing work. The red and blue rectangles represent local artifacts and global texture information respectively. (a) Prediction using local features. (b) Prediction using global features. (c) Joint prediction using local and global features by element-wise sum. (d) Dynamic fusion of local artifacts and global texture for a joint prediction.

researchers in the multimedia forensics community. In the early work, there was a lot of forensics work based on machine learning or lightweight CNNs. For machine learning based approaches, Yang et al. [15] fed the locations of facial landmark points into support vector machine (SVM) to detect the GAN-synthesized face images. Later, Yang et al. [3] also realized face forgery detection by using head posture inconsistency and SVM. Li et al. [16] constructed a feature set based on chrominance components to capture image statistics for identifying face forgery images. Chen et al. [17] extracted rich spatial and spectral features from face images, and trained them with LightGBM classifier, thus realizing lightweight and efficient face forgery detection. For lightweight CNNs based approaches, Afchar et al. [1] proposed two neural networks with only a few layers, which realized face forgery detection by capturing the mesoscopic properties of faces. Matern et al. [2] trained a small neural network to capture visual artifacts in eyes and teeth. Nguyen et al. [18] proposed a lightweight multi-task learning method, which can simultaneously detect the authenticity of face images and locate the manipulation regions. Sun et al. [19] proposed an efficient detection framework by using temporal modeling of precise geometric features.

Although the above-mentioned small model based face forgery detection works can be easily deployed in portable devices, they are usually difficult to achieve high detection accuracy and generalization due to limited feature representation ability. As we know, machine learning methods usually have poor generalization. In addition, it is also difficult for lightweight CNNs to capture the common features left by different face forgeries, which leads to poor performance in cross-dataset evaluation. Different from the previous work, we design some low-complexity mechanisms to capture the key features of face forgery, thus achieving efficient face forgery detection in the intra-dataset and inter-dataset experimental evaluation.

B. Large Model Based Face Forgery Detection

Recently, a lot of large model based detection works have been proposed. Some works promote the existing backbone networks to capture forgery patterns by designing attention mechanisms [20], [21], [22]. For example, Li et al. [7] assumed that there are blending steps in face swapping, and designed a face X-ray algorithm, which can be trained without any fake face images. Du et al. [20] used pixel-level mask to regularize the local representation in the training process, so as to force the detection model to learn the intrinsic representation from the local forgery region. Miao et al. [22] introduced a self-attention module to force the model to focus on the forgery region, and designed a landmark-guided dropout module to destroy the identity features. Miao et al. [23] also used Central Difference Attention and High-frequency Wavelet Sampler to extract subtle forgery clues in spatial and frequency domains. Wang et al. [24] proposed a localization invariance Siamese network, and used the structural similarity index measurement to construct the groundtruth mask to guide the detection model to pay attention to forged regions. Yang et al. [25] proposed a masked relation learning method, which learned attention features from multiple facial regions, and captured global irregularities by using cross-regional relational information.

There are also some works process the face image through preprocessing mechanisms to obtain some key manipulation clues, which are fed into the backbone network for feature learning [26], [27], [28], [29]. For example, Qian et al. [27] used two complementary clues to deeply mine the forgery patterns. Chen et al. [28] realized robust face forgery detection by using dual-color spatial information and improved Xception network. Hu et al. [29] used the frame-level and temporal-level features to accurately detect the compressed face forgery videos. Zhu et al. [8] disentangled the face image into direct light and identity texture as the key clue to expose face forgery. Furthermore, Zhu et al. [30] proposed a composition search strategy to find useful components and effective architectures to expose forged faces. Chen et al. [31] used occluded face images to train the model in the pre-training stage, and used multi-task learning method to fine-tune the model, which achieved high accuracy detection. Li et al. [32] disentangled artifacts from irrelevant information, and designed a new loss function to provide pixel-level supervision for the generator, thus promoting the backbone network to achieve accurate face forgery detection. Yang et al. [33] constructed the manipulation trace from the perspective of image generation to promote the detection of the backbone network. Li et al. [34] fed different facial image patches into the feature extractor, and mapped the symmetrical facial regions into the angular hyperspace, thus exposing face forgery.

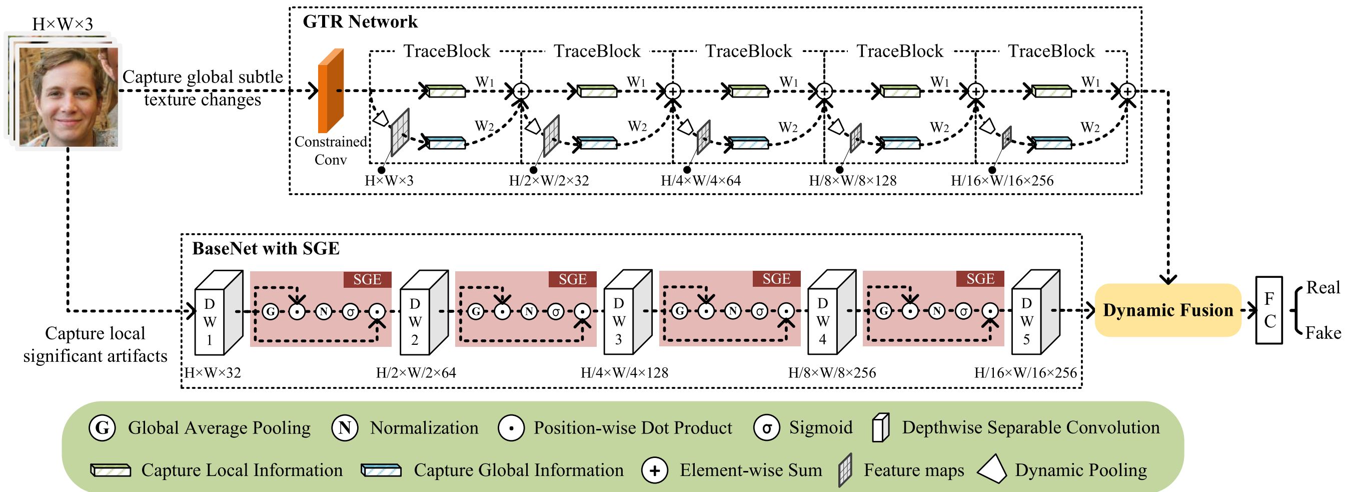


Fig. 4. Overall structure of LDFnet for face forgery detection. LDFnet is a dual-branch network, which captures global subtle texture changes and local significant visual artifacts, respectively. The clues extracted from the two branches are fed into the dynamic fusion mechanism to further optimize the feature representation.

