ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240725

Mutual Local Consistency Learning for Face Forgery Detection

Bosheng Yan[a] **and Chang-Tsun Li**[a,*]

[a]Deakin University, Australia

Abstract. The rapid advancement of face manipulation technology has spurred an urgent need for forgery detection. Existing deepfake detection approaches have achieved impressive performance under the intra-dataset scenario where the same algorithm generates training and testing face data. However, the performance is by no means satisfactory when the methods are applied to unseen forgery datasets. To tackle this problem, in this paper, we propose a new perspective of face forgery detection by considering feature inconsistency in spatial and frequency domains in manipulated images. Specifically, we design a two-stream network equipped with a Multi-scale Mutual Local Consistency Learning module (MMLCL) that consists of a Global Enhancement Module (GEM) combining Mutual Local Consistency Learning (MLCL) to learn local consistency in multi-scale enhanced feature maps. We further exploit the mutual representation to obtain an attention map that serves as guidance of forged regions on the output features for final classification. Extensive experiments demonstrate that our proposed method achieves effectiveness and generalization towards unseen face forgeries.

1 Introduction

The rapid advancement of machine learning and computer vision techniques in recent years has accelerated the technical leap of deepfake creation in general. In particular, face manipulation methods [1, 2, 34] are now sophisticated enough to vividly swap the identity or replace the expression of a target individual with that of a source individual. These methods enable attackers to generate convincing forged face images/videos without professional skills, which causes a series of negative social impacts. For example, a deepfake video of Ukrainian President Zelenskyy was created by malicious attackers encouraging soldiers to surrender. As such, researchers have been motivated to develop effective countermeasures [38, 21, 14, 35, 15] to mitigate such malicious abuse of machine learning and computer vision techniques.

Early deepfake detection methods [3, 30] analyze visual cues left by the post-processing procedures, such as inconsistent head poses [36], abnormal eye blinking frequency [21], face region blending boundary [20], or employ off-the-shelf deep neural network to model the decision boundary between pristine and forged faces. However, the visual artifacts are significantly weakened as advanced deepfake techniques emerge and these methods only perform well in the within-database setting and lack explainability. Some works [38, 6] introduce the attention mechanism to enable networks to locate manipulated regions that could mine inadequate forged clues at some particular regions in the spatial domain and also suffer from the poor between-database generalization problem. It indicates that there is still a noticeable performance gap when testing on unseen manipulated face data.

To tackle the above challenges, we propose a novel framework called: Mutual Local Consistency Learning (MLCL) for manipulated face detection. Conceptually, spatially-local information is expected to be consistent in the same sources. Thus, we assume that the consistency between features in the spatial and frequency domains can be attributed to the inherent characteristics of facial structures. Since the existing state-of-the-art face manipulation techniques only modify the facial region, the original features of other areas (e.g. background, neck) are still preserved. Therefore, a manipulated image contains different features in the spatial domain at different local regions and the forged clues are amplified in the frequency domain. These features are inherent to the frequency domain and are inconsistent with the original features. Due to the features of a pristine image being consistent across all locations, we propose MLCL to learn local-level consistency between RGB and frequency features to identify real or fake faces. Specifically, we design a Multi-scale Mutual Consistency Learning module (MMLCL) that consists of a Global Enhancement Module (GEM) with mutual consistency learning to predict the mutual consistency maps based on each pair of local regions in the feature maps of two streams. The MMLCL models the local region relationship by exploiting the RGB and frequency information and improves the discrimination between the original features and forgery features in multi-scale down-sampled feature maps. Pair-wise cosine measurement is employed to build the mutual consistency maps. To compute the consistency loss for each pair, the mutual consistency maps are constructed with a dynamic ground-truth manipulation mask which indicates if the local region is forged in a self-supervised manner. This means that we penalize the pair of feature vectors in the feature maps from the authentic regions with low similarity and those with diverse features for having high similarity. Moreover, to learn more comprehensive representations, we propose a Mutual Representation Guided Attention Module (MRGA) to fuse RGB and frequency stream features and obtain an attention map for indicating potentially manipulated regions. Then, the output of MRGA is used for classification.

The main contributions of this work are summarized as follows:

- We design a mutual local consistency learning (MLCL) framework for capturing long-range dependencies among local regions and then learning local consistency information between RGB images and their high-pass filtered versions on multi-scale feature maps. It facilitates the model to understand the relationship be-

* Corresponding Author. Email: changtsun.li@deakin.edu.au.

tween real and forged regions.
- We propose a mutual representation guided attention module (MRGA) to combine both RGB and frequency information and collaboratively exploit the mutual representation as guidance to attend to probably forged regions.

The rest of this paper is organised as follows. Section 2 reviews related works on face forgery detection and consistency learning as well as the research gaps. Section 3 presents our methodology for bridging those research gaps. Section 4 demonstrates experimental results to validate our methodology in a comparative manner. Section 5 concludes this work.

2 Related Work

2.1 Deepfake Detection

Recent studies attempt to detect different face manipulations with specific artifacts appearing on the forged faces or discriminative functions learned from data. Li et al. [21] observe that deepfakes have abnormal eye-blinking frequency and propose to check for the inconsistency of head poses [36] to expose manipulated faces. For data-driven approaches, [30] employs powerful neural network Xception [5] to extract separative features. Subsequently, Frank et al. [12] start to pay attention to the information stored in the frequency domain and exploit the peculiar frequency spectrum pattern for deepfake detection. Qian et al. [28] mines forged clues under the frequency domain. Recent works [4, 23] combine the frequency information with the RGB information. F^3-Net [19] reveals discriminative feature by adaptively extracting frequency-related features. Luo et al. [25] adopt SRM filter to extract frequency noise feature to develop a general deepfake detector. However, these methods fail to show generalization to unseen face manipulations. Several works also integrate localization methods into forgery detection to develop a generalized detector. In [27], multi-task learning is employed to integrate the auxiliary localization tasks into the classification of deepfakes. Zhao et al. [38] propose to adopt an attention mechanism to highlight multiple manipulated areas. Face X-ray [20] customizes a face generation pipeline to localize the forged boundary introduced by the image stitching process to improve the generalization ability. To develop a general face detector, Rao et al. [29] introduce domain adoption to avoid the model overfitting to a single domain. Different from these methods, we utilize the feature inconsistency between the pristine and forged images as an effective cue to enhance the generalizability of the model rather than relying on the hand-crafted features that merely target the specific manipulation method.

