# Learnable Local Similarity for Face Forgery Detection and Localization

Lingyun Leng[1], Jianwei Fei[2], Yunshu Dai[3]
College of Cyber Security, Jinan University, Guangzhou, 510632, Guangdong, China[1]
School of Computer, Nanjing University of Information Science & Technology, Nanjing, 210044, Jiangsu, China[2]
School of Cyber Science Technology, Sun Yat-sen University, Shenzhen, 518107, Guangdong, China[3]

***Abstract*—The emergence of many face forgery technologies has led to the widespread of forgery faces on the Internet, causing a series of serious social impacts, thus face forgery detection technology has attracted increasing attention. While many face forgery detection algorithms have demonstrated impressive performance against known manipulation methods, their efficacy tends to diminish severely when applied to unknown forgeries. Previous research commonly viewed face forgery detection as a binary classification problem, disregarding the crucial distinction between real and forged faces, thereby limiting the generalizability of detection algorithms. To overcome this issue, this paper proposes a novel face forgery detection method that utilizes a trainable metric to learn local similarity between local features of facial images, achieving a more generalized detection result. What's more, it incorporate cross-level features to accurately locate forgery regions. After conducting extensive experiments on FaceForensics++, Celeb-DF-v2, and DFD, which demonstrate that the effectiveness of the proposed method is comparable to state-of-the-art detection algorithms.**

***Keywords*—Face forgery detection; local similarity; forgery localization; generalized detection**

## I. Introduction

With the advancements in computer vision and deep learning technology, face forgery technologies have become a growing concern for society. These technologies have matured significantly in recent years, producing counterfeit faces that are indistinguishable to the human eye [1], [2], [3], [4]. Criminal misuse of these technologies can pose severe societal consequences, including pornographic featuring public figures, the spread of political misinformation, and fraudulence that jeopardizes personal and property rights. In response to this challenge, many researchers have developed studies on face forgery detection approaches [5], [6], [7]. Though these approaches perform well in specific scenarios, they are often inadequate when it comes to detecting forgeries unseen in the training data, known as the generalization problem.

Early research considered face forgery detection a binary classification task, employing convolutional neural networks (CNN) to distinguish real and forged faces. While such approaches have shown impressive performances in in-domain settings where training data and testing data are forged by the same algorithm, they cannot be easily applied to unknown domains given that the unknown forgery algorithms have different forgery features. Consequently, some face forgery detection approaches have been proposed to mine generalizable artifact traces [8], [9], [10]. These approaches rely on identifying discrepancies in image features such as brightness, color, and texture to distinguish forged faces. However, they can be affected by the quality of the images and thus may not be suitable for real-world scenarios. To address this issue, emerging approaches centered around data augmentation are being employed [11], [12]. By using various forgery algorithms, these techniques aim to expand the training data, to improve generalization capability. However, it's worth noting that training data depending on forgery detection approaches may become ineffective in light of the continuous advancements in forgery algorithms. On the other hand, some researchers concentrate on the identification details of forged faces, discover identity discrepancies to identify forged faces, and have achieved remarkable success with face replacement [13], [14]. Nevertheless, this type of method cannot detect face images with unchanged identity information. As face forgery algorithms continue to advance, the forgery traces have become increasingly subtle and difficult to detect using previous approaches. Thus, some researchers start paying attention to identifying fine-grained local feature inconsistencies [15], [16], [17].

In this paper, we focus on face forgery detection approach that does not rely on data augmentation but extracts essential features of real and forged faces. Our solution is based on the observation that real faces typically exhibit evenly distributed features and local region similarities [15], whereas forged faces usually exhibit local abnormalities resulting from the blending between real and forged regions. Inspired by this, we propose a novel generalized forgery face detection approach. Our approach improves generalization by leveraging auxiliary constraints between local features of real and forged faces. Unlike existing methods that rely on metrics such as cosine similarity [17], we utilize a learnable network to measure the similarities between local features, which makes the similarity measurement better suited to aligning features extracted by CNN backbones. We also introduce a cross-level forgery localization module that integrates various features across different levels with a lightweight attention module. Our approach enables accurate forgery localization and forged face recognition with strong generalizability. In brief, our contributions are summarized as follows:

- We propose to learn the dense fine-grained similarity between real and forged local features, which greatly improves the generalization of face forgery detection approach.
- We propose a Y-shaped network that achieved accurate face forgery detection and localization with the proposed cross-level attentional feature fusion module.

