LiSiam: Localization Invariance Siamese Network for Deepfake Detection

Jian Wang, Yunlian Sun, and Jinhui Tang, *Senior Member, IEEE*

***Abstract*— Advances in facial manipulation technology have led to increasing indistinguishable and realistic face swap videos, which raises growing concerns about the security risk of deepfakes in the community. Although current deepfake detectors can gain promising performance when handling high-quality faces under within-database settings, most detectors suffer from performance degradation in cross-database evaluation. Moreover, when test faces' quality is different from training faces, the performance degrades even under within-database settings. To this end, we propose a novel Localization invariance Siamese Network (LiSiam) to enforce localization invariance against different image degradation for deepfake detection. Specifically, our Siamese network-based feature extractor takes the original image and the corresponding quality-degraded image as pairwise inputs and outputs two segmentation maps. A localization invariance loss is further proposed to impose localization consistency between the two segmentation maps. In addition, we design a Mask-guided Transformer to capture the co-occurrence between the forgery region and its surroundings. Finally, a multi-task learning strategy is utilized to obtain a robust and discriminative feature representation and jointly optimize multiple objective functions (i.e., segmentation, classification, and localization invariance losses) in an end-to-end manner. Experimental results on two public datasets, i.e., FaceForensics++ and Celeb-DF, demonstrate the superior performance of our proposed method to state-of-the-art methods.**

***Index Terms*— Beepfake detection, localization invariance, Siamese network, attention mechanism, multi-task learning.**

I. Introduction

In RECENT years, with the rapid development of deep learning-based generative models, manipulated media content has swept the world, especially face swap videos and images. Although face swap videos are mainly used for entertainment purposes, the increasingly realistic and indistinguishable fake videos bring huge potential security risks including financial fraud, trust crisis, etc. Therefore, the detection of fake videos has attracted growing concern in the community. In present period, many deepfake detectors [1], [2] have achieved satisfactory performance in uncompressed within-database experiments [3]. However, most methods suffer from a sharp performance drop when conducting cross-quality evaluation in within-database experiments. Moreover, the performance of existing methods drops drastically in cross-database evaluation as it is difficult to cope with unseen forgery methods. Therefore, it is crucial to develop powerful and effective anti-deepfake tools to detect facial forgery against image quality degradation and unseen forgery methods.

To mitigate performance degradation due to image degradation (e.g., compression, blurring, etc.) and unseen forgery methods, one common solution is to utilize various data augmentation strategies for increasing the diversity of training data [4]. However, data augmentation strategies might weaken or even erase some forgery traces, which makes it difficult for the detector to accurately locate forgery traces. Recently, some research attempts to investigate frequency-domain clues to capture intrinsic traces for forgery detection [2], [5], [6]. For example, Qian *et al.* [2] utilized frequency-aware decomposition and local frequency statistics to capture forgery traces in the frequency domain. Li *et al.* [5] designed a single-center loss to guide frequency-aware discriminative feature learning for extracting intrinsic feature representations. However, forgery traces in the frequency domain can be easily weakened by designing a special frequency-domain regularization term for the generator loss [7]. In order to catch slight forgery traces, some researchers focus on the detailed information of local regions. For instance, Chai *et al.* [8] attempted to use a patch-based detector to focus on local artifacts rather than global semantics for improving generalization. These patch-based methods focus on local details, however, do not further explore the fine-grained detection of pixel-level forgery traces.

To overcome these issues, we need to enhance the model to capture more robust and effective forgery clues, especially pixel-level localization of facial forgery. One possible solution is to construct a blending boundary between manipulated regions and original regions, which utilizes the inconsistency of the two regions in manipulated images. For example, Qian *et al.* [9] proposed Face X-ray to reveal the blending boundary between the altered face and background image for locating manipulated traces. However, because the blending boundary is easy to be removed by image compression or image blur, the Face X-ray method suffers from performance degradation when dealing with low-quality data. In addition, the pixel-level segmentation mask is also utilized for the localization of facial

Manuscript received 9 December 2021; revised 30 March 2022 and 8 June 2022; accepted 8 June 2022. Date of publication 27 June 2022; date of current version 5 July 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102002, in part by the National Natural Science Foundation of China under Grant 61732007 and Grant 62076131, and in part by the Open Funding Project of the State Key Laboratory of Communication Content Cognition under Grant 20K03. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Domingo Mery. *(Corresponding authors: Jinhui Tang; Yunlian Sun.)*

Jian Wang and Jinhui Tang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China, and also with the State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing 100733, China (e-mail: wj92@njust.edu.cn; jinhuitang@njust.edu.cn).

Yunlian Sun is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: yunlian.sun@njust.edu.cn).

Digital Object Identifier 10.1109/TIFS.2022.3186803

forgery [1], [10], [11]. For example, Dang *et al.* [10] proposed a novel attention mechanism to highlight manipulated regions with segmentation mask supervision. Shang *et al*. [11] proposed a Pixel-Region Relation Network to separate manipulated regions from authentic regions and then measure the inconsistency between the two regions for deepfake detection. However, slight forgery traces can be also easily removed by quality-degradation methods, which results in inaccurate localization on compressed data. In addition, Zhang *et al.* [12] considered that CNNs are inclined to focus on the most discriminative part of the object for recognition. Nevertheless, some studies [13]–[15] have shown that focusing on the entire object region rather than just the most discriminative part of the object can improve the generalization performance of the model. It should be noted that current deepfake detectors usually ignore the above findings. Considering that no matter how the image is compressed or blurred, the tampered regions of the fake image remain unchanged. Accordingly, it is significant to focus on robust localization of manipulated regions against JPEG compression or image blur.

In this paper, we attempt to enforce localization consistency across images with different image quality degradation and exploit pixel-level localization for improving generalization performance. The consistency of forgery localization is expected to be robust to different image quality degradation for deepfake detection. To this end, we propose Localization Invariance Siamese Networks (LiSiam) for imposing localization invariance constraints between segmentation maps corresponding to input images of different qualities. Specifically, we firstly utilize data augmentation (e.g., JPEG Compression, Gaussian Blurring, Resizing, etc.) to obtain degraded images. Siamese networks then take both the raw image and quality-degraded image as inputs and output two segmentation maps. A novel localization invariance loss function is proposed to enforce localization invariance across images with different degrees of degradation. Moreover, a mask-guided transformer (MT) is designed to capture co-occurrence between the suspected manipulated region and its surroundings. Next, a multi-layer perceptron (MLP) head is employed to take co-occurrence features as inputs and output the final binary decision. Finally, we use a multi-task learning strategy to optimize the network in an end-to-end manner.