2.2 Consistency Learning

Inconsistency detection has long been adopted for image forensics [26], where the affinity is calculated among image patches. Dong et al. [9] propose a transformer-based network to detect identity consistency based on high-level semantics between the inner face and the outer face. Hu et al. [16] capture global and local inconsistency information from a pair of adjacent frames for deepfake video detection. Zhou et al. [40] use a two-stream network to extract spatial and steganalysis features for detecting forged faces and shallow-level inconsistencies. However, their methods mainly focus on the inconsistency information in the spatial domain. In this paper, we propose a two-stream architecture that learns multi-scale internal relations among regions within the face images by estimating self-consistency between extracted spatial and high-frequency features. In addition, we employ ground-truth masks to guide the self-consistency adaptively in a self-supervised manner.

3 Proposed Method

In this section, we illustrate our proposed Mutual Local Consistency Learning network (MLCL), as shown in Fig 1. Our objective is to identify if the face region is tampered with the face of another individual. Based on the observation that the features of manipulated regions are diverse from the original features, we measure the feature consistency within a face image for capturing the forged traces. In our framework, we exploit frequency information to assist the network in learning robust representation due to the artifacts that can be captured in the frequency domain [28]. More specifically, our proposed MLCL consists of two streams, where an RGB image and its high-pass filtered version are taken as input. The multi-scale mutual local consistency learning module is adopted to predict the mutual consistency maps for showing the relationship between the spatial and the frequency features across different resolutions. Ground-truth masks are used to supervise the consistency maps adaptively. The mutual representation guided attention module is applied to guide the classifier to focus on the forged traces. The model is an end-to-end learning architecture. The details of these two modules are presented in the following subsections.

3.1 Frequency-aware Information

To fully exploit the frequency information without loss of generality, we transform $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ which denotes the RGB input with height H and width W to the frequency domain. This can be formulated as:

$$\mathbf{X}_f = \mathcal{F}(\mathcal{D}(\mathbf{X}), \beta), \tag{1}$$

where $\mathbf{X}_f \in \mathbb{R}^{H\times W\times 1}$, $\mathcal{D}$ stands for the Discrete Cosine Transform (DCT). DCT is commonly used to separate the frequency distribution of an image into different frequency bands, while the low-frequency components are concentrated in the top-left corner, and the high-frequency components are located in the bottom-right corner. $\mathcal{F}$ denotes high-pass filtering. To be specific, $\mathcal{F}$ shifts the low-frequency information to the center of $\mathcal{D}(\mathbf{X})$ and filters it out by setting a circular region to 0, where the circle takes the center of $\mathcal{D}(\mathbf{X})$ as center point with radius (i.e. $\beta \times W$). We employ a high-frequency filter to suppress the low-frequency content to amplify the artifacts hidden at the high-frequency band. However, the conventional frequency domain image does not inherit the shift-invariance and local consistency from the natural image, so we use $\mathcal{D}^{-1}$ to convert $\mathbf{X}_f$ back into the RGB color space to get the intended representation in the frequency domain. The process is shown below:

$$\mathbf{X}^f_{input} = \mathcal{D}^{-1}(\mathbf{X}_f), \tag{2}$$

where $\mathbf{X}^f_{input} \in \mathbb{R}^{H\times W\times 3}$.

3.2 Multi-scale Mutual Local Consistency Learning

Our carefully designed Multi-scale Mutual Local Consistency Learning module (MMLCL) is built with the Global Enhancement Module (GEM) followed by the Mutual Local Consistency Learning operation (MLCL). To improve the discrimination of global feature vectors in the input of each stream, we propose a GEM to explore the

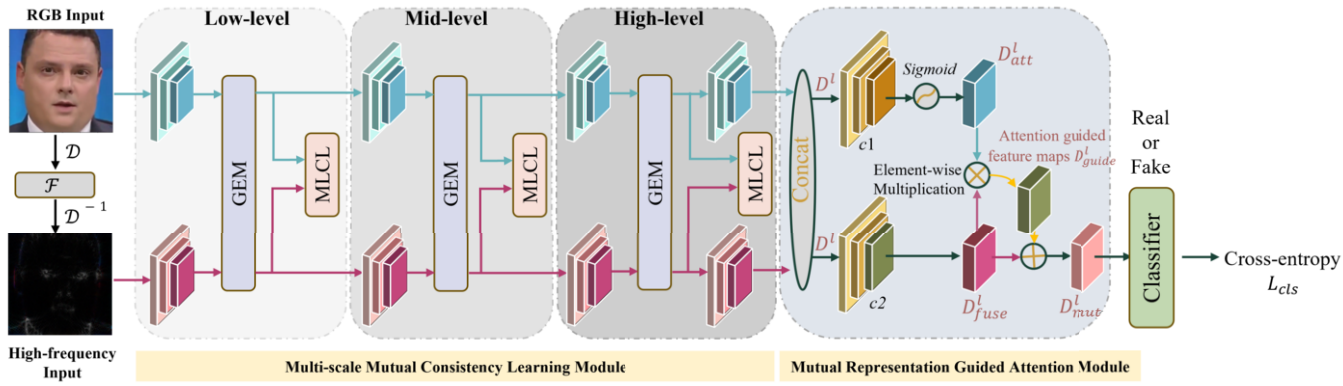

Figure 1: The overview of the proposed Mutual Local consistency Learning architecture for deepfake detection. Our proposed MLCL consists of RGB and frequency streams. Two modules are proposed: the Multi-scale Mutual Consistency Learning module captures the forgery clues and makes the model more robust to the variations in the size of the forgery region; the Mutual Representation Guided Attention module forces the network to focus more on the potential forged regions.