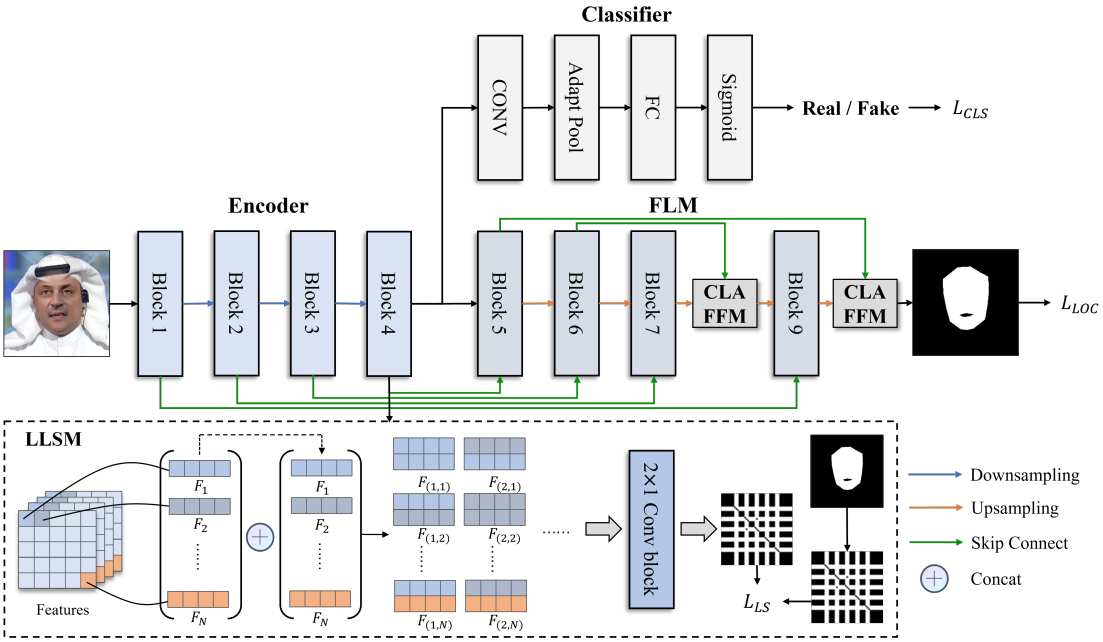


Fig. 1. The overall architecture of the proposed approach.

- Extensive experiments validate the superior detection performance and impressive generalization ability of our approach.

## II. Related Work

### A. Face Forgery Algorithms

Face forgery technology has advanced quickly in recent years. It can be categorized into three categories: face swap, face editing, and face generation, based on the forgery objects. Early face swap algorithms are mostly accomplished using graphics techniques [18], which are complex and challenging. With the rapid development of deep learning, several novel face swap algorithms have emerged, significantly reducing the difficulty of face swapping [3], [4]. The advance of Generative Adversarial Networks (GAN) has further improved the realism of the forged faces [1], [2].

### B. Face Forgery Detection Algorithms

Early face forgery algorithms were not very mature, resulting in flawed faces with obvious artifacts. Therefore, early detection algorithms depend mostly on CNN to catch these probable artificial traces, such as color artifacts [8], blink rate [19], head position [14], synthetic artifacts [9], and so on. However, with the advance of face forgery algorithms, these problems have largely been resolved, forcing researchers to develop new detection methods capable of identifying generic forgery clues.

On the one hand, the researchers discovered that forgery faces have invariable forgery patterns in the frequency domain, giving rise to frequency-aware forgery detection approaches. Qian *et al.* [20] proposed a two-stream collaborative learning framework that leverages frequency-aware image decomposition and local frequency statistics to extract forgery patterns. Meanwhile, Chen *et al.* [21] tackled face forgery detection through local relational learning, integrating RGB and frequency information using an attention module to improve generalization. On the other hand, temporal inconsistencies of forged videos have become a significant clue for forgery detection. Time-aware models that extract temporal features from numerous single-frame inputs have been introduced to detect the authenticity of face videos. Güera *et al.* [22] used Long Short-Term Memory (LSTM) to extract temporal features from numerous single-frame. Gu *et al.* [23] detected local dynamic inconsistencies induced by tiny movements in forged videos. In addition, approaches based on advanced semantic information are proposed. Haliassos *et al.* [24] suggested detecting forgery videos by analyzing differences in lip movements between real and forgery videos, while Dong *et al.* [13] developed an identity consistency transformer to safeguard celebrities by detecting identity inconsistencies between inner and outer faces. However, approaches based on common forgery clues may not be suitable for real-world scenarios with unknown forgery clues, and approaches based on temporal inconsistency may be limited by video quality.

Due to the flaws in splicing, blending, and editing procedures present in the majority of available face forgery algorithms, forged faces may contain features from multiple sources, leading to local inconsistencies. Shang *et al.* [15] proposed to capture pixel-level and region-level differences for face forgery detection. Zhao *et al.* [16] proposed pairwise self-consistency learning of local features, which achieved excellent performance in generalization. Sun *et al.* [25] enhanced the generalization by creating positive and negative sample pairings and contrastively learning real and forged regions. The studies mentioned above highlight the importance of local inconsistency in improving the generalization of face forgery detection. They mainly use fixed metric measures when calcu-

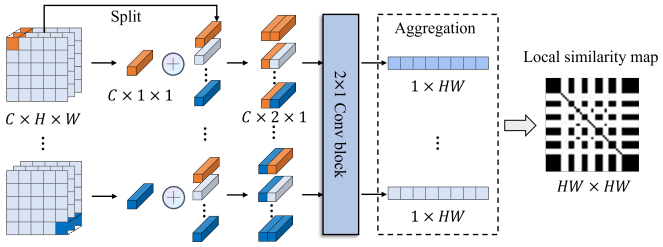


Fig. 2. The proposed learnable local similarity module (LLSM).