In summary, the main contributions of this work are fivefold:

- A novel framework, LiSiam, is proposed to implement pixel-level localization of facial forgery and enforce localization invariance for improving performance in both cross-database and cross-quality evaluation.
- We propose a mask-guided transformer to finely capture the co-occurrence between manipulated regions and their surroundings. The co-occurrence is expected to contain rich and distinctive information of facial forgery and can better guide the model to capture forgery traces.
- To improve forgery localization of low-quality images, we design a localization invariance loss to enforce the localization consistency between the raw image and quality-degraded image.
- A multi-task learning strategy is adopted to optimize the network in an end-to-end manner.
- We conduct extensive experiments on two public datasets to evaluate the performance of the proposed method. Experimental results show the superior performance of our proposed method to the state-of-the-art methods.

The rest of the paper is organized as follows: Sec. II reviews related work of Deepfake Detection, Siamese Network, and Manipulation Localization. The proposed LiSiam architecture is detailed in Sec. III. Experimental setup, results, and analysis are described in Sec. IV. Finally, the paper is concluded in Sec. V.

II. Related Work

In this section, we give a brief review to Deepfake Detection, Siamese Network, and Manipulation Localization.

A. Deepfake Detection

Previously, some researchers focused on investigating camera characteristics and hand-crafted features to capture forged traces for tampered face detection [16]–[18]. For instance, Zhou *et al.* [16] captured low-level fingerprint-like camera characteristics to detect tampered faces. In [17], inconsistency of head poses is estimated via facial landmarks to detect deepfakes. Although these methods achieved sound performance at that time, they could not meet the current requirements of deepfake detection, especially in the face of increasingly advanced deepfake technologies. Recently, there are some approaches attempting to explore information like frequency-aware clues [2], [5], [7] and pixel-level segmentation map [9]–[11] for deepfake detection. Durall *et al.* [7] conducted experiments to prove that common up-convolution operations could result in high-frequency distortions in CNN-generated images. The frequency-based detectors [2], [5] captured artifacts in the frequency domain and not just in the RGB domain. However, these detectors could be easily bypassed by designing a special frequency-domain regularization term for the generator loss. Zhao *et al.* [19] proposed a multi-attentional network to capture fine-grained information for deepfake detection. However, this method lacks powerful supervision and only depends on the attention mechanism to detect artifacts, which can not well capture slight forgery traces for quality-degraded forgery detection.

Besides, there are also some methods focusing on exploiting spatial or temporal clues from local facial regions [20], such as eye blinking [18], lip movement [21], and so on. However, these early methods struggled to detect the increasingly realistic deepfakes. To better focus on local details and the global context of face forgery, some researchers [22], [23] attempted to combine Vision Transformers with CNNs for deepfake detection. These methods achieved promising generalization performance in cross-database evaluation, but the combination of CNNs and Vision Transformers involves a large number of parameters resulting in high computational complexity. Besides most of them also ignored the robustness of the detector against quality-degraded forgery. At present, some researchers began to pay attention to this valuable

and meaningful issue [11]. However, no specific approach has been proposed to systematically address the problem of quality-degraded face forgery detection. In this work, we design a localization invariance loss function to allow the model to learn effective feature representation against the quality-degraded images.

B. Siamese Network

Siamese network is a neural network architecture consisting of twin subnetworks with shared weights. Siamese network was widely used in the computer vision field, including image recognition [24], object tracking [25], semantic segmentation [26], unsupervised visual representation learning [27], and others. Wang *et al.* [26] proposed a self-supervised equivariant attention mechanism (SEAM) which is implemented by a Siamese network with equivariant cross regularization loss. SEAM was designed to enforce Class Activation Mapping (CAM) predicted from input images with various scales to be consistent. Chen *et al.* [27] proposed a simple Siamese network with simple designs to model invariance for learning meaningful representations. Mayer *et al.* [28] proposed Siamese Networks-based forensic similarity network to determine whether two image patches contain the same or different manipulated characteristics. Besides, Han *et al.* [1] proposed a state-of-the-art Co-teaching approach similar to Siamese Networks for robust training of deep networks with noisy labels. The Co-teaching approach can simultaneously train two networks with non-weight sharing scheme and enable them to teach each other in each mini-batch. Inspired by the above attempts, we propose a unified architecture based on Siamese Networks that model localization invariance for cross-quality deepfake detection.

C. Manipulation Localization

Previously, there are some approaches trying to perform the localization of tampered regions for image forgery detection [29]–[31]. Zhou *et al.* [30] proposed a two-stream Faster R-CNN network to localize tampered regions by discovering noise inconsistency between tempered and authentic regions. Bappy *et al.* [31] utilized both frequency and spatial domain features to localize tampered regions. These approaches were used to detect image forgeries (e.g., copy-move forgery), not face forgeries. For the deepfake detection task, some researchers focused on pixel-level fake face segmentation to capture manipulation traces. For example, Dang *et al.* [10] employed an attention mechanism to automatically detect forged traces in face images. Although their method obtained the promising performance, the generation of groundtruth (GT) manipulation masks needs to be improved for better guiding the localization of facial forgery. In this work, we utilize a novel GT segmentation mask to better guide the model to focus on the forged region. Besides, Shang *et al.* [11] proposed to segment tampered regions using a spatial attention mechanism and then use multiple metrics to measure the inconsistency between authentic regions and tampered regions. Li *et al.* [9] proposed a Face X-ray method to locate the boundary trace between tampered regions and authentic regions. However, this method suffered performance degradation due to the change of image quality, because image compression or image blur can erase some boundary traces.Yun *et al.* [13] claimed that the generalization performance of the model can be improved by focusing on the entire object region rather than just the most discriminative part of the object. To this end, we design a localization invariance loss to enforce forgery localization invariance across different image degradation for fine-grained localization.