Due to the powerful feature representation ability, these large model based face forgery detection works can achieve high detection accuracy. However, with the increase in detection performance, the improvement of model parameters and computational complexity is unacceptable for practical application scenarios. In our work, we only use very low model parameters and calculation cost while improving the detection performance, which realizes efficient face forgery detection.

III. Methodology

The proposed LDFnet is designed from two directions. That is, capturing local significant artifacts and global subtle texture changes. To efficiently fuse the learned local and global manipulation features, we also design a DFM. Thus, this section will elaborate from three aspects: local artifact representation, global texture representation and dynamic fusion mechanism.

A. Local Artifact Representation

To efficiently model the feature representation of local artifacts in face forgery images, we construct a BaseNet by using five depthwise separable convolution layers, and use the SGE module as supervision to promote the BaseNet to focus on the local artifacts.

In our BaseNet, the depthwise separable convolution is exploited to generate a set of feature maps $T \in \mathbb{R}^{C\times H\times W}$, where C and $H \times W$ represent the channel number and spatial dimension of the feature map, respectively. Then, we use SGE module to divide T into N groups along the channel dimension. For a group feature map X, each position in the feature space has a vector representation, that is, $X = \{x_1, x_2, \ldots, x_m\}$, $m = H \times W$. Since the inevitable noise and similar patterns in the feature space, it is difficult for vanilla convolution layers to obtain well-distributed feature responses. Thus, the SGE module first uses the spatial average function $F_{gp}(\cdot)$ to obtain the overall information of entire group space:

$$g = F_{gp}(X) = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{1}$$

Then, using the dot product between the global feature g and the local feature x_i, the corresponding coefficient c_i can be generated for each feature in X.

$$c_i = g \cdot x_i \tag{2}$$

Coefficient c_i measures the similarity between global feature g and local feature x_i to some extent. To obtain the enhanced feature vector $\hat{x}_i$, the normalized coefficient $\hat{c}_i$ is first passed through sigmoid function $\sigma(\cdot)$. Then, the original feature x_i is scaled by dot products to enhance the learning of semantic features in local regions, as follows:

$$\hat{x}_i = x_i \cdot \sigma(\hat{c}_i) \tag{3}$$

The SGE module was originally designed to enhance the representation of semantic features in feature space [12]. Since significant visual artifacts can also be regarded as important semantic features for face forgery detection, the SGE module can be used to enhance the feature representation of local visual artifacts. In our design, after each depthwise separable convolution, an SGE module is embedded to promote BaseNet to pay attention to local manipulation clues (see Fig. 4).

B. Global Texture Representation

Let x_j be a pixel in the input image. Thus, the pixel y_i in the output image calculated by CNNs can be expressed as follows:

$$y_i = \sum_{j\subseteq N(i)} x_j * w_{ij} \tag{4}$$

where $i, j = \{1, 2, 3, \ldots, H \times W\}$, and w_{ij} is the learnable weight between the i-th pixel and the j-th pixel. In addition, j is the index of some possible pixels related to the i-th pixel. Note that the bias is omitted to simplify the equation. For CNNs, j-th pixel belongs to the neighborhood $N(i)$ of the i-th pixel. For example, $N(i)$ in 3×3 convolution contains a set of eight neighboring pixels except the i-th pixel itself. Due to the limited receptive field, the local inference mode of CNNs is difficult to capture the manipulation clues in the global scope.

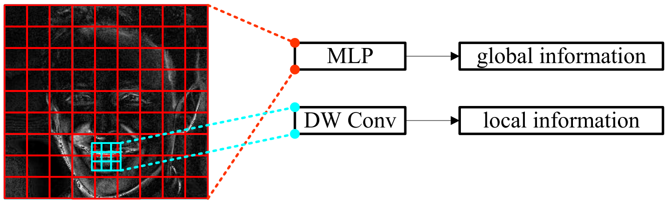


Fig. 5. A diagram that captures global and local manipulation traces.

To realize global inference, we design a TraceBlock to improve the defects of CNNs. Specifically, TraceBlock contains two branches (see Fig. 4), which capture local information and global information respectively. In the first branch, a layer of depthwise separable convolution is used to capture local information. In the second branch, we introduce a Multi-Layer Perceptron (MLP), which allows more pixels to interact with the given i-th pixel, to perform global inference. On the one hand, the size of feature map is variable in the process of hierarchical feature extraction. Thus, the size of global information extracted by MLP needs to be consistent with local information, so as to realize feature fusion in a TraceBlock. On the other hand, manipulation traces are usually presented in the form of gray scale changes. Thus, the max pooling is a suitable operation to capture these traces. In the second branch of TraceBlock, we use max pooling function $F_{max}(\cdot)$ to adaptively reduce the size of x_j according to the stride of convolution kernel in the first branch, which makes the global information consistent with the extracted local information. Thus, the output y_i of TraceBlock can be expressed as:

$$y_i = W_1 \sum_{j \subseteq N(i)} x_j * w_{ij} + W_2 \sum_{\forall j} F_{max}(x_j) * u_{ij} \tag{5}$$

where w_{ij} and u_{ij} are both learnable weights between the i-th pixel and the j-th pixel. But the difference is that the j-th pixel in u_{ij} belongs to all pixels in the feature map, not just the local neighborhood, as shown in Fig. 5. The design of TraceBlock is very flexible, we assign parameters $W_1, W_2 \in [0, 1]$ to local information and global information learned from x_j, respectively. In the process of back propagation, the parameters $\{W_1, W_2\}$ are updated to adaptively fuse local&global features and switch different inference modes. For example, when $W_1 = 0$ and $W_2 = 1$, TraceBlock degrades to MLP, and when $W_1 = 1$ and $W_2 = 0$, TraceBlock degrades to CNNs that only capture local information.