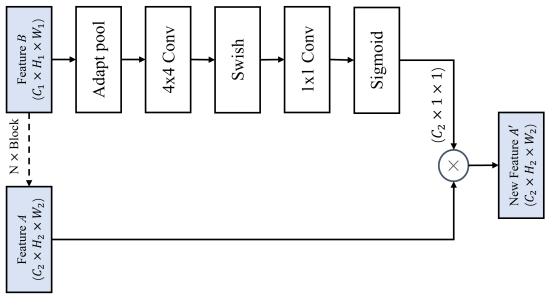


Fig. 3. The proposed cross-level attentional feature fusion module (CLAFFM).

lating consistency or similarity, which may result in feature misalignment issues. For example, when using normalized cosine similarity to measure the similarity of local features, it directly transforms the feature onto a unit hypersphere. However, the most generalizable features extracted from the backbone network may not satisfy the similarity requirement measured by the cosine similarity metric.

In this paper, we propose a plug-and-play learnable local similarity module that densely imposes fine-grained similarity constraints on the features extracted from the backbone network. Unlike existing approaches, the similarities are not directly *calculated* but learned using ConvNets, and we turn the optimization target into binary classification between features.

## III. Methodology

### A. Overview

The proposed approach is shown in Fig. 1. It is a Y-shape network with an encoder, a decoder, and a classifier head, where a learnable local similarity module works in collaboration with the encoder, a cross-level attention feature fusion model connects the encoder and decoder that performs pixel-level forgery localization. The classifier head implements image-level binary classification of authenticity (real/fake) for the input faces.

### B. Learnable Local Similarity Module

Deepfake faces contain subtle artifacts or feature conflicts from several sources, this difference is frequently displayed at the fine-grained feature level and is difficult to detect. Therefore, we build a specific fine-grained local feature similarity map to mine fine-grained feature inconsistencies by computing its feature similarity with all locations based on local features. We constructed a learnable local similarity module for the augmented model to further enhance the model's capacity for generalization by mining fine-grained local differences.

The learnable local similarity module (LLSM) is a plug-and-play module that takes as input the deep features extracted by the backbone and calculates the similarity between local features. Given a face image $I$, we first pass through the encoder to obtain its middle layer feature $F \in R^{H\times W\times C}$ where $C$, $H$, and $W$ represent channel, length, and width respectively. The middle layer feature $F$ is then fed into LLSM, which predicts a single-channel local similarity map. As shown in Fig. 2, the 3-$d$ tensor $F$ is first unfold into a 2-$d$ matrix of shape $\mathbb{R}^{(H\times W)\times C}$, $F_{(i,j)}\,(0 \le i < H, 0 \le j < W)$ denote the local feature, where $i$ and $j$ are the spatial index. A shallow ConvNet with a single kernel of size $2\times1\times C$, denoted by $f$ then perform convolution operation on each local feature pairs. For $\forall i, m \in [1, H], \forall j, n \in [1, W]$, to predict the local similarity map, we use the following equation to calculate the similarity of any local features on the feature $F$:

$$M_{i*H+j,m*H+n} = \sigma(f(F_{(i,j)}, F_{(m,n)})), \tag{1}$$

where $M$ is the predicted 2-$d$ map of size $\mathbb{R}^{(H\times W)\times(H\times W)}$. That means each local feature is compared with all local features including itself by $f$, which outputs a binary prediction on the similarity of arbitrary two local features. After all the local features are calculated and arranged according to the corresponding positions, we can obtain a feature inconsistency map $M^F$ with the shape of $(HW)\times(HW)$. On the optimization strategy for the feature inconsistency map, we calculating the ground truth local similarity map are as follows: First, downsample the mask until its dimensions match those of the middle layer feature map. Then expand the mask map into a one-dimensional tensor $m^{'}$, calculate the Cartesian product of all positional features, and return the local similarity map of $(HW)\times(HW)\times 2$ after adjusting the shape. After that, it is divided into two identically sized feature maps $m_1$,$m_2$ based on the third dimension, and the ultimate ground truth local similarity map $M^{GT}$ is obtained by binarizing the two after an absolute value difference. We adopt the following BCE loss to supervise the training:

$$M^F = LSC\,(F)\,, \tag{2}$$

$$m_1, m_2 = split\left(Cartesia\_prod\left(m^{'}, m^{'}\right)\right), \tag{3}$$

$$M^{GT} = binary\,(|m_1 - m_2|)\,, \tag{4}$$

$$L_{LS} = \frac{1}{N}\sum_{i=1}^{N} BCE\left(M_i^F, M_i^{GT}\right), \tag{5}$$

where LSC stands for the method of predicting the local similarity map, $split$ means split features by dimension, $Cartesia\_prod$ means Cartesian product operation, and $binary$ means binarization operation.