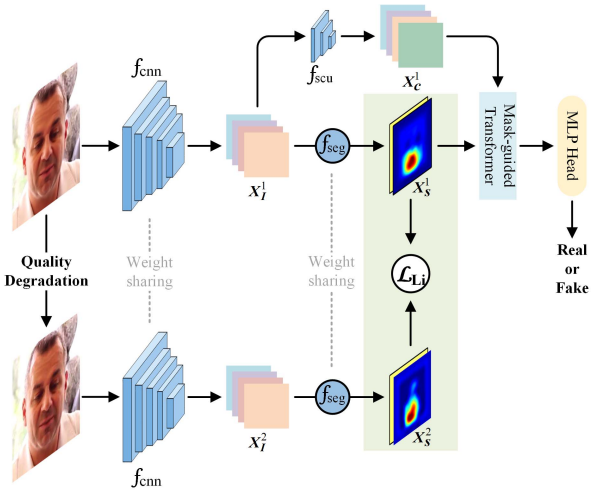

Fig. 1. Overview of the proposed method. The proposed LiSiam network takes the original image and quality-degraded image as the inputs of the feature extractor (f_{cnn}) based on Siamese networks. Both output feature maps of f_{cnn} are fed to the weight-sharing segmentation decoder (f_{seg}). The output segmentation maps from f_{seg} are utilized to enforce localization consistency by localization invariance loss ($\mathcal{L}_{\text{Li}}$). The proposed mask-guided transformer takes feature maps $\boldsymbol{X}_c^1$ and segmentation maps $\boldsymbol{X}_s^1$ as inputs and outputs discriminative co-occurrence features. Finally, the co-occurrence features are used as inputs to the classifier based on MLP to perform the binary classification.

III. The Proposed Method

In this section, we first introduce Localization invariance Siamese (LiSiam) architecture in Sec. III-A and then present Mask-guided Transformer, designed loss function and multi-task learning strategy in Sec. III-B, Sec. III-C, and Sec. III-D, respectively. Finally, we describe our implementation details in Sec. III-E.

A. LiSiam Architecture

To detect facial forgery, we propose a novel framework to enforce localization invariance across images of different compression and blur degrees, as shown in Fig. 1. The proposed LiSiam networks mainly consist of two sub-network branches which are designed to process the original input image and the quality-degraded input image, respectively. Each branch has a Feature Extractor (f_{cnn}) and a Segmentation Decoder (f_{seg}). A CNN-based Feature Extractor is used to extract a feature representation $\boldsymbol{X}_{\text{I}}^n \in \mathbb{R}^{C\times H_0\times W_0}$ ($n = 1, 2$ indicates n^{th} branch) from the input image $\boldsymbol{I}^n$. The Segmentation Decoder is used to obtain pixel-level localization of the forged region. The Siamese Segmentation Decoder outputs two segmentation maps which are used to calculate the localization invariance

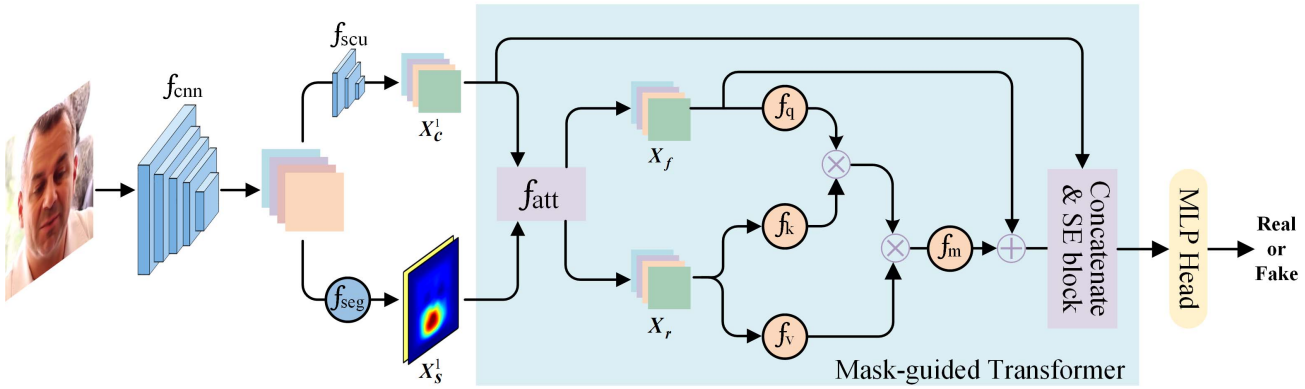

Fig. 2. The structure of Mask-guided Transformer. "⊗" and "⊕" denote matrix multiplication and element-wise sum, respectively.

loss. With the supervision of this loss, we can train our networks to extract more robust features against image quality degradation. Apart from f_{cnn} and f_{seg}, the first branch includes also a Separable Convolution Unit (f_{scu}), Mask-guided Transformer (f_{mt}), and Multilayer Perceptron (MLP) Head (f_{mlp}). The Separable Convolution Unit consists of separable convolutions [32], batch normalization, ReLU, and dropout layers, which is used to further encode feature representation for the classification task. The Mask-guided Transformer takes feature maps and segmentation maps as inputs and extracts co-occurrence features. The prediction operation is implemented by the MLP head with two linear layers. During inference, since our proposed LiSiam is a weight-sharing siamese network, only the first branch is utilized to make the final binary decision $\boldsymbol{\alpha}$ for a test image, which can be formulated as:

$$\boldsymbol{\alpha} = f_{\text{mlp}}(f_{\text{mt}}(f_{\text{seg}}(f_{\text{cnn}}(\boldsymbol{I})), f_{\text{scu}}(f_{\text{cnn}}(\boldsymbol{I})))). \tag{1}$$