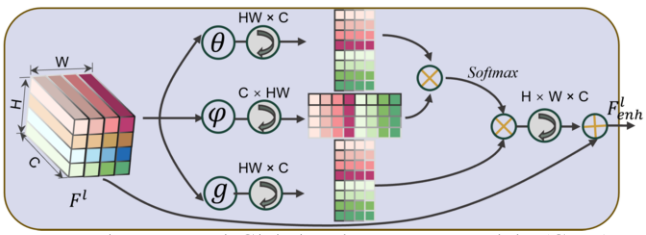

Figure 2: The proposed Global Enhancement Module (GEM) explores the long-range dependencies of the images with self-attention.

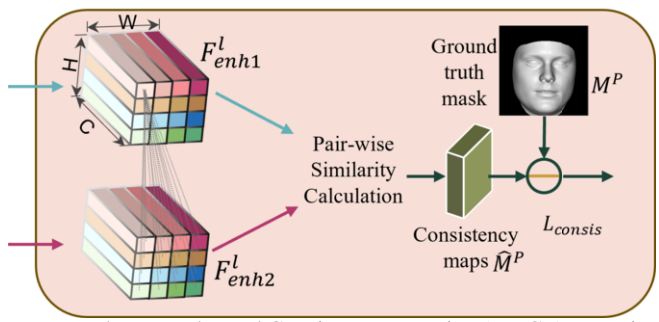

Figure 3: The Mutual Local Consistency Learning (MLCL) operation measures the consistencies of local features of RGB and frequency streams.

long-range dependencies across multi-scales in the images with self-attention, inspired by [11]. Specifically, GEM takes the intermediate feature maps $\mathbf{F}^l \in \mathbb{R}^{h_l \times w_l \times c_l}$ of both RGB and frequency streams as input, where l indicates the l-th layer. The details of the GEM are demonstrated in Fig 2. To be specific, $\mathbf{F}^l$ is projected using embedding functions θ and φ and the weight matrix $\mathbf{W}^l$ is calculated as:

$$\mathbf{W}^l = softmax((\theta(\mathbf{F}^l))^T \otimes \varphi(\mathbf{F}^l))), \tag{3}$$

where each vector in $\mathbf{W}^l$ represents the weight assigned to local features and demonstrates the relevance among each local vector in a feature map, given that local features of authentic regions are less relevant to that of forged regions and vice versa. We multiply $\mathbf{W}^l$ with $\mathbf{F}^l$ transformed by embedding function g to produce feature maps $\hat{\mathbf{F}}^l_{enh}$ which is then added with input feature maps $\mathbf{F}^l$ spatially to obtain enhanced feature map $\mathbf{F}^l_{enh}$:

$$\mathbf{F}^l_{enh} = \mathbf{F}^l + \mathbf{W}^l \otimes g(\mathbf{F}^l). \tag{4}$$

As shown in Fig 3, the enhanced feature maps of each stream (i.e., $\mathbf{F}^l_{enh1}$, $\mathbf{F}^l_{enh2}$) are sent to the MLCL operation for predicting the mutual consistency maps.

To effectively capture the forgery clues and make the model more robust to the variations in the size of the forgery region, we utilize the multi-scale features for constructing different-level consistency maps. Across multiple scales of features, shallow-level features with large spatial dimensions facilitate localization; mid-level features contribute to identifying subtle discrepancies with fine-grained analysis of the facial features; high-level features employed for identification are rich in semantic information. More specifically, we compare each patch $P^1_{p,q}$ in the enhanced feature maps $\mathbf{F}^l_{enh1}$ against all the patches P^2 in the $\mathbf{F}^l_{enh2}$ to reflect their feature affinities and acquire a patch-corresponding consistency map M_c with the same size of a enhanced feature map. The subscript means the position index of the patch and the superscript indicates the patch from which stream. For any pair of patches P^1_i and P^2_j, we adopt cosine similarity to calculate their similarity score:

$$S(P_i, P_j) = \frac{\frac{P_i}{\|P_i\|^1} \cdot \frac{P_j}{\|P_j\|^2} + 1}{2}, \tag{5}$$

where the similarity score falls into the range of [0, 1]. This process is repeated over all patches $\{P^l_{p,q} \mid 0 \le \mathrm{p} \le h_l, 0 \le \mathrm{q} \le w_l\}$ in the $\mathbf{F}^l_{enh1}$. The result of this operation is a set of 2D consistency maps M^P of size $h_l \times w_l$ with $h_l \times w_l$ channels.

We employ the ground-truth manipulation mask M with size $H \times W$ which highlights the forged region, and transform it dynamically to guide the multi-scale consistency maps. To construct the ground-truth 2D consistency mask $M^{P_{p,q}}$ for the consistency map of (p, q)-th patch correspondingly, we divide M into $h_l \times w_l$ patches $\mathbf{m}^l_i \in \mathbb{R}^{H_l \times W_l}$ corresponding to specific scale of MMLCL, where $H_l = (\frac{H}{h_l})$ and $W_l = (\frac{W}{w_l})$, $i \in \{1, 2, 3, ..., h_l \times w_l\}$. We sum up all pixels in each patch $\mathbf{m}^l_i$ and obtain the relation between a pair of patches by computing their differences:

$$M^{P_{p,q}} = 1 - (|\mathbf{m}^l_{p,q} - Sum(\mathbf{m}^l_i)|/(h_l \times w_l)), \tag{6}$$

where $\mathbf{m}^l_{p,q}$ is the sum value of the (p, q)-th patch, and $M^{P_{p,q}} \in \mathbb{R}^{h_l \times w_l}$ is obtained by iterating over all patches in M. Each entry $s_i \in [0, 1]$ in $M^{P_{p,q}}$ indicates the probability of forgery, where the value close to 1 means the two patches are coherent, and vice versa. We compute $M^{P_{p,q}}$ for all patches to get ground-truth consistency

masks for all 2D consistency maps. It is worth noting that each value in the ground-truth mask M for the real image should be 0. We formulate the consistency learning loss function as:

$$\mathcal{L}_{consis} = \frac{1}{h_l \times w_l} \sum_{j=1}^{h_l \times w_l} ||M_j^{P_{p,q}} - \widehat{M}_j^{P_{p,q}}||, \quad (7)$$

where j denotes the index of the ground-truth mask $\widehat{M}^{P_{p,q}}$ to the corresponding consistency map $M^{P_{p,q}}$. $h_l \times w_l$ is the number of channels of 2D consistency maps $\widehat{M}^P$.