### C. Forgery Localization Module with Multi-level Feature Fusion

Previously, the forgery localization module usually had a single structure or could not adequately integrate the feature

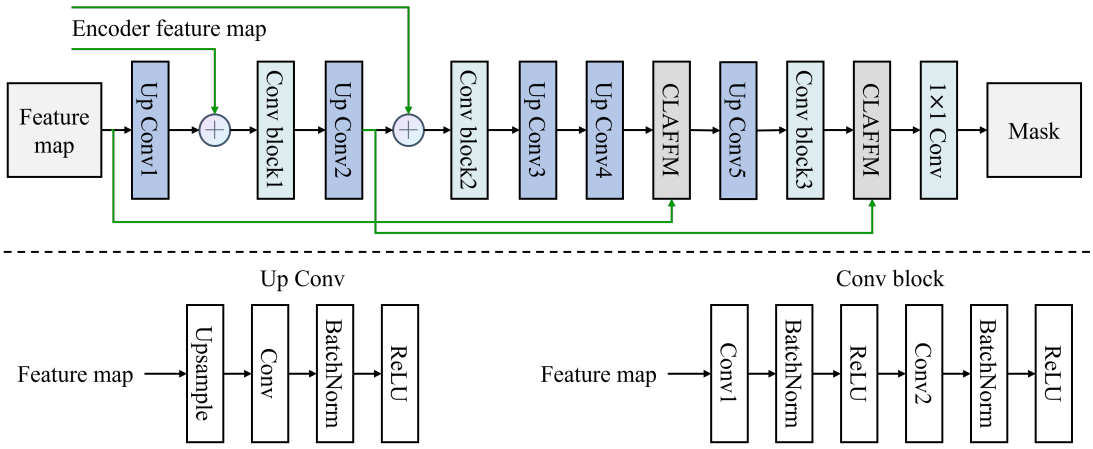


Fig. 4. The detailed architecture of the proposed forgery localization module (FLM).

information of each level, resulting in unsatisfactory outcomes in the forgery region localization task. Therefore, we propose a forgery localization module based on multi-level feature fusion, which made full use of the feature extractor and the feature information of each level of the forgery locator itself, resulting in improved forgery localization effect and detector generality. As shown in Fig. 4, our forgery localization module with multi-level feature fusion receives as input the feature $f_i$ of various block layers of the encoder. Following the up-sampling stage of the forgery locator, the features from different layers of the encoder are coupled with the current features via channels, and more convolution operations are conducted on them to feed into the next up-sampling block. We obtain the forgery localization mask $M^{prd} \in R^{1\times H\times W}$ for the final output prediction after several layers of upsampling.

Skip connection is a commonly used feature fusion technique in neural networks which can aggregate different levels of semantic information. In addition to combining semantic features at various levels in the feature extractor, we additionally introduce a novel attentional fusion module for combining localization feature information at various levels in the forged localization module. Specifically, our proposed cross-level attentional feature fusion module does not only aggregate features. It is able to obtain a set of adaptive learning weights by adaptively learning the attention of low-level semantic features on the channel. These weights are subsequently utilized to highlight the channel attention of high-level semantic features. This attention technique strengthens the robust local similarity information learned by high-level semantic features while still preserving the local similarity details that low-level semantic features value. Its structure is shown in Fig. 3. For feature $A$, get previous feature $B$, first adopt adaptive pooling processing, then perform $4 \times 4$ convolution operation and activate it. After that, perform a $1\times1$ convolution, Sigmoid activate to obtain the attention weight of the same number of channels as the current feature $A$ and output a new feature after multiplying with the feature $A$. With a very tiny amount of calculation, this module may bring the previous feature's attention information to the present feature and increase the positioning accuracy of the fusion feature based on the channel attention level. The forgery localization

TABLE I. CROSS-DATASET EVALUATIONS ON CELEB-DF AND DFD.