B. Mask-Guided Transformer

The attention mechanism has achieved great success in both natural language processing [33], [34] and computer vision [35], [36]. Therefore, we design a novel Mask-guided Transformer to extract robust co-occurrence features for deepfake detection, as shown in Fig. 2. The Mask-guided Transformer is an "attention on attention" operation. One attention is used to generate pixel-level predictions of the suspected forged region. We then add another attention operation to extract robust attended features by computing the similarity between fake features and real features. Specifically, we first extract feature representations for the suspected fake regions and real regions by an attention operation, respectively:

$$\boldsymbol{X}_{\text{I}}^n = f_{\text{cnn}}(\boldsymbol{I}^n), \tag{2}$$
$$\boldsymbol{X}_{\text{s}}^n = f_{\text{seg}}(\boldsymbol{X}_{\text{I}}^n), \tag{3}$$
$$\boldsymbol{X}_{\text{c}}^1 = f_{\text{scu}}(\boldsymbol{X}_{\text{I}}^1), \tag{4}$$
$$\boldsymbol{X}_{\text{f}}, \boldsymbol{X}_{\text{r}} = f_{\text{att}}(\boldsymbol{X}_{\text{s}}^1, \boldsymbol{X}_{\text{c}}^1), \tag{5}$$

where $\boldsymbol{X}_s^n \in \mathbb{R}^{2\times H_0\times W_0}$ denotes segmentation maps obtained by the Segmentation Decoder. $\boldsymbol{X}_c^1 \in \mathbb{R}^{C\times H_1\times W_1}$ represents feature maps extracted by the Separable Convolution Unit. f_{att} is a soft attention operation which aims to separate features of forged regions from those of real regions in $\boldsymbol{X}_c^1$ using segmentation maps. f_{att} operation is defined as follows:

$$\boldsymbol{X}_{\text{s}'}^1 = f_{\text{bi}}(\text{softmax}(\boldsymbol{X}_{\text{s}}^1)), \tag{6}$$
$$\boldsymbol{X}_{\text{r}} = \boldsymbol{X}_{\text{c}}^1 \odot \boldsymbol{X}_{\text{s}'}^1[0, :, :], \tag{7}$$
$$\boldsymbol{X}_{\text{f}} = \boldsymbol{X}_{\text{c}}^1 \odot \boldsymbol{X}_{\text{s}'}^1[1, :, :], \tag{8}$$

where $\odot$ indicates element-wise multiplication. f_{bi} is a bilinear interpolation operation, which is used to obtain attention maps $\boldsymbol{X}_{\text{s}'}^1$ of the same spatial size as $\boldsymbol{X}_{\text{c}}^1$. $\boldsymbol{X}_{\text{r}} \in \mathbb{R}^{C\times H_1\times W_1}$ and $\boldsymbol{X}_{\text{f}} \in \mathbb{R}^{C\times H_1\times W_1}$ are real and suspected fake features, respectively.

Next, we use one 1×1 convolution to map the suspected fake feature into $\boldsymbol{Q}$ (Query), and use two 1×1 convolutions to map the real feature into $\boldsymbol{K}\&\boldsymbol{V}$ (Key and Value).

$$\boldsymbol{q}_i = f_{\text{q}}(\boldsymbol{x}_{f,i}), \tag{9}$$
$$\boldsymbol{k}_j = f_{\text{k}}(\boldsymbol{x}_{r,j}), \tag{10}$$
$$\boldsymbol{v}_j = f_{\text{v}}(\boldsymbol{x}_{r,j}), \tag{11}$$

where $\boldsymbol{q}_i \in \boldsymbol{Q}$, $\boldsymbol{k}_j \in \boldsymbol{K}$ and $\boldsymbol{v}_j \in \boldsymbol{V}$ are the i^{th} Query, j^{th} Key and Value pairs, respectively. $\boldsymbol{x}_{\text{f},i}$ is the i^{th} feature position of the fake feature $\boldsymbol{X}_{\text{f}}$. $\boldsymbol{x}_{\text{r},j}$ is the j^{th} feature position of the real feature $\boldsymbol{X}_{\text{r}}$. $f_{\text{q}}(\cdot)$, $f_{\text{k}}(\cdot)$, and $f_{\text{v}}(\cdot)$ are the corresponding 1×1 convolutions.

Then, we obtain the transformed feature $\boldsymbol{y}_i$ (i.e., the i^{th} feature position of $\boldsymbol{Y}$) based on $\boldsymbol{v}_j$ and the attention weight $a_{i,j}$ computed by the dot product similarity as follows:

$$a_{i,j} = \text{softmax}(\boldsymbol{q}_i^{\text{T}}\boldsymbol{k}_j), \tag{12}$$
$$\boldsymbol{y}_i = f_{\text{m}}(a_{i,j}\boldsymbol{v}_j), \tag{13}$$

where $f_{\text{m}}(\cdot)$ is implemented by 1×1 convolutions.

After that, we concatenate original classification features $\boldsymbol{X}_{\text{c}}^1$ and the transformed features $\boldsymbol{Y}$ (added with $\boldsymbol{X}_f$). Finally, features are re-weighted along the channel dimension using a Squeeze-and-Excitation (SE) [37] block (f_{se}). The process described above is defined as follows:

$$\boldsymbol{Z} = f_{\text{se}}(\text{concat}(\boldsymbol{Y} + \boldsymbol{X}_f, \boldsymbol{X}_c^1)). \tag{14}$$

C. Localization Invariance Loss

In the deepfake detection task, obtaining a discriminative and robust feature representation is crucial for performance improvement. Current deep learning-based methods usually

utilize cross-entropy loss or center loss [5] to optimize networks. Although these methods based on the above loss achieve promising performance when dealing with faces of seen quality, they suffer performance degradation on unseen quality (i.e., cross-quality) evaluation. To this end, we propose a localization invariance loss to allow the model to learn robust feature representations for cross-quality deepfakes. Localization invariance loss aims to enforce forgery localization invariance across different image degradation and improve performance for cross-quality deepfake detection. The localization invariance loss is defined as:

$$\mathcal{L}_{\text{li}} = \|\boldsymbol{X}_{\text{s}}^{1} - \boldsymbol{X}_{\text{s}}^{2}\|_{1}. \tag{15}$$

D. Multi-Task Learning Strategy

The multi-task learning strategy aims to perform joint learning of multiple tasks by sharing information among multiple related tasks [38]. In our proposed LiSiam, three learning tasks are involved: (1) forgery localization, (2) binary classification, and (3) localization invariance constraint. To this end, we design three loss functions to optimize the network including segmentation loss, classification loss, and localization invariance loss. Deepfake detection is essentially a binary classification task. Same as most previous work, we utilize the general cross-entropy loss to optimize the parameters of the networks. Furthermore, we also use the general cross-entropy loss as our segmentation loss for the segmentation task. The two loss functions are defined as follows:

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{ce}}(\boldsymbol{\alpha}, \boldsymbol{\alpha}_{\text{gt}}), \tag{16}$$

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{ce}}(\boldsymbol{X}_{\text{s}}^{1}, \boldsymbol{X}_{\text{gt}}), \tag{17}$$

where $\boldsymbol{\alpha}_{\text{gt}}$ is the ground-truth classification label and $\boldsymbol{X}_{\text{gt}}$ refers to pixel-level ground-truth. The predicted score $\boldsymbol{\alpha}$ is obtained as:

$$\boldsymbol{\alpha} = f_{\text{mlp}}(f_{\text{aap}}(\boldsymbol{Z})) \tag{18}$$

where f_{aap} is the Adaptive Average Pooling operation which aims to aggregate the feature maps.