3.3 Mutual Representation Guided Attention

In this section, we develop a mutual representation guided attention module (MRGA) to force the network to focus more on the potential forgery regions, which facilitates later face forgery detection. With the discriminative global features enhanced by GEM, we fuse both spatial and frequency information to facilitate feature representation learning. To take full advantage of the knowledge that RGB information helps to reveal abnormal textures and frequency information magnifies subtle forged clues, we combine both spatial and frequency features at deep semantic layers, contributing to learning an attention map. As shown in Fig 1, deep semantic feature maps of the RGB stream and frequency stream at the *l*-th layer of the network (i.e, $\mathbf{D}_1^l \in \mathbb{R}^{h_l \times w_l \times c_l}$ and $\mathbf{D}_2^l \in \mathbb{R}^{h_l \times w_l \times c_l}$, where h_l, w_l and c_l are the height, width, and channel number, *l* denotes high-level for simple representation) are concatenated in the channel dimension to obtain $\mathbf{D}^l \in \mathbb{R}^{h_l \times w_l \times 2c_l}$. We apply the residual learning mechanism to our designed MRGA to enhance the representation of features by focusing on important areas while preserving the integrity of the original input. Specificaly, we compute the attention map $\mathbf{D}_{att}^l$ based on $\mathbf{D}^l$ and transform $\mathbf{D}^l$ to acquire $\mathbf{D}_{fuse}^l$ through a shallow convolutional neural network. Finally, we add $\mathbf{D}_{guided}^l$ to $\mathbf{D}_{fuse}^l$ spatially to output the mutual feature maps $\mathbf{D}_{mut}^l$:

$$\mathbf{D}_{guided}^l = \sigma(c1(\mathbf{D}^l)) \otimes c2(\mathbf{D}^l), \quad (8)$$

$$\mathbf{D}_{mut}^l = \mathbf{D}_{guided}^l + \mathbf{D}_{fuse}^l, \quad (9)$$

where σ represents the sigmoid function, *c1*, *c2* specify the convolutional operation, and $\otimes$ denotes the element-wise multiplication.

3.4 Loss Function

In the last stage, the outputs of MRGA are sent to the classifier to obtain probability y_{pred} for face forgery detection. We employ the widely-used Cross-Entropy loss as the objective function:

$$\mathcal{L}_{ce} = -\frac{1}{N} \sum_{i}^{N} [y_{gt} \log(y_{pred}) + (1 - y_{gt}) \log(1 - y_{pred})], \quad (10)$$

where y_{gt} is the ground truth label and is set to 1 for fake face images, otherwise it is set to 0.

By integrating the Cross-Entropy loss with the consistency learning loss, we have the overall loss function defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{consis}, \quad (11)$$

where λ_1 and λ_2 are the weight parameters for balancing these two losses. By default, we set $\lambda_1 = \lambda_2 = 1$ in our experiments.

4 Experiment Results and Analysis

4.1 Experimental Setup

Dataset. To evaluate our proposed method, we conduct experiments on four widely used benchmark datasets: FaceForensis++ (FF++) [30], Celeb-DF [22], DeepFakeDetection (DFD) [10] and Deepfake Detection Challenge (DFDC) [8] datasets. FF++ is a face manipulation detection dataset containing 1000 pristine videos and each video has its fake version with generating by 4 different algorithms (i.e., DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT)). The videos in FF++ dataset have three different-level qualities (i.e., raw, high-quality (HQ), and low-quality (LQ)). All performances of our proposed method are based on high-quality videos unless otherwise stated. Celeb-DF is a large-scale deepfake detection dataset comprising 590 real videos and 5639 realistic fake videos manipulated by advanced DeepFake algorithms. The DFD is produced by Google, including 3068 fake videos created by using Deepfake techniques of 28 actors in various scenes. To evaluate the generalization ability of our method, we also test our method on DFDC which is a large-scale deepfake dataset designed for the Facebook DeepFake Detection Challenge.

Implementation Details. We employ EfficientNet-B2 [33] pretrained on ImageNet as the backbone for the two-stream network. We divide the backbone into multi-scale feature extractors (i.e., low-level, mid-level, and high-level) as shown in Fig. 1. For the face forgery detection datasets that do not come with ground-truth manipulated masks, we self-define the ground-truth masks for forged face images by using RetinaFace [7] to detect the face region and set 1.3 times the size of the bounding box to preserve enough background areas. All frames are resized to 256×256 and the β in Eq. 1 is empirically set to 0.3. The network is trained using Adam optimizer [18] with a learning rate of 4e-4, and the learning rate is decreased to half every 10 epochs. We set the batch size to 8, and the number of training epochs to 40.

Evaluation metrics. In our experiment, we apply the Accuracy score (Acc), Area Under Receiver Operating Characteristic (AUC), and Equal Error Rate (EER) as our evaluation metrics following the convention [30, 28].