Based on TraceBlock, a GTR Network is constructed. Firstly, we introduce a well-known manipulation trace extraction module, namely the constrained convolution layer [14], which is placed at the front end of the GTR Network to extract low-level forgery clues. Then, we feed the low-level trace features into a hierarchical feature extraction network composed of five TraceBlocks to learn the high-level global texture representation features. The diagram of GTR Network is shown in Fig. 4.

C. Dynamic Fusion Mechanism

Different from the traditional fusion method, that is, the features are combined in a fixed way in the feature space.

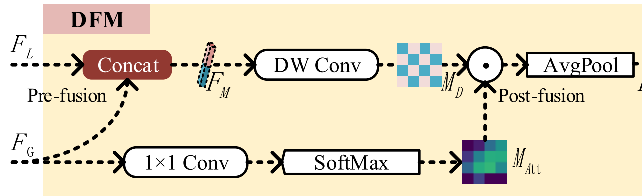


Fig. 6. Detailed structure of DFM. Where F_L and F_G represent local features and global features, respectively. And F_D represents the discriminant feature after dynamic fusion.

We feed two types of features into the proposed DFM (see Fig. 6), which can dynamically generate and refine the final feature representation via back propagation.

In the pre-fusion, we first dynamically generate the joint feature representation of local artifacts and global texture information. Given the local feature $F_L \in \mathbb{R}^{C\times H\times W}$ and the global feature $F_G \in \mathbb{R}^{C\times H\times W}$, they are first concatenated together along the channel direction to form the mixed feature $F_M \in \mathbb{R}^{2C\times H\times W}$, as follows:

$$F_M = Concat(F_L, F_G) \tag{6}$$

F_M contains all the manipulation features learned from the backbone network. To extract the high-dimensional representation of mixed features from F_M, we use a layer of depthwise separable convolution to dynamically generate feature map $M_D \in \mathbb{R}^{C\times H\times W}$ in the process of back propagation.

In the post-fusion, we use the global information to dynamically refine the feature map M_D. As subtle texture changes are widely distributed in the global feature space, we use 1×1 convolution for cross-channel interaction and information integration at each spatial position of F_G, and use SoftMax function to generate attention matrix $M_{Att} \in \mathbb{R}^{C\times H\times W}$. The above process can be expressed as:

$$M_{Att} = SoftMax(\sum_{i=1}^{N} F_G * \omega) \tag{7}$$

where ω represents 1×1 convolution kernel, and N is equal to $H \times W$, which denotes all feature space positions. Finally, the M_{Att} is projected onto the M_D by a point-wise dot product to further refine M_D. In addition, the refined features are fed into the spatial average function $F_{gp}(\cdot)$, and then the dynamic fusion features $F_D \in \mathbb{R}^{C\times 1\times 1}$ are generated by layer normalization $LN(\cdot)$ and activation function $ReLU(\cdot)$, as follows:

$$F_D = ReLU(LN(F_{gp}(M_D \cdot M_{Att}))) \tag{8}$$

IV. Experiments

A. Datasets

In our experiments, four challenging face forgery datasets, such as FF++ [35], DFDC [36], CelebDF [37] and HFF [38], are selected for experimental evaluation. Since the first three datasets are all video datasets, we need to perform some necessary preprocessing steps for them. Specifically, we first extract the face region from the video frame by using the Face Recognition Library.[2] Then, all extracted face images are

[2] https://github.com/ageitgey/face_recognition

resized to the size of 256 × 256 for training and evaluation. To avoid the similarity between consecutive frames, we extract N_{frames} face images from each video sequence at interval N_{iter}. In addition, considering the difference of the number of frames in the video sequence, the specific details performed on different datasets are as follows:

FF++ [35] contains 1,363 real video sequences and 4,000 fake video sequences generated by four typical face forgery methods, including FaceSwap [40], DeepFake,[3] Face2Face [41] and NeuralTextures [42]. For real video sequences, the $N_{iter} = 2$ and $N_{frames} = 50$ are used to obtain the real face images. For fake video sequences, the face images are extracted from each video with $N_{iter} = 2$ and $N_{frames} = 16$. As a result, the number of both real and fake face images is more than 60k. To balance the data, we randomly select 60k real faces and 60k fake faces from the extracted images for experiments. The ratio of training set to evaluation set is 5:1. In addition, FF++ contains three versions: raw dataset, compressed high-quality dataset (FF-HQ) and compressed low-quality dataset (FF-LQ). We adopted two more challenging FF-HQ and FF-LQ datasets in the experiment.

DFDC [36] is a large face forgery dataset containing more than 100k face video sequences. In our experiment, 2,891 real face videos and 20,210 fake face videos are randomly selected for the experiment. In real videos, parameters N_{iter} and N_{frames} are set to 2 and 35, respectively. For fake videos, parameters $N_{iter} = 10$ and $N_{frames} = 5$. As a result, we also randomly select 60k real faces and 60k fake faces from the extracted images to balance the data, and divide the training set and the evaluation set at the ratio of 5:1.

CelebDF [37] contains 890 real face videos and 5,639 fake face videos. For real videos, parameters $N_{iter} = 2$ and $N_{frames} = 100$. For fake videos, parameters $N_{iter} = 2$ and $N_{frames} = 16$. CelebDF is a dataset with high visual quality, which is usually used for cross-dataset evaluation. Therefore, we randomly select 120k face images (real: 60k, fake: 60k) from the extracted images as the evaluation set to verify the generalization of the detection model.

HFF [38] is an image dataset composed of a variety of generative face forgeries. In our experiment, a total of 155k face images are divided into training set and evaluation set at the ratio of 4:1. All face images are also resized to 256 × 256 for experiments.