To achieve joint learning of multiple tasks, we sum up losses from the above three tasks. The total loss of LiSiam is defined as:

$$\mathcal{L} = \lambda_{\text{li}}\mathcal{L}_{\text{li}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}, \tag{19}$$

where λ_{li}, λ_{cls}, and λ_{seg} are the weights representing the importance of each loss.

E. Implementation Details

Specifically, our feature extractor consists of the backbone network and the OCR [39] block. We take Xception as our backbone network, which is widely used in the deepfake detection task. To reduce both the computational complexity and the number of parameters in the network, we utilize only the first 12 blocks of Xception as our backbone network for feature extraction. Besides, to obtain finer segmentation maps for fine-grained detection, we remove the last MaxPool2d layer of Xception. Therefore, the modified backbone network produces output feature maps of size $1024 \times 37 \times 37$. The OCR block then takes the output feature maps of the backbone network as input and outputs feature maps of size $512 \times 37 \times 37$. The output dimension $2 \times H_0 \times W_0$ of the Segmentation Decoder is $2 \times 37 \times 37$. The dimension $C \times H_1 \times W_1$ of the feature map $\boldsymbol{X}_{\text{c}}^{1}$ is $512 \times 13 \times 13$. Segmentation loss, classification loss, and localization invariance loss are equally important to our task. Thus, to balance the three losses, λ_{li}, λ_{cls}, and λ_{seg} are empirically set to 1.0 in the loss function.

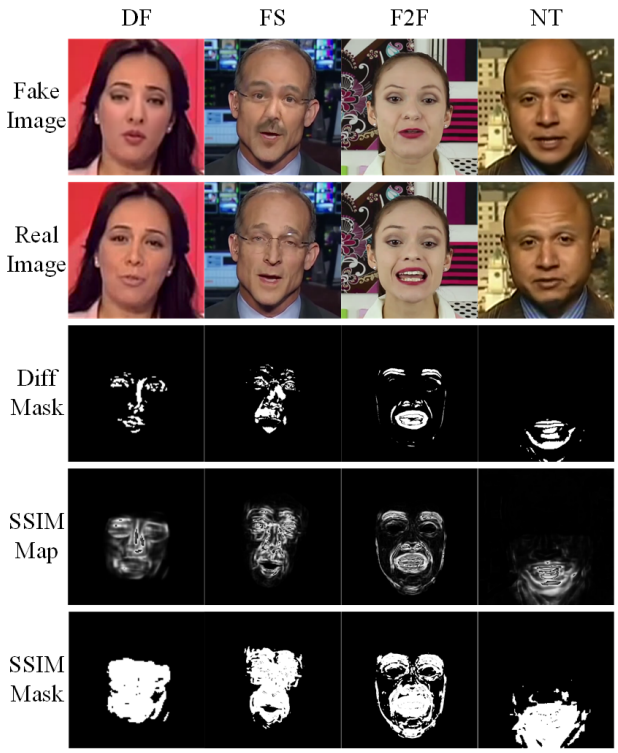

Fig. 3. Examples of different segmentation masks. Face images are from FF++ [3] with four forgery methods, i.e., DeepFakes(DF), FaceSwap (FS), Face2Face (F2F), and NeuralTextures (NT). The DIFF mask is the difference between the real and fake images. The SSIM map is generated by computing the structural similarity between the real and fake images. The SSIM mask is the binary image of the SSIM map.

The Structural Similarity Index Measure (SSIM) [40] has been widely used to measure the structural difference between two images. The pixel-by-pixel structural difference between the real face and the corresponding fake face can locate the forged region of the fake image. Therefore, in this work, we utilize the SSIM map as the groundtruth segmentation map, which better guides the model to focus on the forged region. Specifically, the original SSIM map multiplied by 255 is converted into a binary mask by empirically applying a threshold of 20. The obtained binary mask is used as the final groundtruth of the fake image. Besides, we use an all-zero map as the segmentation mask for the real image. As shown in Fig. 3, we compare our generated SSIM mask with difference-based (DIFF) mask [10]. It is indicated that our SSIM mask provides richer and more comprehensive details than the difference-based mask.

IV. EXPERIMENTS

In this section, we evaluate the performance of our proposed method on the FaceForensics++ (FF++) database [3] and the Celeb-DF database [41], [42].

A. Datasets and Settings

1) FaceForensics++: FaceForensics++ (FF++) is a large-scale public database, which is widely adopted in the deepfake detection task. The database consists of 1000 real videos from YouTube and 4000 fake videos manipulated by four methods: DeepFakes(DF) [43], FaceSwap (FS) [44], Face2Face (F2F) [45] and NeuralTextures (NT) [46]. Besides, each video in FF++ has three versions in terms of compression level: original version (RAW or C0), high-quality versions (HQ or C23), and low-quality version (LQ or C40).

2) Celeb-DF: Celeb-DF is a challenging database, which is widely used for evaluating the generalization performance of deepfake detectors. The Celeb-DF database has two versions: Celeb-DF-v1 [41] and Celeb-DF-v2 [42]. The Celeb-DF-v1 includes 408 real videos and 795 fake videos. The Celeb-DF-v2 consists of 890 real videos and 5,639 fake videos.

3) Data Preparation: To extract face data from the whole image, we use CenterFace [47] for detecting the face region. The face image is resized to 299×299 pixels as the input of the networks. On FF++, we sample only 30 frames per video for training. Compared with other methods [3], our methods need less training data. Following [3], we extract 100 frames and 100 frames per video for validation and test, respectively.

Besides, we utilize various degradation methods to obtain quality-degraded images, including JPEG Compression, Resizing, and Gaussian Blur. These methods are performed with a probability of 50% in random order. The quality-degraded images using different degradation methods are shown in Fig. 4.