4.2 Experimental Results

Intra-dataset evaluation. We compare the proposed MLCL with representative prior methods on the FF++ (HQ) and the Celeb-DF datasets to demonstrate its effectiveness. Some of these methods [11, 25] either adopt consistency learning or exploit frequency information for deepfake detection. Intra-dataset evaluation means the training and testing on the same dataset. From Table 1, we can observe that the proposed MLCL achieves the best performance (96.58%) in Acc and closes to SOLA (drop 0.2%) in AUC when tested on the FF++ dataset while outperforming other reference methods by a considerable margin in terms of Acc and AUC on Celeb-DF datasets.

Cross-dataset evaluation. The generalization ability poses a challenge to existing deepfake detection methods due to new forgery methods emerging. To show the generalization capability of our proposed MLCL, we conduct the cross-dataset evaluation by training the model on the FF++ (HQ) dataset while testing on Celeb-DF, DFDC and DFD respectively.

Table 2 demonstrates the cross-dataset evaluation results compared with other methods exploiting high-frequency features on

Table 1: Intra-dataset comparison on the FF++(HQ) and Celeb-DF datasets, respectively. The best result is highlighted in bold, and the second-best is underlined. The last row (†) complements the F1 score (%) of our method.

Method	FF++ (HQ)		Celeb-DF	
	Acc (%)	AUC (%)	Acc (%)	AUC (%)
Xception [5]	95.73	96.30	94.10	97.65
Face X-ray [20]	89.25	90.40	-	-
F^3-Net [28]	96.07	98.53	93.95	96.53
MAT [38]	96.20	99.09	93.72	98.50
PCL [39]	95.80	99.18	97.80	<u>98.93</u>
DCL [32]	95.78	99.30	-	-
EN-B4 [33]	95.24	98.21	96.63	98.21
GFF [25]	94.93	97.86	-	-
PEL [13]	96.13	99.32	-	-
SOLA [11]	<u>96.25</u>	**99.60**	<u>97.97</u>	98.79
Ours	**96.58**	<u>99.40</u>	**99.11**	**99.80**
Ours† (F1 Score)	97.54		98.48	

Table 2: Cross-dataset evaluation on other datasets by training on FF++(HQ) in terms of AUC (%) and EER (%). The last row (†) complements the F1 score (%) of our method.

Method	Celeb-DF		DFDC		DFD	
	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
Xception [5]	60.70	43.75	62.51	31.45	87.86	21.04
F^3-Net [28]	63.41	42.28	65.30	41.78	86.10	26.17
MAT [38]	69.08	37.80	68.21	37.15	87.58	21.73
Local-relation [39]	77.12	31.98	73.52	35.97	73.52	35.97
DIFL [37]	76.28	32.52	73.17	36.28	-	-
GFF [25]	75.25	33.01	70.28	35.23	85.51	25.64
Ours	**77.53**	**31.50**	**76.21**	**33.83**	**89.68**	**19.35**
Ours† (F1 Score)	72.63		70.78		84.60	

Celeb-DF, DFDC and DFD. For cross-dataset evaluation on Celeb-DF, our proposed MLCL outperforms all other state-of-the-art and achieves 77.53% in AUC. It is worth noting that the performance of our model is on par with Local-relation [39] which exploits the consistency information in the spatial domain. The gain of MLCL benefits from learning the intrinsic relationship between the RGB and frequency domains, which is a significant cue for detecting manipulated faces with various algorithms. For the results of a more challenging dataset DFDC, MLCL achieves convincing results with 76.21% in AUC, which is superior to all listed methods. The results indicate that our designed MLCL operation explicitly discriminates the intrinsic differences between real and forged regions, which improves the generalization ability of the model to unknown manipulation techniques.

Cross-manipulation evaluation. To comprehensively evaluate the generalization capability of the proposed MLCL to unseen forgeries, we further conduct cross-manipulation experiments on four different methods in the FF++(HQ) dataset (i.e. Deepfake (DF), Face2Face (F2F), FaceSwap (FS), NeuralTexture (NT)). Following [32, 25, 24, 17], the model is trained on one method and test it on all four methods. We reimplement PCL [39] with EfficientNet-B2 backbone for a fair comparison. As tabulated in Table 3, our proposed MLCL outperforms other state-of-the-art methods, especially PCL [39] in most cases under both intra-manipulation and cross-manipulation settings. Specifically, when the models are trained on NeuralTexture and tested on FaceSwap, our MLCL achieves nearly 5% performance improvement in terms of AUC. Since the PCL exploits only RGB features for consistency learning to detect forged faces, our method explores the local feature differences in both RGB and frequency domains, which facilitates the localization of the modified regions and learns the feature differences between real and manipulated faces for improving the performance of the cross-dataset evaluation.

Table 3: Cross-manipulation evaluation on FF++(HQ) in terms of AUC(%). Diagonal results report the intra-manipulation performance. * indicates the reproduced result

Train Set	Method	Backbone	Test Set (AUC (%))			
			DF	F2F	FS	NT
DF	Xception [5]	Xception	99.85	69.34	46.05	67.65
	EN-B2 [33]	EfficientNet-B2	99.90	72.21	47.21	66.54
	PCL* [39]	EfficientNet-B2	99.97	71.81	48.84	70.78
	Ours	EfficientNet-B2	**99.98**	**75.78**	**54.65**	**74.27**
F2F	Xception [5]	Xception	77.32	99.01	57.82	65.34
	EN-B2 [33]	EfficientNet-B2	76.67	99.20	60.24	65.18
	PCL* [39]	EfficientNet-B2	78.13	99.12	**62.15**	67.48
	Ours	EfficientNet-B2	**82.16**	**99.62**	61.98	**69.56**
FS	Xception [5]	Xception	65.21	66.45	98.95	47.68
	EN-B2 [33]	EfficientNet-B2	63.24	67.38	99.58	51.10
	PCL* [39]	EfficientNet-B2	68.36	**72.14**	99.86	55.28
	Ours	EfficientNet-B2	**72.71**	71.58	**99.98**	**60.27**
NT	Xception [5]	Xception	80.21	52.28	68.26	95.56
	EN-B2 [33]	EfficientNet-B2	**84.08**	52.64	70.62	96.90
	PCL* [39]	EfficientNet-B2	82.97	55.38	69.95	96.85
	Ours	EfficientNet-B2	83.68	**59.16**	**75.86**	**97.50**

4.3 Ablation Study

To investigate the effectiveness of the components of our proposed framework, we conduct the following variants: 1) the baseline model only takes RGB images as input, 2) our method only contains RGB and high-frequency inputs and concatenates the final feature maps of two streams directly, 3) MLCL that only adopts MGRA. 4) our methods w/o MLCL operation. 5) Full MLCL with all carefully designed components. All the variants are trained on FF++ and tested on FF++ and Celeb-DF.