| Methods | FF++(C23) | Celeb-DF | DFD |
|---|---|---|---|
| | AUC | AUC | AUC |
| Xception [26] | 99.09 | 65.27 | 87.86 |
| TI2Net [27] | **99.95** | 68.22 | 72.03 |
| FRLM [28] | 99.50 | 70.58 | 68.17 |
| F3Net [20] | 98.10 | 71.21 | 86.10 |
| DMGTN [29] | 99.80 | 72.30 | — |
| Face-X-ray [30] | 87.40 | 74.20 | 85.60 |
| MLDG [31] | 98.99 | 74.56 | 88.14 |
| GFF [32] | 98.36 | 75.31 | 85.51 |
| SFDG [33] | 99.53 | 75.83 | 88.00 |
| SOLA [34] | 99.25 | 76.02 | — |
| MultiAtt [35] | 99.27 | 76.65 | 87.58 |
| BiG-Arts [36] | 99.39 | 77.04 | 89.92 |
| LTW [37] | 99.17 | 77.14 | 88.56 |
| FAAFF [38] | 99.27 | 77.59 | — |
| Local-Relation [17] | 99.46 | 78.26 | 89.24 |
| DCL [25] | 99.30 | **82.30** | 91.66 |
| Ours | 98.54 | 80.56 | **96.01** |

loss is defined as follows:

$$L_{LOC} = \frac{1}{N}\sum_{i=1}^{N} BCE\left(M_i^{prd}, M_i\right), \tag{6}$$

where $M^{prd}$ is the prediction mask output by the multi-level forgery localization module, and $M$ is the ground-truth mask.

### D. Classifier

In addition to the modules mentioned above, we must additionally include a classifier to receive feature input and verify the image's authenticity. Specifically, the features that the decoder outputs are average pooled and input to a fully connected network classifier for classification. For classifier predictions, we use BCE loss for supervised training:

$$L_{CLS} = \frac{1}{N}\sum_{i=1}^{N} BCE\left(y_i^{'}, y_i\right), \tag{7}$$

where $y^{'} \in [0, 1]$ denotes the label of the input image, $y \in \{0, 1\}$ denotes the prediction of the classifier. The overall loss

TABLE II. RESULTS OF IN-DATASET EVALUATIONS ON FF++ C23 AND C40

| Methods | FF++(C23) | | FF++(C40) | | Avg | |
|---|---|---|---|---|---|---|
| | Acc | AUC | Acc | AUC | Acc | AUC |
| Face-X-ray [30] | — | 87.40 | — | 61.60 | — | 74.5 |
| MesoNet [5] | 83.10 | 84.30 | 70.47 | 72.62 | 76.79 | 78.46 |
| Multi-task[6] | 85.65 | 85.43 | 81.30 | 75.59 | 83.48 | 80.51 |
| Xception-ELA [39] | 93.86 | 94.80 | 79.63 | 82.90 | 86.75 | 88.85 |
| SPSL [40] | 91.50 | 95.32 | 81.57 | 82.82 | 86.54 | 89.07 |
| CFFs [11] | — | 97.21 | — | 86.56 | — | 91.89 |
| M2TR [41] | 91.86 | 96.75 | 83.89 | 87.15 | 87.88 | 91.95 |
| Xception [26] | 95.73 | 96.30 | 86.86 | 89.30 | 91.30 | 92.80 |
| Two-branch [42] | **96.43** | 98.70 | 86.34 | 86.59 | **91.39** | 92.65 |
| HFI-Net [43] | 91.87 | 97.07 | 58.69 | 88.40 | 88.78 | 92.74 |
| RFM [44] | 95.69 | **98.79** | 87.06 | 89.83 | 91.38 | 94.31 |
| FST-Matching [45] | 94.05 | 98.27 | **87.38** | 90.44 | 90.72 | 94.36 |
| Ours | 92.84 | 98.54 | 81.13 | **91.93** | 86.99 | **95.24** |

TABLE III. RESULTS OF CROSS-DATESET EVALUATIONS ON FF++ (AUC)

| Methods | Train | DF | F2F | FS | NT | Avg |
|---|---|---|---|---|---|---|
| FDFL [21] | **DF** | 98.91 | 58.90 | **66.87** | 63.61 | 72.07 |
| MultiAtt [35] | | 99.92 | 75.23 | 40.61 | 71.08 | 71.71 |
| GFF [32] | | 99.87 | **76.89** | 47.21 | 72.88 | 74.21 |
| Ours | | **99.99** | 73.06 | 51.86 | **76.09** | **75.25** |
| FDFL [21] | **F2F** | 67.55 | 93.06 | 55.35 | 66.66 | 70.66 |
| MultiAtt [35] | | 86.15 | 99.13 | 60.14 | 64.59 | 77.50 |
| GFF [32] | | **89.23** | 99.10 | **61.30** | 64.77 | **78.60** |
| Ours | | 77.70 | **99.24** | 60.13 | **71.52** | 77.15 |
| FDFL [21] | **FS** | **75.90** | 54.64 | 98.37 | 49.72 | 69.66 |
| MultiAtt [35] | | 64.13 | 66.39 | 99.67 | 50.10 | 70.07 |
| GFF [32] | | 70.21 | 68.72 | **99.85** | 49.91 | 72.17 |
| Ours | | 70.24 | **70.67** | 99.65 | **54.82** | **73.85** |
| FDFL [21] | **NT** | 79.09 | 74.21 | 53.99 | 88.54 | 73.96 |
| MultiAtt [35] | | 87.23 | 48.22 | **75.33** | 98.66 | 77.36 |
| GFF [32] | | **88.49** | 49.81 | 74.31 | **98.77** | 77.85 |
| Ours | | 83.91 | **79.37** | 56.64 | 97.27 | **79.30** |

function is as follows:

$$L = L_{CLS} + \alpha L_{LS} + \beta L_{LOC}, \tag{8}$$

where $\alpha$ and $\beta$ are loss weights and range between [0, 1].