4) Evaluation Metrics: To effectively evaluate the performance of our proposed method, we adopt two commonly used metrics in the deepfake detection task, including Accuracy (Acc) and Area Under the ROC Curve (AUC). We report our experimental results at the frame level by default and specify the fashion otherwise.

B. Training Details

We implement our proposed method on the PyTorch platform and use Stochastic Gradient Descent (SGD) to update the parameters of the networks. The initial learning rate and mini-batch size are set to 0.002 and 32, respectively. A poly learning rate policy with a power of 0.9 is employed to update the learning rate. To speed up the training of our network, we use ImageNet weights for the initialization of network weights. After weight initialization, we first freeze classification-related layers (i.e., Separable Convolution Unit, Mask-guided Transformer, and MLP Head) and only train segmentation modules (consisting of Feature Extractor and Segmentation Decoder) for about $146k$ iterations. Then, we freeze segmentation-related layers and only train classification-related layers for about $16k$ iterations. Finally, we train all layers of the whole network for $324k$ iterations.

C. Within-Database Evaluation

In this section, we compare our methods with prior state-of-the-art methods in within-database evaluation. And these methods are trained and tested using the same-quality images (e.g., trained on FF++ C23 and tested on FF++ C23). We report both frame-level and video-level results for fair comparisons on the FF++ database. Table I summarizes the frame-level results of 12 different state-of-the-art detectors in terms of accuracy and AUC. As can be seen, our method surpasses the state of the art in C23 (high quality) evaluation. Although our proposed method is trained with fewer training samples than the common setting [3], our methods still obtain the best results. For the more challenging C40 (low quality) scenario, our proposed method enhances the robustness against compressed data owing to localization invariance and greatly improves the detection performance. Specifically, the proposed LiSiam improves AUC from 87.26% to 91.44% in FF++ (C40) evaluation compared with the current state-of-the-art method [19].

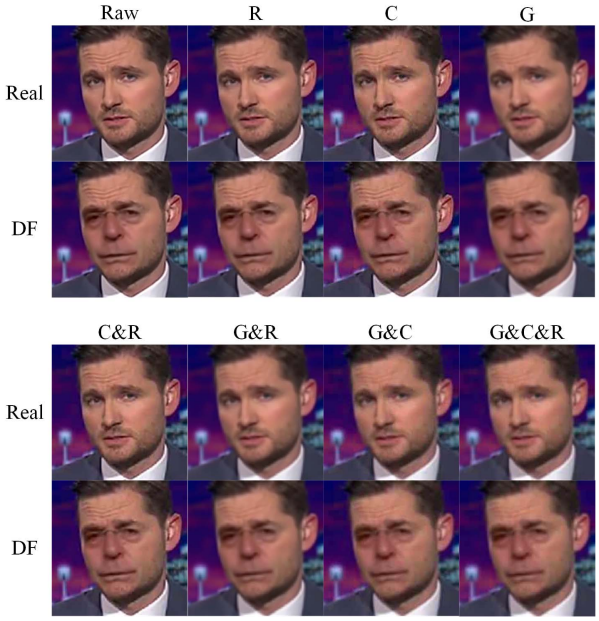

Fig. 4. Examples of images with different quality-degradation methods. We introduce the comparison of different quality degradation between real and DeepFakes (DF) images. "R", "C", and "G" indicate Resizing, JPEG Compression, and Gaussian Blur, respectively. "&" means two or more quality degradation methods are applied at the same time.

TABLE I

FRAME-LEVEL PERFORMANCE COMPARISON AMONG 12 DIFFERENT METHODS ON FF++. * INDICATES THE MODEL IS TRAINED BY US USING THE OFFICIAL CODE

Model	C23		C40	
	Acc	AUC	Acc	AUC
Steg. Features [48]	70.97	-	70.97	-
Cozzolino et al. [49]	78.45	-	58.69	-
Bayar and Stamm. [50]	82.97	-	66.84	-
Rahmouni et al. [51]	79.08	-	61.18	-
MesoNet [52]	83.10	-	70.47	-
Face X-ray [9]	-	87.35	-	61.60
Two-branch [6]	-	98.70	-	86.59
SPSL [53]	92.39	94.32	81.57	82.82
PRRNet [11]	96.15	-	86.13	-
Zhao et al. [19]	96.37	98.97	86.95	87.26
Xception* [3]	94.90	98.37	85.50	85.39
LiSiam	**96.51**	**99.13**	**87.81**	**91.44**

TABLE II

VIDEO-LEVEL PERFORMANCE COMPARISON AMONG 5 DIFFERENT METHODS ON FF++. * INDICATES THE MODEL IS TRAINED BY US USING THE OFFICIAL CODE

Model	C23		C40	
	Acc	AUC	Acc	AUC
Xception* [3]	97.00	99.31	88.71	91.93
Two-branch [6]	-	99.12	-	91.10
F^3-Net [2]	97.52	98.10	90.43	93.30
FDFL [5]	96.69	99.30	89.00	92.40
LiSiam	**97.57**	**99.52**	**91.29**	**94.65**

TABLE III

PERFORMANCE COMPARISON AMONG 12 DIFFERENT METHODS IN CROSS-DATABASE EVALUATION. "PD" MEANS PRIVATE DATA

Model	Training Set	Test Set (AUC)	
		Celeb-DF-v1	Celeb-DF-v2
Xception [3]	FF++	38.7	65.5
FWA [54]	PD	53.8	56.9
DFFD [10]	PD & FF++	71.2	-
FakeSpotter [55]	PD & FF++	-	66.8
Face X-ray [9]	PD & FF++	80.58	-
Face X-ray [9]	PD	74.76	-
Two-branch [6]	FF++	-	73.41
SPSL [53]	FF++	-	76.88
Zhao et al. [19]	FF++	-	67.44
GFFD [56]	FF++	79.4	-
MTD-Net [57]	FF++	-	70.12
LiSiam	FF++	**81.14**	**78.21**

Some frame-level methods report video-level results by averaging the accuracy and AUC of each frame in a video. Therefore, we use the same video-level evaluation metrics for fair comparisons. Table II lists video-level results of our method and other four state-of-the-art approaches. Similarly, our proposed LiSiam improves AUC by 1.35% compared with the current advanced method (i.e., F^3-Net with 93.30% v.s. LiSiam with 94.65%).