Effect of different components. All quantitative results are reported in Table 4. The comparison between variant 1 and variant 2 shows that the employment of high-frequency information boosts the performance in both intra-dataset and cross-dataset settings. It is worth noting that by adding MRGA, the AUCs considerably increase to 97.84% and 73.01% in the FF++ and Celeb-DF datasets, which demonstrates the merit of the attention map learned by mutual representation. Comparing variant 3 with variant 4 shows that employing the GEM improves 0.37% and 2.2% in within-dataset and cross-dataset evaluation, due to the exploration of the global long-range inconsistency information. In variant 5, the baseline model equipped with all the proposed components achieves the best performance, 99.40% and 77.53% AUCs for in-dataset and cross-dataset testing. The improved performance indicates that the MLCL operation facilitates the discrimination between original features and artifact features and plays a significant role in generalized deepfake detection.

Effect of different levels of MMLCL. To evaluate the effectiveness of each level in the multi-scale mutual local consistency learning, we split each level of local consistency learning separately and conduct a series of experiments for verification. As shown in Table 5, employing single-scale local consistency learning leads to similar AUC but causes performance degradation compared to other variants (i.e. multi-scales), as the forgery clues captured by consistency maps with features of one scale size are insufficient. Due to the shallow-level

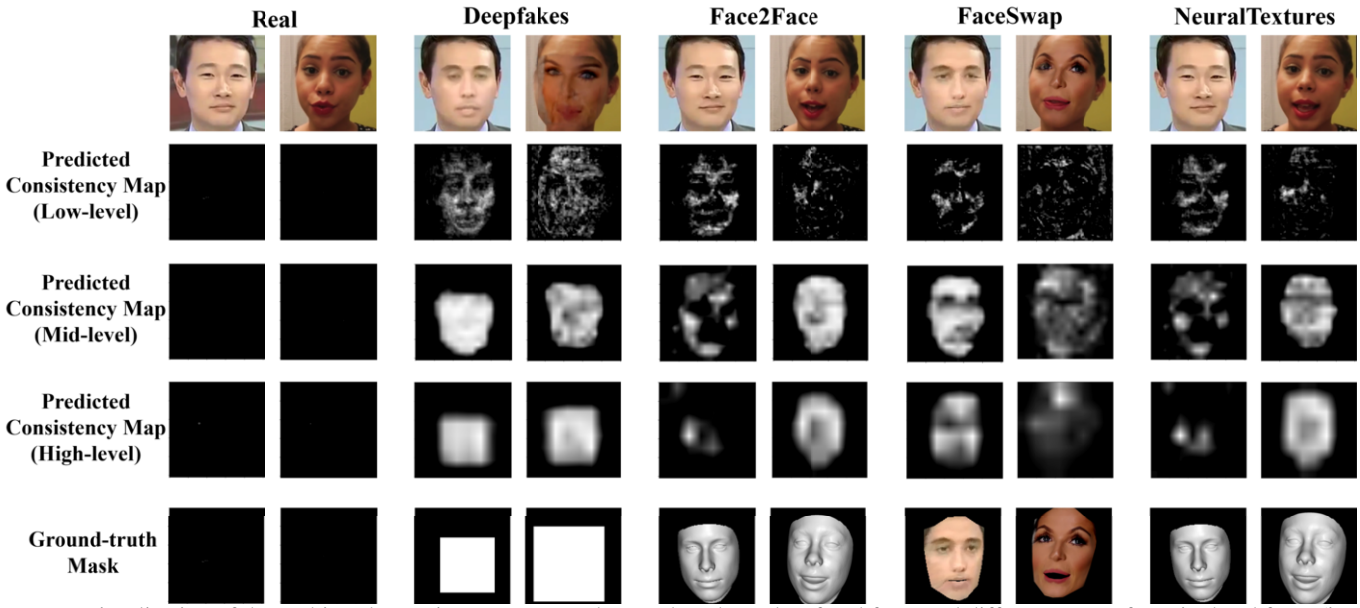

Figure 4: Visualization of the multi-scale consistency maps and ground-truth masks of real faces and different types of manipulated faces (i.e. Deepfakes, Face2Face, FaceSwap, NeuralTextures).

Table 4: Ablation study on FF++(HQ) and Celeb-DF with different model components in terms of AUC(%).

ID	Freq	GEM	MLCL	MRGA	Train set	Test set	
						FF++	Celeb-DF
1	–	–	–	–	FF++	96.65	71.20
2	✓	–	–	–	FF++	97.12	72.84
3	✓	–	–	✓	FF++	97.84	73.01
4	✓	✓	–	✓	FF++	98.21	75.21
5	✓	✓	✓	✓	FF++	**99.40**	**77.53**

features with high resolution being helpful for subtle differences localization, it performs slightly better than the other two single-scale local consistency learning with 0.35%, 0.78% AUC gains on FF++. By combining different levels of local consistency learning, we can observe that the performance is considerably improved by providing more inconsistency information from multi-scale features. This is because the consistency maps produced at different levels excavate diverse forged regions, which facilitates learning the discriminative representation. The performance reaches the peak for intra and cross-dataset testing when all levels of consistency learning are employed and is achieved at 99.40%, 77.53%, and 89.68% for AUC on FF++, Celeb-DF, and DFD. This shows the capability of MMLCL for promoting consistency learning across multi-scale features.