## IV. EXPERIMENTS

### A. Experimental Settings

*1) Datasets:* We conduct experiments on several face forery datasets FaceForensics++ (FF++), Celeb-DF-V2, and Deepfake Detection Dataset (DFD) [46]. FF++ is a large-scale public face forgery dataset that contains 1,000 real videos and 4,000 forged videos created using 4 forgery algorithms, including Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). Additionally, FF++ has three compression levels: original version (Raw), high-quality (C23), and low-quality (C40). CelebDF-V2 is a more challenging dataset consists of 569 original videos and 5,639 forged videos. DFD is a large dataset containing 363 real videos and 3,068 forged videos in various scenarios.

*2) Implementation details:* All faces are detected, cropped, normalized and resized to 224×224 pixels. We use pre-trained Xception as the backbone. All experiments use Adam optimizer with the learning rate set to 1e-4. The batch size is 32, and each epoch has 200 iterations. For the metrics, we utilize binary classification accuracy (Acc.) and area under the curve (AUC.) as the metrics to evaluate model performances.

### B. Evaluations

*1) In-dataset evaluations:* We first evaluate the in-dataset effectiveness of our approach on the FF++, where the network is trained and tested on the same dataset. We only use C23 and C40 versions of FF++ since detecting compressed forged faces is more challenging. We also compare with some state-of-the-art (SOTA) approaches. The results are presented in Table II. We can observe that the proposed approach has promising performances compared to the SOTA approaches when faced with highly compressed forgery faces. The AUC on C23 is close to SOTA [44], while the AUC on C40 exceeds SOTA [45] by 1.49%. This result indicates our method has excellent detection potential. Our proposed method is distinct

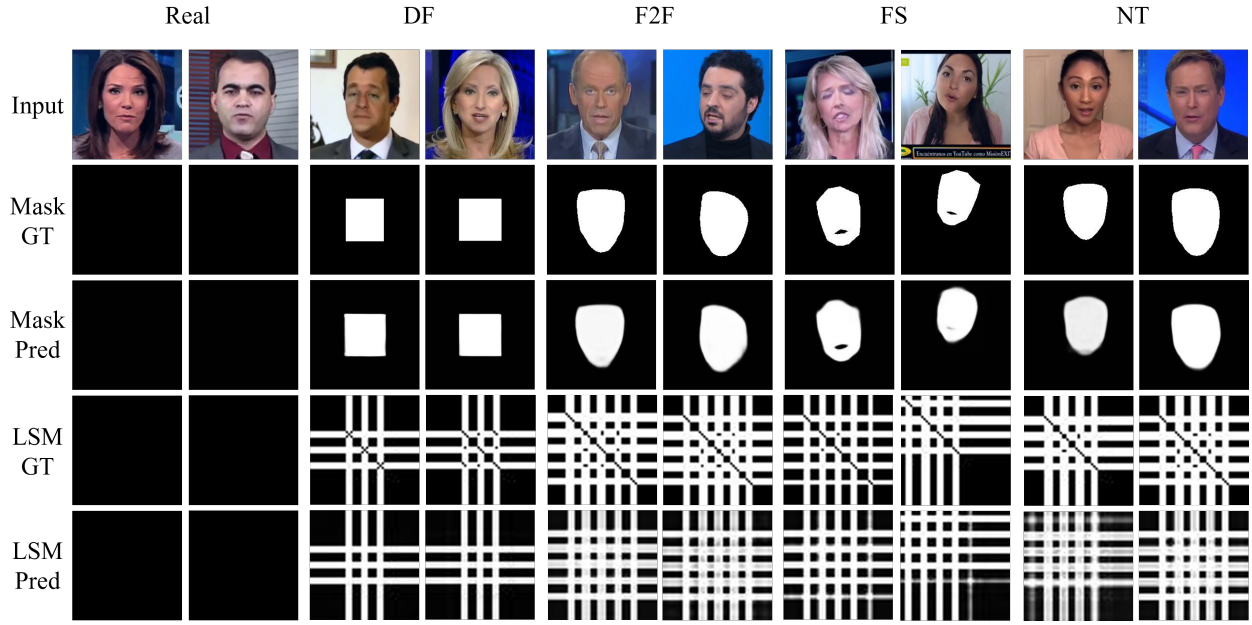


Fig. 5. Visualization of our approach on the FF++. From top to bottom are the original faces, ground-truth of pixel-level mask (Mask GT), predictions of face forgery localization (Mask Pred), ground-truth of local similarity map (LSM GT), and predictions of local similarity map (LSM pred).