B. Experimental Settings

1) Evaluation Metrics: Face forgery detection is essentially a binary classification task, so the accuracy rate (ACC) and the area under the receiver operating characteristic curve (AUC) are used to evaluate all detection models in the experiment. In addition, LDFnet is a lightweight detection model. To better demonstrate the superiority of LDFnet, we also provide model parameters (Param.) and floating point operations (FLOPs) in the paper. Since the FLOPs only measures the theoretical computational complexity of the model, the inference speed in real scenarios will also be affected by hardware equipment and optimization algorithms. Thus, we also provide the actual inference speed of the detection model for reference.

2) Implementation Details: Our experiments are implemented under the PyTorch framework. In the training stage, we set the random seed to 7 to ensure that the initialization parameters of the model are consistent in any experiment. In addition, the detection model is trained for 20 epoches in each group of experiments, and the batch size is set to 64. We use Adam optimizer with default parameters for training, and the initial learning rate is 10^{-3}. After each training epoch, the learning rate decays to half.

C. Comparison With Face Forgery Detection Works

1) Intra-Dataset Evaluation: To make a fair comparison, we select seven representative detection works with open source codes, such as Meso-Incep [1], Multi-task [18], XceptionNet [35], F^3-Net [27], AMTENnet [38], M2TR [10] and GocNet [39], for experimental evaluation. All detection networks are trained from scratch on four datasets (FF-HQ [35], FF-LQ [35], DFDC [36] and HFF [38]) and evaluated with the same testing set.

Table I reports the comparison results of LDFnet and existing detection methods on four datasets. Among them, the HFF dataset mainly contains face forgery images generated by GANs. The generative models will leave forgery clues in the global face image. These forgery clues have not been laundered by image compression or other operations. Thus, the existing detection method, even for lightweight models (i.e., Meso-Incep, Multi-task and AMTENnet), can capture the anomalies between real and fake face images and achieve high ACC and AUC scores. For three video datasets, most manipulation traces in compressed video frames are usually laundered, which makes it difficult to detect. Generally, the large model (i.e., XceptionNet, F^3-Net, M2TR and GocNet) can effectively mine the identification features by using complex model structure and mechanism, and obtain better detection results than the lightweight model. Especially, M2TR achieves competitive detection results on four datasets by combining CNNs and multi-scale transformer.[4] Compared with M2TR, GocNet uses tensor preprocessing module and manipulation trace attention module to further improve the detection performance of backbone network while keeping relatively low FLOPs. Although these works have achieved good detection results, their model parameters and calculation costs are still large. In our work, LDFnet enhances the local representation of manipulation features and captures the global manipulation trace features by using SGE module and GTR network, respectively. In addition, LDFnet optimizes the features learned from the two-branch network by using DFM. Thus, by paying attention to important clues and realizing efficient dynamic fusion, LDFnet can obtain very competitive face forgery detection results with very little model parameters and calculation cost (Parameters: 963K & FLOPs: 801M). In the actual model inference, we can observe that LDFnet can

[3] https://github.com/deepfakes/faceswap

[4] Since it is difficult to calculate the model parameters and FLOPs of M2TR, we estimate the minimum value according to the backbone network (XceptionNet + Transformer) for reference.

TABLE I
COMPARISON RESULTS (%) OF FACE FORGERY DETECTION METHODS FOR INTRA-DATASET EVALUATION

Methods	Year	Publication	Param.	FLOPs (Mac)	Inference speed (frames/second, FPS)	FF-HQ		FF-LQ		DFDC		HFF	
						ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Meso-Incep [1]	2018	WIFS	28.52K	60.18M	1087	74.58	84.65	73.53	77.98	76.91	86.01	96.19	97.41
Multi-task [18]	2019	BTAS	305K	146M	1201	81.44	91.00	77.50	86.51	80.05	87.89	96.30	99.53
XceptionNet [35]	2019	ICCV	20.81M	6.00G	185	86.72	92.02	83.53	89.61	89.83	94.86	98.64	99.73
F^3-Net [27]	2020	ECCV	21.17M	8.49G	153	80.02	88.00	75.21	83.00	79.86	87.50	89.65	95.71
AMTENnet [38]	2021	CVIU	9.88M	575M	201	89.25	90.63	85.09	87.64	89.81	92.64	99.87	99.91
M2TR [10]	2022	ICMR	>20.81M	>6.00G	139	86.88	94.50	82.31	92.10	91.27	97.20	91.14	97.00
GocNet [39]	2023	ESWA	28.71M	4.76G	135	94.34	97.75	89.46	94.94	92.52	96.87	99.90	99.98
LDFnet	-	-	963K	801M	236	**96.01**	**98.92**	**92.32**	**96.79**	**93.15**	**97.20**	**99.94**	**99.99**

TABLE II
COMPARISON RESULTS (%) OF FACE FORGERY DETECTION METHODS FOR CROSS-DATASET EVALUATION

Methods		Backbone	Year	Publication	Training Dataset	CelebDF Dataset [37]
Big Models	Two-stream [43]	GoogLeNet	2017	CVPRW	Private dataset	53.8
	FWA [44]	ResNet50	2019	CVPRW	UADFV dataset	56.9
	Xception-LQ [35]	XceptionNet	2019	ICCV	FF++ dataset	65.5
	Capsule [45]	Capsule Network	2019	ICASSP	Private dataset	57.5
	DSP-FWA [44]	ResNet50	2019	CVPRW	UADFV dataset	64.6
	SMIL [46]	XceptionNet	2020	MM	FF++ dataset	56.3
	F^3-Net [27]	XceptionNet	2020	ECCV	FF++ dataset	65.2
	MADD [47]	EfficientNet-b4	2021	CVPR	FF++ dataset	67.4
	LTW [48]	EfficientNet-b0	2021	AAAI	FF++ dataset	64.1
	M2TR [10]	XceptionNet + Transformer	2022	ICMR	FF++ dataset	65.7
	GocNet [39]	ResNet18	2023	ESWA	FF++ dataset	**67.4**
Small Models	Meso-4 [1]	Lightweight CNN	2018	WIFS	Private dataset	54.8
	MesoInception-4 [1]	Lightweight CNN	2018	WIFS	Private dataset	53.6
	HeadPose [3]	SVM	2019	ICASSP	UADFV dataset	54.6
	VA-MLP [2]	Lightweight CNN	2019	WACVW	Private dataset	55.0
	VA-LogReg [2]	Lightweight CNN	2019	WACVW	Private dataset	55.1
	Multi-task [18]	Lightweight CNN	2019	BTAS	FF++ dataset	54.3
	LRNet [19]	Lightweight RNN	2021	CVPR	FF++ dataset	56.9
	DefakeHop++ [17]	LightGBM	2022	arXiv	FF++ dataset	60.5
	LDFnet	Lightweight CNN	-	-	FF++ dataset	**65.7**

surpass the existing large model detection methods at the speed of 236 frames per second, which realizes real-time inference.