Our proposed method achieves promising performance on FF++, which can illustrate the effectiveness of our proposed LiSiam architecture. Although the image degradation process can eliminate some of the forgery traces and make forged faces more difficult to detect, our method can still capture slight forgery traces due to superior robustness against the image degradation process.

D. Cross-Database Evaluation

In this section, we conduct extensive experiments to examine the generalization capability of the proposed method. For the deepfake detection task, examining the generalization capability of different methods is usually carried out by cross-database evaluation. Therefore, we train our LiSiam on FF++ and test it on Celeb-DF-v1 [41] and Celeb-DF-v2 [42] to verify its generalization capability. The results are reported in Table III. It is indicated that the generalization performance of our proposed method is superior to state-of-the-art methods and improves about 1.74% AUC (i.e., LiSiam with 81.14% v.s. GFFD with 79.4%) towards a fair comparison on Celeb-DF-v1. Although the face X-ray method utilizes extra data to achieve high performance, our method still slightly outperforms the method. On Celeb-DF-v2, our proposed method achieves 1.33% AUC improvement over the current advanced methods (i.e., SPSL). The promising results on both databases demonstrate that the proposed method has higher detection performance and better generalization capability compared to the current state-of-the-art methods in cross-database evaluation.

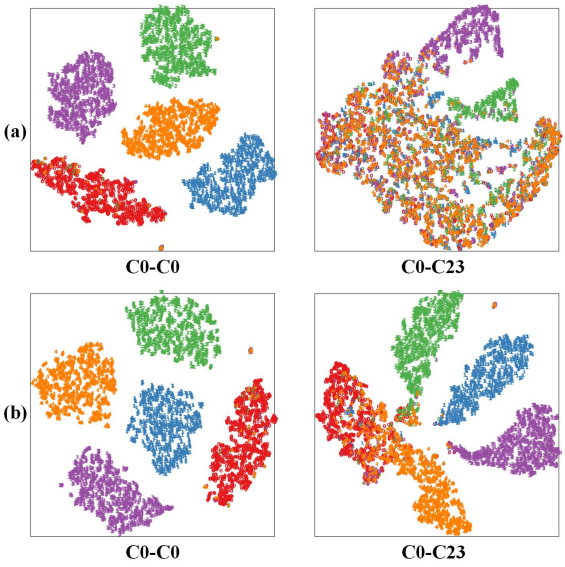

Fig. 5. t-SNE visualization of features derived from different models on the test set of FF++: (a) Xception, (b) LiSiam. C†-C‡ indicates that the model is trained on FF++ (C†) and tested on FF++ (C‡). The red, blue, green, purple, and orange color represent features from the original, DeepFakes, FaceSwap, Face2Face, and NeuralTextures images, respectively.

TABLE IV

PERFORMANCE COMPARISON WITH STATE-OF-THE-ART MODELS IN CROSS-QUALITY EVALUATION. * INDICATES THE MODEL IS TRAINED BY US USING THE OFFICIAL CODE

Model	Training Set	Test Set (AUC)		
		C0	C23	C40
Xception* [3]	C0	99.89	68.95	56.48
Zhao et al.* [19]	C0	99.82	95.26	72.97
CEViT* [22]	C0	99.77	**98.34**	81.48
LiSiam	C0	**99.93**	98.12	**83.52**
Xception* [3]	C23	99.33	99.31	80.74
Zhao et al.* [19]	C23	99.47	99.35	85.92
CEViT* [22]	C23	98.75	98.44	87.58
LiSiam	C23	**99.50**	**99.52**	**87.82**
Xception* [3]	C40	80.68	88.62	91.93
Zhao et al.* [19]	C40	90.94	91.77	93.02
CEViT* [22]	C40	89.11	88.20	90.97
LiSiam	C40	**93.88**	**93.80**	**94.65**

E. Cross-Quality Evaluation

To show the generalization performance of the proposed method for deepfake detection, we perform cross-quality experiments on the FF++ database with three image qualities. To verify the robustness of the proposed method on cross-quality deepfake detection, we train LiSiam using images of one quality and test it using images of the other two qualities. For example, we train our LiSiam with the raw images and then test it with high-quality and low-quality images. Table IV provides the video-level AUC com-

TABLE V

PERFORMANCE COMPARISON WITH THE BASELINE MODEL USING OUR IMAGE QUALITY DEGRADATION (IQD) IN CROSS-QUALITY EVALUATION. * INDICATES THE MODEL IS TRAINED BY US USING THE OFFICIAL CODE

Model	Training Set	Test Set (AUC)		
		C0	C23	C40
Xception*+IQD	C0	99.64	97.74	78.82
LiSiam	C0	**99.93**	**98.12**	**83.52**
Xception*+IQD	C23	99.31	99.07	86.23
LiSiam	C23	**99.50**	**99.52**	**87.82**
Xception*+IQD	C40	92.14	92.15	91.31
LiSiam	C40	**93.88**	**93.80**	**94.65**

parisons with state-of-the-art methods including Xception, Zhao *et al.* [19], and CEViT [22], showing that our proposed method obtains promising performance. Note that Xception achieves promising results in same-quality evaluation, but suffers from performance drop in cross-quality evaluation, especially in quality-degraded evaluation (e.g., trained on FF++ C0 and tested on FF++ C40). Compared with state-of-the-art methods, our proposed LiSiam achieves more promising results in cross-quality evaluation. It should be noted that on the FF++ database, compressed videos are obtained by the H.264 codec which is a widely used video-level compression technique. In our work, we use only three common frame-level quality-degraded operations to roughly simulate the effect of the video-level compression. The two control methods instead utilize various data augmentation to enhance the detection performance. For example, CEViT [22] utilized ImageCompression, GaussNoise, HorizontalFlip, IsotropicResize, PadIfNeeded, RandomBrightnessContrast, FancyPCA, HueSaturationValue, ToGray, and ShiftScaleRotate as data augmentation strategies.

To further demonstrate the performance improvement of our model is brought by Siamese structure with an extra quality-degraded branch, we compare our method with the baseline Xception in cross-quality evaluation, by making it use our simple image quality degradation (IQD) operations. The results are reported in Table V. From the obtained results, we observed that our method outperforms Xception when using the same data augmentation operations. This illustrates that the performance improvement in cross-quality evaluation benefits from the proposed architecture.