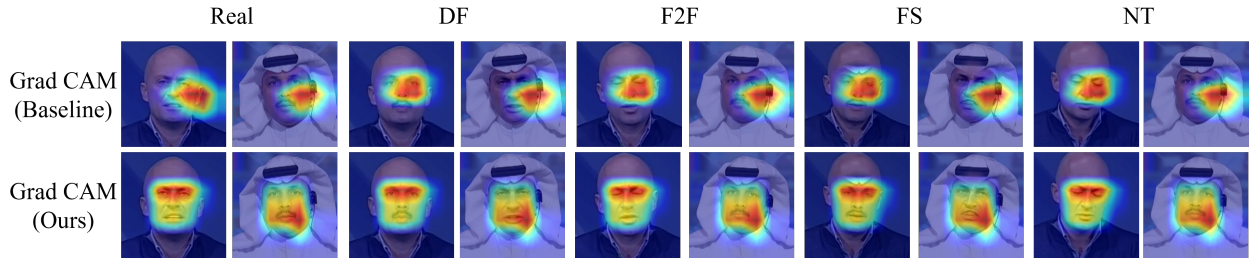


Fig. 6. Visualization of the heatmaps extracted by grad CAM.

from the prior deep learning detection method since it focuses on the fine-grained inconsistency that is shared by various forgery faces rather than just learning the distribution of real and forgery faces, which can effectively improve detection accuracy.

Although we do not achieve the best performance in each setting, we have a clear advantage in high compression scenario (C40) measured by AUC. Moreover, we achieve the best averaged AUC when facing both levels of compression.

*2) Cross-dataset evaluations:* As we all know, there are countless face manipulation techniques in real scenes, but the face manipulation techniques contained in the samples for training detection models are limited. Therefore, the generalization of detection models based on different manipulation methods is of great practical significance. Cross-dataset evaluations directly reflect the generalization of detectors. Table III presents the cross-dataset evaluation results on FF++(C23). We use only one subset of FF++ for training while the remaining 3 subsets for testing. Our approach achieves the highest average AUC across the three training scenarios. We further evaluate the generalizability of our approach on other datasets. As shown in Table I, we train the model on FF++ (C23) and test it on Celeb-DF and DFD. We can observe that our approach can outperform most recent SOTAs by 2% to 10.00% in terms of AUC while maintaining a promising in-dataset performance, with a gap of less than 1.00% on FF++(C23) compared with SOTAs. On DFD, our approach achieves 96.01% AUC which is 4.35% higher than [25] and nearly 10% better than other approaches on average. The results demonstrate that the proposed learnable local similarity can significantly improve generalization capabilities across different forgeries.

TABLE IV. ABALATION STUDIES ON LLSM AND FLM. THE MODEL IS TRAINED ON FF++(C23) AND EVALUATED ON CELEB-DF (AUC)

| Baseline | LLSM | FLM | Celeb-DF |
|---|---|---|---|
| ✓ | | | 72.15 |
| ✓ | ✓ | | 75.11 |
| ✓ | | ✓ | 76.67 |
| ✓ | ✓ | ✓ | **80.56** |

## C. Ablation Study

To demonstrate the effectiveness of each module of our approach, we conduct the following ablation studies: 1) Base-

line (Xception) without any of the proposed modules; 2) Baseline with the proposed learnable local similarity module; 3) Baseline for the proposed forgery localization module with multi-level feature fusion; 4) The complete network. We present the cross-dataset results trained on FF++ (C23) in Table IV. We can observe that the utilization of LLSM leads to a 2.96% improvement in the performance on Celeb-DF, which confirms the effectiveness of local similarity learning in enhancing generalization. The proposed framework, when incorporated with FLM, demonstrated a 4.52% improvement in performance, affirming its superiority. It is noteworthy that the combined use of FLM and LLSM results in a remarkable performance improvement.

### D. Visualization

To provide further evidence of our approach's effectiveness, we evaluate the visualization results using the FF++. As shown in Fig. 5, we compare the forgery localization results and pixel level ground-truth of four different forgery types. The results demonstrate that our approach is capable of performing high-precision forgery localization. Additionally, we present the predictions of LLSM and the corresponding ground-truth similarity map. For real faces, given that the local features share the same source, the similarity maps do not exhibit any abnormal patterns. However, the similarity map of the forged face reveals clear abnormal patterns, which differ depending on the forged region. The LLSM can also achieve accurate predictions on the local similarity patterns. This provides further evidence of the effectiveness of our proposed approach.

We also visualize the heatmaps extracted using Grad CAM [47] to demonstrate the effectiveness of the proposed approach. In these attention maps, the warmer color indicates the areas more significant for predictions or localization. As shown in Fig. 6, we observe that the baseline model is not accurate in identifying the manipulated region, whereas our approach successfully directs the network's attention to the forged facial region.