2) Cross-Dataset Evaluation: To verify the generalization of detection model in cross-dataset evaluation, we introduce a widely used evaluation benchmark, namely CelebDF [37]. This benchmark contains many large model based detection methods, such as MADD [47], LTW [48], and M2TR [10]. It also contains many small model based detection methods, such as VA-LogReg [2] and Multi-task [18]. To reduce the evaluation error on CelebDF benchmark as much as possible, we extracted 120k face images from all videos in CelebDF dataset. As we know, large models can usually rely on complex structures and powerful feature representation ability to learn rich common features, thus achieving better generalization. However, it is often difficult to achieve good generalization for small models with limited feature representation ability. Thus, we divide the detection methods in CelebDF benchmark according to model parameters and FLOPs,[5] so as to compare generalization more fairly.

[5]We define the detection model with parameters below 2M and FLOPs below 1G as a small model, and vice versa.

Table II reports the cross-dataset evaluation results of LDFnet and 19 comparison methods. Among them, most detection methods use FF++ dataset as training set. Thus, LDFnet is consistent with most methods, training on FF++ dataset and testing on CelebDF dataset. From Table II, we can observe that large model detection methods generally achieve good generalization, while small model detection methods often have poor cross-dataset evaluation results. However, LDFnet still has good generalization while maintaining its lightweight, which is mainly due to the fact that LDFnet not only captures local significant artifacts and global subtle texture changes, but also optimizes the extracted two types of features via DFM, thus capturing rich common features.

D. Ablation Study of LDFnet

LDFnet includes three important components: SGE module, GTR network and DFM. To verify the effectiveness of these modules, we conduct ablation experiments on three challenging datasets, such as FF-LQ, FF-HQ and DFDC.

Table III reports the quantitative results of the detection network under different configurations. From it, we can observe that the performance of the detection network is

TABLE III
QUANTITATIVE RESULTS (%) OF ABLATION STUDIES ON LDFNET

Network	BaseNet	SGE Module	GTR Network	DFM	FF-HQ [35]		FF-LQ [35]		DFDC [36]	
					ACC	AUC	ACC	AUC	ACC	AUC
LDFnet	✓				82.87	89.11	79.72	86.20	80.43	86.75
	✓	✓			84.34	91.50	81.17	87.07	82.07	88.08
	✓		✓		87.25	93.82	83.31	89.70	82.88	88.86
	✓	✓	✓		87.81	94.18	83.36	90.13	83.72	90.04
	✓	✓	✓	✓	**96.01**	**98.92**	**92.32**	**96.79**	**93.15**	**97.20**

improved when only SGE module or GTR network is assembled. Among them, the SGE module can promote the backbone network to capture local artifacts by suppressing possible noise and highlighting correct artifact regions with almost no additional parameters and calculation costs, thus improving the detection accuracy. The GTR network can effectively improve the locality problem in CNNs by adaptively switching different inference modes, so as to realize the global inference of manipulation clues. Although the detection performance can be improved by capturing local artifacts or global texture changes, these results are far from satisfactory. To further optimize the feature representation, the common way is to use local and global features to make joint prediction by element-wise sum. However, we find that even if SGE module and GTR network are assembled at the same time, the improvement of detection performance is still limited (see the fourth row of Table III). The reason behind this may be that the local and global features cannot be fully integrated by pixel-by-pixel addition. Thus, we introduce DFM to adaptively generate the optimal feature representation in a dynamic fusion way, thus further boosting the detection performance of lightweight networks.

E. Design of DFM

To further evaluate the design of DFM, we modify some structures and configurations of DFM, and name the modified versions as A, B, C and D, as shown in Fig. 7. To verify the importance of concatenated features in pre-fusion, we first remove the Concat operation between local features and global features in DFM_A. Then, the attention matrix generated from global features is removed in DFM_B to verify its effectiveness. Finally, we replace 'DW Conv' and 'AvgPool' components with '1×1 Conv' and 'MaxPool' respectively, as shown in DFM_C and DFM_D in Fig. 7. These ablation experiments are also conducted on FF-LQ, FF-HQ and DFDC datasets.

Table IV reports the quantitative results of five versions of DFM. From this, we can observe that it is important to splice local features and global features along the channel direction. The combined features can generate a more discriminating feature representation through convolution operation. Then, we compare the differences between 3×3 convolution and 1×1 convolution in extracting information from the combined features. We note that 3×3 convolution is obviously more suitable for capturing the correlation between manipulation