Table 5: Ablation study on the effect of different level local consistency learning of MMLCL on intra-dataset evaluation (FF++(HQ)) and cross-dataset evaluation (Celeb-DF, DFD, DFDC) in terms of AUC(%).

Module	Low-level	Mid-level	High-level	Train set	Test Set (AUC (%))		
					FF++	Celeb-DF	DFD
MMLCL *w/*	✓			FF++	96.02	72.29	86.60
		✓		FF++	95.67	72.18	86.51
			✓	FF++	95.24	72.25	86.46
	✓	✓		FF++	97.96	75.68	88.32
	✓		✓	FF++	97.86	76.02	88.29
		✓	✓	FF++	97.24	75.17	88.16
	✓	✓	✓	FF++	**99.40**	**77.53**	**89.68**

Choice of hyper-parameter. We evaluate the effectiveness of using different filter ratios. Conceptually, smaller radius high-pass filters highlight more high-frequency information (e.g. edges and textures) in the face image. However, primarily focusing on local details and textures limits the ability of the high-pass filter to capture and enhance features in the image. Moreover, small radius filters are sensitive to noise and minor variations in pixel value, which may amplify noise in the image, resulting in a noisy appearance and inducing image quality degradation. Thus, we train the MCLC using frequency input generated with different filter radios and conduct intra-dataset evaluation and cross-dataset evaluation on FF++, and Celeb-DF, respectively. Table 6 shows the results of different radius ratios of the high-pass filter in the frequency stream to indicate the suitable β for our proposed method. Additionally, to validate the advantage of our proposed MMLCL, we experiment with different combinations of λ_1 and λ_2 in Eq. 11, as shown in Table 7. We can observe that the best performance is achieved by training with $\lambda_1 = \lambda_2 = 1$. Especially, the performance on Celeb-DF gains 1.55% in AUC compared to the second-best results. The results prove that balancing λ_1 and λ_2 is beneficial to improve the generalization ability.

Table 6: Ablation study on the effect of different ratios of the high-pass filter in intra-dataset(FF++) and cross-dataset settings in terms of AUC (%).

Filter Ratio (β)	FF++	Celeb-DF
0.2	99.00	77.21
0.3	**99.40**	**77.53**
0.5	98.55	76.35
0.7	98.10	74.90

Table 7: Ablation study on the effect of different combinations of $\mathcal{L}_{cls}$ and $\mathcal{L}_{consis}$ on FF++, Celeb-DF, and DFDC datasets.

Method	Hyper-parameter	Backbone	Train Set	Test Set AUC(%)		
				FF++	Celeb-DF	DFDC
MLCL	λ_1=0.7, λ_2=0.3	EN-B2	FF++	99.05	74.38	74.50
	λ_1=1, λ_2=1			**99.40**	**77.53**	**76.21**
	λ_1=0.3, λ_2=0.7			99.12	75.98	74.32

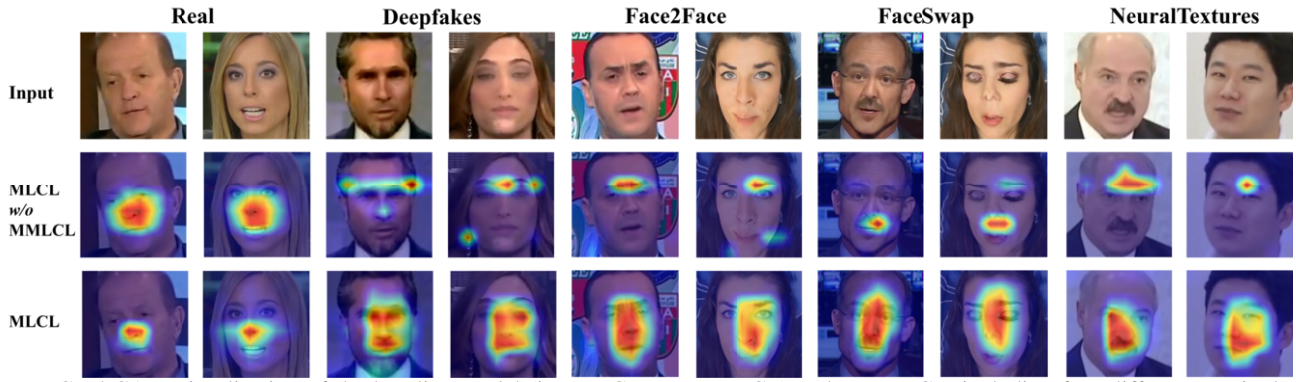

Figure 5: Grad-CAM visualization of the baseline model (i.e. MLCL *w/o* MMLCL) and our MLCL, including four different manipulation methods in FF++..

4.4 Visualization

Interpretability of consistency maps. For an intuitive understanding of the multi-scale mutual local consistency learning, we visualize some examples of the predicted consistency maps along with the corresponding input for the inconsistent region demonstration, as shown in Fig. 4. We can observe that the predicted multi-scale consistency maps generated by MMLCL are blank for real images and are adequately close to the ground-truth manipulation masks for fake face images. The blank maps indicate that the spatial and frequency features are consistent. However, taking the Deepfakes method which manipulates the entire face region as an example, all level consistency maps illustrate the central face region as inconsistent areas, which matches the forged region. Moreover, we also calculate the average value and obtain 0.963, 0.976, and 0.986 for multi-scale consistency maps from real face images. However, for some faces with evident signs of forgery (e.g. FaceSwap), the feature dramatically varies in unnatural regions. High-level feature maps of RGB and frequency stream would be attentive to the unnatural transition regions for this kind of forged faces because the features in unnatural transition regions are significantly variant compared to other regions. So, the predicted consistency maps show a smaller area of inconsistency. As discussed, these statistical results demonstrate that our proposed MMLCL effectively learns the intrinsic relationship between the RGB and frequency domains and predicts all entries in the consistency maps with satisfactory confidence.