## V. Conclusion

In this paper, we propose a dual-task approach that achieves generalized face forgery detection and accurate forgery localization. Our approach takes advantage of the feature similarity between the internal parts of the forged image. The learnable local similarity module successfully enhances the difference traces between real and forged features and improves the generalization of the model. Furthermore, from a multi-tasking learning view, we present a forgery localization module with cross-level attentional feature fusion strategy, which improves the detection capability even further. We conduct extensive experiments, and the results fully demonstrate the effectiveness of our approach. However, our suggested learnable local similarity module relies on fine-grained local feature calculation, which requires more computational overhead and has feature size restrictions. Moreover, our approach has limited robustness for very low-quality faces. In further studies, we will consider calculating local Inconsistencies for a few local features instead of the entire feature to reduce computational overhead. Additionally, we will design novel image enhancement algorithms to improve the robustness of the detection model.

In the future, exploring the connections between local inconsistencies, identity inconsistencies, and inter-frame inconsistencies may further improve forged face detection performance.

## References

[1] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[2] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[3] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3677–3685.

[4] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

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

[6] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "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)*. IEEE, 2019, pp. 1–8.

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

[8] P. He, H. Li, and H. Wang, "Detection of fake images via the ensemble of deep representations from multi color spaces," in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 2299–2303.

[9] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. IEEE, 2019, pp. 83–92.

[10] Z. Liu, X. Qi, and P. H. Torr, "Global texture enhancement for fake face detection in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8060–8069.

[11] P. Yu, J. Fei, Z. Xia, Z. Zhou, and J. Weng, "Improving generalization by commonality learning in face forgery detection," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 547–558, 2022.

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

[13] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo, "Protecting celebrities from deepfake with identity consistency transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9468–9478.

[14] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes." in *CVPR workshops*, vol. 1, 2019, p. 38.

[15] Z. Shang, H. Xie, Z. Zha, L. Yu, Y. Li, and Y. Zhang, "Prrnet: Pixel-region relation network for face forgery detection," *Pattern Recognition*, vol. 116, p. 107950, 2021.

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

[17] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, "Local relation learning for face forgery detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 1081–1088.

[18] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: automatically replacing faces in photographs," in *ACM SIGGRAPH 2008 papers*, 2008, pp. 1–8.

[19] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.

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

[21] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6458–6467.

[22] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2018, pp. 1–6.

[23] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, and L. Ma, "Delving into the local: Dynamic inconsistency learning for deepfake video detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 744–752.

[24] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5039–5049.

[25] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji, "Dual contrastive learning for general face forgery detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 2316–2324.

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

[27] B. Liu, B. Liu, M. Ding, T. Zhu, and X. Yu, "Ti2net: Temporal identity inconsistency network for deepfake detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4691–4700.

[28] C. Miao, Q. Chu, W. Li, S. Li, Z. Tan, W. Zhuang, and N. Yu, "Learning forgery region-aware and id-independent features for face manipulation detection," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 71–84, 2021.

[29] B. Liang, Z. Wang, B. Huang, Q. Zou, Q. Wang, and J. Liang, "Depth map guided triplet network for deepfake face detection," *Neural Networks*, vol. 159, pp. 34–42, 2023.

[30] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001–5010.

[31] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

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

[33] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng, "Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7278–7287.

[34] J. Fei, Y. Dai, P. Yu, T. Shen, Z. Xia, and J. Weng, "Learning second order local anomaly for general face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 270–20 280.

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

[36] H. Chen, Y. Li, D. Lin, B. Li, and J. Wu, "Watching the big artifacts: Exposing deepfake videos via bi-granularity artifacts," *Pattern Recognition*, vol. 135, p. 109179, 2023.

[37] K. Sun, H. Liu, Q. Ye, Y. Gao, J. Liu, L. Shao, and R. Ji, "Domain general face forgery detection by learning to weight," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 2638–2646.

[38] C. Tian, Z. Luo, G. Shi, and S. Li, "Frequency-aware attentional feature fusion for deepfake detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[39] T. S. Gunawan, S. A. M. Hanafiah, M. Kartiwi, N. Ismail, N. F. Za'bah, and A. N. Nordin, "Development of photo forensics algorithm by detecting photoshop manipulation using error level analysis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 7, no. 1, pp. 131–137, 2017.

[40] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 772–781.

[41] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, "M2tr: Multi-modal multi-scale transformers for deepfake detection," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022, pp. 615–623.

[42] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020, pp. 667–684.

[43] C. Miao, Z. Tan, Q. Chu, N. Yu, and G. Guo, "Hierarchical frequency-assisted interactive networks for face manipulation detection," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3008–3021, 2022.

[44] C. Wang and W. Deng, "Representative forgery mining for fake face detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 923–14 932.

[45] S. Dong, J. Wang, J. Liang, H. Fan, and R. Ji, "Explaining deepfake detection by analysing image matching," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*. Springer, 2022, pp. 18–35.

[46] N. Dufour and A. Gully, "Contributing data to deepfake detection research," *Google AI Blog*, vol. 1, no. 3, 2019.

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