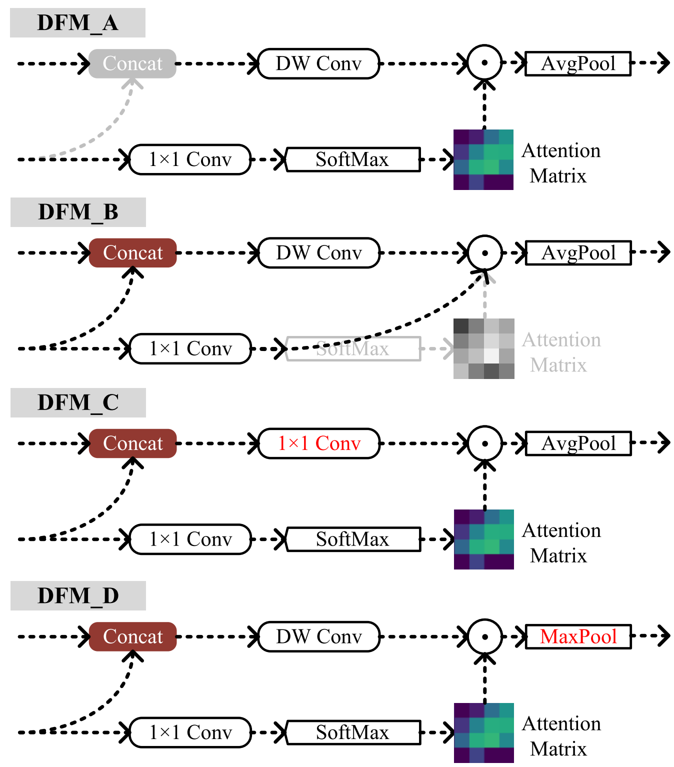


Fig. 7. Comparison of DFM with different structures. The removed components are indicated by gray lines, and the replaced components are marked by red fonts.

traces in spatial dimension and channel dimension. However, 1×1 convolution can only achieve cross-channel information integration at different pixel positions, which makes it impossible to generate a more suitable joint feature representation from local and global features. In addition, it can be observed that it is also meaningful to generate attention matrix from global features. The feature representation can be refined in a point-wise dot product way by using the attention matrix. Finally, we compare AvgPool with MaxPool, and verify that AvgPool can better optimize feature representation.

F. Visualization Result

As we know, Class Activation Map (CAM) [49] and Gradient-weighted Class Activation Mapping (Graded-CAM) [50] are two commonly used visualization methods, which visualize the attention of the model to different regions of the input image by generating a heat map. However, the number and distribution position of manipulation artifacts left

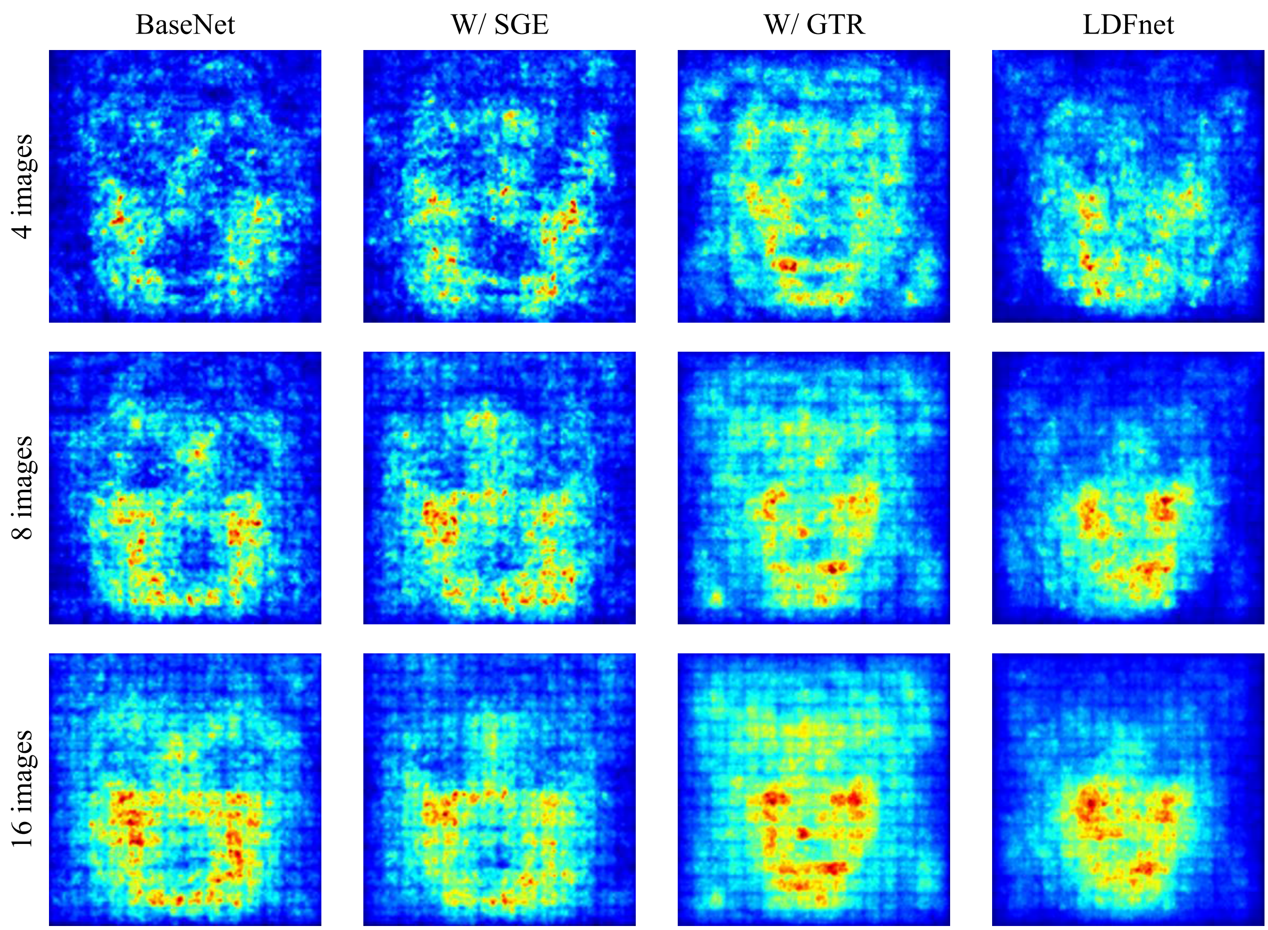


Fig. 8. Visualization of the attention regions. 'W/ SGE' stands for BaseNet equipped with SGE module, which is used to verify whether SGE module can promote the detection network to pay attention to local artifact regions. 'W/ GTR' denotes BaseNet equipped with GTR network to verify whether the detection network captures the global subtle texture changes.