Visualization for class activation. To explore the region of interest of our classifier for various manipulation methods, we utilize Grad-CAM [31] to generate class activation maps. The warm color indicates the areas that the classifier strongly responds to and makes the prediction based on. We compare our proposed MLCL with the model without equipping MMLCL module which we call it as baseline model in the following context. As shown in Fig 5, both two models have almost no response to the real faces. However, our method can capture a comprehensive face region for manipulated faces while the baseline model only highlights a small portion of forged areas or even fails to capture proper forged regions. For instance, the heatmaps for Face2Face (F2F) and NeuralTextures (NT) produced by baseline are prominent around the eyes region, where NT mainly modifies the mouth region and F2F changes the expression of the facial area. On the other hand, the focused areas in the heatmaps of MLCL are matched with the manipulation mask, which indicates that MLCL is capable of locating the forged regions with multi-scale local consistency maps. The results validate the effectiveness of MCLC for local consistency learning.

We also conduct the cross-dataset evaluation on analyzing the decision region of the detector for unseen manipulation methods and provide visualization results, as illustrated in Fig 6. The heatmaps generated by baseline model focus on the central facial region for all original images and fail to locate the modified area for forged face images, owing to a lack of inconsistency information for capturing discriminative feature regions. In contrast, our MLCL generates distinguishable heatmaps for both pristine and fake faces, where the salient regions accurately match the manipulated areas. The results indicate that the local feature consistency learning is generalizable to unseen forgery techniques.

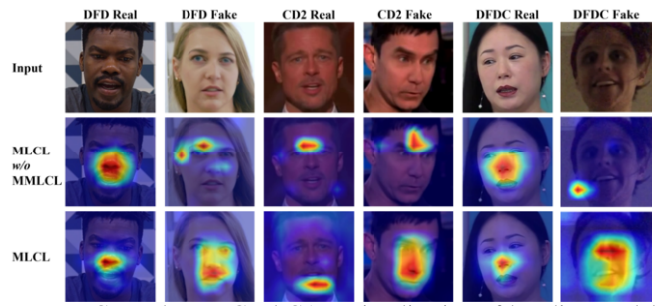

Figure 6: Cross-dataset Grad-CAM visualization of baseline model (i.e. MLCL *w/o* MMLCL) and our MLCL on DFD, Celeb-DF (CD2) and DFDC.

5 Conclusion

In this paper, we propose a Mutual Local Consistency Learning framework (MLCL) for face manipulation detection. A novel MMLCL is designed to localize forged regions by modelling the relationship of local regions on the feature maps of the RGB and frequency domains and utilizing the cue of the inconsistency between original and artifact features. Besides, we employ a self-supervised strategy to guide the MMLCL with adaptive ground-truth manipulation masks. Moreover, MGRA is introduced to collaboratively combine features of two streams and learn a comprehensive attention map as guidance for further predicting forged regions. Extensive experiments show that MLCL is competitive against state-of-the-art methods and demonstrates the inter-dataset generalization ability on challenging face forgery datasets.

References

[1] Deepfakes github. https://github.com/deepfakes/faceswap, 2019. Accessed on:15/12/2021.

[2] Faceswap github. https://github.com/MarekKowalski/FaceSwap, 2019. Accessed on:15/12/2021.

[3] 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)*, pages 1–7. IEEE, 2018.

[4] 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*, volume 35, pages 1081–1088, 2021.

[5] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[6] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020.

[7] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020.

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

[9] 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*, pages 9468–9478, 2022.

[10] N. Dufour and A. Gully. Contributing data to deepfake detection research. *Google AI Blog*, 1(2):3, 2019.

[11] 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*, pages 20270–20280, 2022.

[12] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020.

[13] Q. Gu, S. Chen, T. Yao, Y. Chen, S. Ding, and R. Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 735–743, 2022.

[14] J. Guan, H. Zhou, Z. Hong, E. Ding, J. Wang, C. Quan, and Y. Zhao. Delving into sequential patches for deepfake detection. *Advances in Neural Information Processing Systems*, 35:4517–4530, 2022.

[15] X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, and X. Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023.

[16] Z. Hu, H. Xie, Y. Wang, J. Li, Z. Wang, and Y. Zhang. Dynamic inconsistency-aware deepfake video detection. In *IJCAI*, pages 736–742, 2021.

[17] B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang, and D. Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2023.

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] 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*, pages 6458–6467, 2021.

[20] 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*, pages 5001–5010, 2020.

[21] 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)*, pages 1–7. IEEE, 2018.

[22] Y. Li, P. Sun, H. Qi, and S. Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, Seattle, WA, United States, 2020.

[23] 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*, pages 772–781, 2021.

[24] A. Luo, C. Kong, J. Huang, Y. Hu, X. Kang, and A. C. Kot. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19:1168–1182, 2023.

[25] 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*, pages 16317–16326, 2021.

[26] O. Mayer and M. C. Stamm. Exposing fake images with forensic similarity graphs. *IEEE Journal of Selected Topics in Signal Processing*, 14 (5):1049–1064, 2020.

[27] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019.

[28] 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*, pages 86–103. Springer, 2020.

[29] Y. Rao and J. Ni. Self-supervised domain adaptation for forgery localization of jpeg compressed images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15034–15043, 2021.

[30] 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*, pages 1–11, 2019.

[31] 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*, pages 618–626, 2017.

[32] 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*, volume 36, pages 2316–2324, 2022.

[33] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[34] 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*, pages 2387–2395, 2016.

[35] T. Wang and K. P. Chow. Noise based deepfake detection via multi-head relative-interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14548–14556, 2023.

[36] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[37] J. Zhang and J. Ni. Domain-invariant feature learning for general face forgery detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2321–2326. IEEE, 2023.

[38] 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*, pages 2185–2194, 2021.

[39] 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*, pages 15023–15033, 2021.

[40] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pages 1831–1839. IEEE, 2017.