TABLE IV
QUANTITATIVE RESULTS (%) OF DFM WITH DIFFERENT STRUCTURES

Dynamic Fusion	FF-HQ		FF-LQ		DFDC	
	ACC	AUC	ACC	AUC	ACC	AUC
A	94.04	98.19	88.12	94.15	91.20	96.29
B	95.77	98.51	90.28	95.52	91.96	96.67
C	91.72	96.95	85.56	91.43	87.11	93.28
D	93.56	97.77	88.67	94.01	89.58	94.95
Ours	**96.01**	**98.92**	**92.32**	**96.79**	**93.15**	**97.20**

by face forgery technology are often different. Visualizing a single image only by CAM or Grad-CAM will have some deviation. Thus, we adopt the Average Forgery Attention Maps (AFAMs) [9] for visualization to reveal the attention changes of the detection networks under different configurations. These AFAMs are calculated on 4, 8 and 16 fake face images respectively, so as to more accurately visualize the attention region of the detection network.

Fig. 8 shows the visualization results. Among them, from the first to third rows are AFAMs calculated on different numbers of fake face images. The first column is the visual result of BaseNet. It can be observed that when the SGE module is embedded, the detection network generally pays more attention to the local artifact region. In addition, we can observe from the third column that GTR network can effectively help the detection network to capture the subtle texture changes in the global scope. However, only paying attention to all the clues in the global scope can not fully improve the detection performance in the intra-dataset and cross-dataset evaluation. For the task of face forgery detection, intra-dataset evaluation with high accuracy needs to extract more discriminating features, while cross-dataset evaluation with high generalization needs to learn more common features. From the visualization results of LDFnet, LDFnet can focus on the important facial regions containing common features and more discriminating manipulation clues by dynamically aggregating local and global features, thus achieving better intra-dataset and cross-dataset evaluation results. These results reveal the intrinsic reason why LDFnet can achieve better face forgery detection while maintaining lightweight from a visual perspective.

G. Limitations

Even though the research shows that LDFnet can achieve lightweight and high-performance face forgery detection with the support of SGE, GTR and DFM, there are still some limitations. For cross-dataset evaluation, although LDFnet has obvious advantages compared with small model methods, it is still difficult to obtain competitive detection results compared with large model methods. LDFnet is still insufficient in capturing the common features left by different forgery technologies. In the future research, it is still necessary to design around the common features in face forgery to promote the generalization performance of the detection network. In addition, we should also consider deploying the network in portable devices to evaluate the energy consumption of the network in the real environment.

V. Conclusion

In this work, a novel LDFnet is proposed, which dynamically aggregates local artifact features and global texture information, for face forgery detection. Its main advantage is to achieve accurate detection while maintaining lightweight architecture. Firstly, we introduce SGE module into the detection network, which hardly needs additional parameters and calculation cost, to supervise the feature learning process of local artifacts. Secondly, we design a TraceBlock by improving the locality problem in CNNs, and build a GTR network based on constrained convolution layer and TraceBlocks to capture the subtle texture changes in the global scope. Finally, to fully fuse the extracted local features and global features, we design a lightweight dynamic fusion mechanism. The extensive experiments confirm that the proposed LDFnet has achieved promising results. This research seeks to encourage future work to realize face forgery detection in the direction of lightweight and strong generalization, so as to better realize the landing application of detection technology.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.

[2] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose Deepfakes and face manipulations," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2019, pp. 83–92.

[3] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8261–8265.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[5] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[6] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–22.

[7] L. Li et al., "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5000–5009.

[8] X. Zhu, H. Wang, H. Fei, Z. Lei, and S. Z. Li, "Face forgery detection by 3D decomposition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2928–2938.

[9] C. Wang and W. Deng, "Representative forgery mining for fake face detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14918–14927.

[10] J. Wang et al., "M2TR: Multi-modal multi-scale transformers for Deepfake detection," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2022, pp. 615–623.

[11] Z. Guo, G. Yang, D. Wang, and D. Zhang, "A data augmentation framework by mining structured features for fake face image detection," *Comput. Vis. Image Understand.*, vol. 226, Jan. 2023, Art. no. 103587.

[12] X. Li, X. Hu, and J. Yang, "Spatial group-wise enhance: Improving semantic feature learning in convolutional networks," 2019, *arXiv:1905.09646*.

[13] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[14] B. Bayar and M. C. Stamm, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2691–2706, Nov. 2018.

[15] X. Yang, Y. Li, H. Qi, and S. Lyu, "Exposing GAN-synthesized faces using landmark locations," in *Proc. ACM Workshop Inf. Hiding Multimedia Secur.*, Jul. 2019, pp. 113–118.

[16] H. Li, B. Li, S. Tan, and J. Huang, "Identification of deep network generated images using disparities in color components," *Signal Process.*, vol. 174, Sep. 2020, Art. no. 107616.

[17] H.-S. Chen, S. Hu, S. You, and C.-C. Jay Kuo, "DefakeHop++: An enhanced lightweight Deepfake detector," 2022, *arXiv:2205.00211*.

[18] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *Proc. IEEE 10th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2019, pp. 1–8.

[19] Z. Sun, Y. Han, Z. Hua, N. Ruan, and W. Jia, "Improving the efficiency and robustness of Deepfakes detection through precise geometric features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3608–3617.

[20] M. Du, S. Pentyala, Y. Li, and X. Hu, "Towards generalizable Deepfake detection with locality-aware AutoEncoder," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manag.*, Oct. 2020, pp. 325–334.

[21] Z. Guo, G. Yang, J. Chen, and X. Sun, "Exposing Deepfake face forgeries with guided residuals," *IEEE Trans. Multimedia*, early access, Jan. 18, 2023, doi: 10.1109/TMM.2023.3237169.

[22] C. Miao et al., "Learning forgery region-aware and ID-independent features for face manipulation detection," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 1, pp. 71–84, Jan. 2022.

[23] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, and N. Yu, "F^2Trans: High-frequency fine-grained transformer for face forgery detection," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1039–1051, 2023.

[24] J. Wang, Y. Sun, and J. Tang, "LiSiam: Localization invariance Siamese network for Deepfake detection," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2425–2436, 2022.

[25] Z. Yang, J. Liang, Y. Xu, X. Zhang, and R. He, "Masked relation learning for Deepfake detection," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1696–1708, 2023.

[26] G. Pang, B. Zhang, Z. Teng, Z. Qi, and J. Fan, "MRE-Net: Multi-rate excitation network for Deepfake video detection," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jan. 25, 2023, doi: 10.1109/TCSVT.2023.3239607.

[27] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 86–103.

[28] B. Chen, X. Liu, Y. Zheng, G. Zhao, and Y. Shi, "A robust GAN-generated face detection method based on dual-color spaces and an improved Xception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3527–3538, Jun. 2022.

[29] J. Hu, X. Liao, W. Wang, and Z. Qin, "Detecting compressed Deepfake videos in social networks using frame-temporality two-stream convolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1089–1102, Mar. 2022.

[30] X. Zhu et al., "Face forgery detection by 3D decomposition and composition search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8342–8357, Jul. 2023.

[31] H. Chen, Y. Lin, B. Li, and S. Tan, "Learning features of intra-consistency and inter-diversity: Keys toward generalizable Deepfake detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1468–1480, Mar. 2023.

[32] X. Li, R. Ni, P. Yang, Z. Fu, and Y. Zhao, "Artifacts-disentangled adversarial learning for Deepfake detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1658–1670, Apr. 2023.

[33] J. Yang, S. Xiao, A. Li, W. Lu, X. Gao, and Y. Li, "MSTA-Net: Forgery detection by generating manipulation trace based on multi-scale self-texture attention," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4854–4866, Jul. 2022.

[34] G. Li, X. Zhao, and Y. Cao, "Forensic symmetry for Deepfakes," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1095–1110, 2023.

[35] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[36] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The Deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.

[37] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for Deepfake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3204–3213.

[38] Z. Guo, G. Yang, J. Chen, and X. Sun, "Fake face detection via adaptive manipulation traces extraction network," *Comput. Vis. Image Understand.*, vol. 204, Mar. 2021, Art. no. 103170.

[39] Z. Guo, G. Yang, D. Zhang, and M. Xia, "Rethinking gradient operator for exposing AI-enabled face forgeries," *Exp. Syst. Appl.*, vol. 215, Apr. 2023, Art. no. 119361.

[40] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3697–3705.

[41] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2387–2395.

[42] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, Aug. 2019.

[43] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1831–1839.

[44] Y. Li and S. Lyu, "Exposing Deepfake videos by detecting face warping artifacts," in *Proc. CVPR Workshops*, 2019, pp. 1–7.

[45] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2307–2311.

[46] X. Li et al., "Sharp multiple instance learning for Deepfake video detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1864–1872.

[47] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional Deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2185–2194.

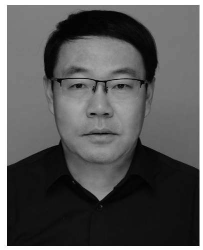
[48] K. Sun et al., "Domain general face forgery detection by learning to weight," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 2638–2646.

[49] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[50] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

Zhiqing Guo received the Ph.D. degree in computer science and technology from Hunan University in 2022. Currently, he has been introduced to Xinjiang University as an A-Level Young Talent, where he engages in teaching and scientific research. He has published research articles in international journals, such as IEEE TRANSACTIONS ON MULTIMEDIA, *ACM Transactions on Multimedia Computing Communications and Applications*, *Expert Systems with Applications*, and *Computer Vision and Image Understanding*. His current research interests include digital media forensics, computer vision, and deep learning.

Liejun Wang received the Ph.D. degree from the School of Information and Communication Engineering, Xi'an Jiaotong University, in 2012. He is currently a Full Professor with the School of Information Science and Engineering, Xinjiang University. He has published research articles in international journals, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON CYBERNETICS, and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. His current research interests include wireless sensor networks, computer vision, and natural language processing.



Wenzhong Yang received the Ph.D. degree from the School of Computer Science, Wuhan University, in 2011. He is currently a Full Professor with the School of Information Science and Engineering, Xinjiang University. He is the PI of several projects, such as the Natural Science Foundation of China (NSFC) and the Leading Talent Project of Autonomous Region. His current research interests include cyberspace security, wireless networks, and machine learning.

Gaobo Yang received the Ph.D. degree in communication and information systems from Shanghai University in 2004. He is currently a Professor with Hunan University, China. He made an academic visit to the University of Surrey, U.K., from August 2010 to August 2011. He is the PI of several projects, such as the Natural Science Foundation of China (NSFC), the Special Pro-phase Project on the National Basic Research Program of China (973), and the Program for New Century Excellent Talents (NCET) in university. He has published research articles in international journals, such as IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON BROADCASTING, and *ACM Transactions on Multimedia Computing, Communications, and Applications*. His current research interests include image and video signal processing and digital media forensics.

Keqin Li (Fellow, IEEE) is currently a SUNY Distinguished Professor of computer science with The State University of New York and a National Distinguished Professor with Hunan University, China. He has authored or coauthored more than 810 journal articles, book chapters, and refereed conference papers. He holds more than 60 patents announced or authorized by the Chinese National Intellectual Property Administration. His current research interests include cloud computing, fog computing and mobile edge computing, energy-efficient computing and communication, embedded systems and cyber-physical systems, heterogeneous computing systems, big data computing, high-performance computing, CPU-GPU hybrid and cooperative computing, computer architectures and systems, computer networking, machine learning, and intelligent and soft computing. He was a recipient of several best paper awards. He is among the world's top ten most influential scientists in distributed computing based on a composite indicator of Scopus citation database. He has chaired many international conferences. He is currently an Associate Editor of *ACM Computing Surveys* and *CCF Transactions on High Performance Computing*. He was on the editorial boards of IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON CLOUD COMPUTING, the IEEE TRANSACTIONS ON SERVICES COMPUTING, and IEEE TRANSACTIONS ON SUSTAINABLE COMPUTING.