Discriminative feature learning is usually expected to enhance the robustness and improve the detection performance of the model. To further show the discriminative power of the feature representations learned by our method, we utilize t-distributed stochastic neighbor embedding (t-SNE) [58] to visualize the feature derived from the baseline model (Xception) and our LiSiam on the test set of FF++, as shown in Fig. 5. To better separate embedded features, both models perform five-class classification tasks (i.e., one real class and four forgery classes). From the visualization, we can clearly see that the two methods are both promising in FF++ (C0-C0), but features of our LiSiam are more separable than the baseline model in FF++ (C0-C23). LiSiam benefits from localization invariance that enhances the model to learn more discriminative features for deepfake detection. The above observation demonstrates that the proposed method is robust against quality-degraded forgery.

F. Ablation Study

To evaluate the effectiveness of each component in LiSiam, we study seven variants of our network:

1) **Model-A**. Our proposed LiSiam.

2) **Model-B**. Based on LiSiam, we further add classification loss and segmentation loss to the second branch.

3) **Model-C**. LiSiam without pixel-level segmentation map supervision.

4) **Model-D**. Based on LiSiam, we remove classification loss and segmentation loss from the first branch and instead add classification loss and segmentation loss to the second branch.

5) **Model-E**. A single branch of LiSiam. We remove the first branch and localization invariance loss and add classification loss, segmentation loss, and Mask-guided Transformer to the second branch. Model-E is a version of LiSiam without localization invariance loss.

6) **Model-F**. LiSiam without the Mask-guided Transformer.

7) **Model-G**. Model-G is a single branch of LiSiam, where we remove the second branch & localization invariance loss, and retain classification loss & segmentation loss & Mask-guided Transformer. For the input of Model-G, we include both the original and quality-degraded images.

The evaluation results on the FF++ (C40) database are shown in Table VI. To evaluate the effectiveness of the multi-task learning strategy, we evaluate various multi-task strategies to find the best practices for robust feature learning, as shown in Table VI. Specifically, we first study the effectiveness of each of the three loss functions (i.e., segmentation loss $\mathcal{L}_{seg}$, classification loss $\mathcal{L}_{cls}$, and localization invariance loss $\mathcal{L}_{li}$). We compare our proposed LiSiam with Model-B, Model-C, Model-D, Model-E, Model-F, and Model-G. We can see that the removal of classification loss $\mathcal{L}^2_{cls}$ and segmentation loss $\mathcal{L}^2_{seg}$ from the second branch improves our performance, indicating that the performance improvement comes mainly from the localization invariance loss $\mathcal{L}_{li}$ and the segmentation loss $\mathcal{L}^1_{seg}$ from the first branch. It is worth noting that Model-E is a version of LiSiam without localization invariance loss, whose performance decreases significantly. Although the second branch contains quality-degraded operations, Model-D suffers from performance degradation. In [4], the authors conducted extensive experiments to show that data augmentation improves the generalization of the model, but may remove forgery clues, hence resulting in performance degradation on some databases. The authors attempted to alleviate this problem by decreasing the probability of data augmentation. However, the empirical setting of the frequency also increases the complexity of training. To further show our proposed localization invariance loss is indeed better than simple data augmentation, we compare LiSiam with Model-G. From the results, we can clearly see that our proposed LiSiam outperforms Model-G.

In this work, we present a novel solution to maintain a good balance between cross-database generalization and

TABLE VI

ABLATION STUDY RESULTS OF DIFFERENT MODULES ON THE FF++ (C40). $\mathcal{L}^i_{seg}$ AND $\mathcal{L}^i_{cls}$ INDICATE SEGMENTATION LOSS AND CLASSIFICATION LOSS OF THE i^{th} SUB-NETWORK BRANCH OF SIAMESE NETWORKS, RESPECTIVELY. * INDICATES THE MODEL IS TRAINED BY US USING THE OFFICIAL CODE

Model	$\mathcal{L}^1_{seg}$	$\mathcal{L}^2_{seg}$	$\mathcal{L}^1_{cls}$	$\mathcal{L}^2_{cls}$	$\mathcal{L}_{li}$	MT	video-level Acc	video-level AUC	frame-level Acc	frame-level AUC
FF++ - Xception* [3]							88.71	91.93	85.50	85.39
Model-A	✓		✓		✓	✓	**91.29**	**94.65**	**87.81**	**91.44**
Model-B	✓	✓	✓	✓	✓	✓	88.43	93.13	86.57	89.41
Model-C			✓		✓	✓	88.86	93.99	86.50	89.02
Model-D		✓		✓	✓	✓	89.42	93.24	86.11	90.33
Model-E		✓		✓		✓	89.43	92.94	86.21	89.24
Model-F	✓		✓		✓		89.43	94.01	87.17	90.65
Model-G	✓		✓			✓	90.00	93.91	87.18	90.51

TABLE VII

PERFORMANCE COMPARISON BETWEEN SSIM AND DIFF MASKS ON TWO COMPRESSED VERSIONS OF THE FF++ DATABASE

Method	C23 Acc	C23 AUC	C40 Acc	C40 AUC
LiSiam (DIFF)	97.00	99.41	89.85	93.96
LiSiam (SSIM)	**97.57**	**99.52**	**91.29**	**94.65**

within-database performance. Our proposed method contains two sub-network branches. The two sub-network branches take the original image without data augmentation and quality-degraded image with data augmentation as inputs, respectively. The first branch adopts a classification loss and localization invariance loss and thus improves the robustness of the detector against image quality degradation. The two loss functions are jointly used for the training of our model. Finally, comparing Model-F with our proposed method, we observe that the video-level accuracy of Model-F decreases by 1.86% due to the lack of Mask-guided Transformer.

Besides, to evaluate the effectiveness of our ground-truth segmentation map, we compare our SSIM mask with the difference (DIFF) mask [10]. As shown in Fig. 3, our SSIM mask contains richer details (e.g., cheek texture) compared to the DIFF mask. Table. VII lists performance comparison between SSIM and DIFF masks in terms of video-level AUC and accuracy on two compressed versions of the FF++ databases, i.e., FF++ (C23) and FF++ (C40). From Table. VII, we can observe that our LiSiam based on the SSIM mask outperforms that based on the DIFF mask, especially in terms of the video-level accuracy on FF++ (C40). Experimental results demonstrate that the SSIM mask contains richer details, which allows the model to focus more on texture information. In [59], authors also noticed the importance of global textures.

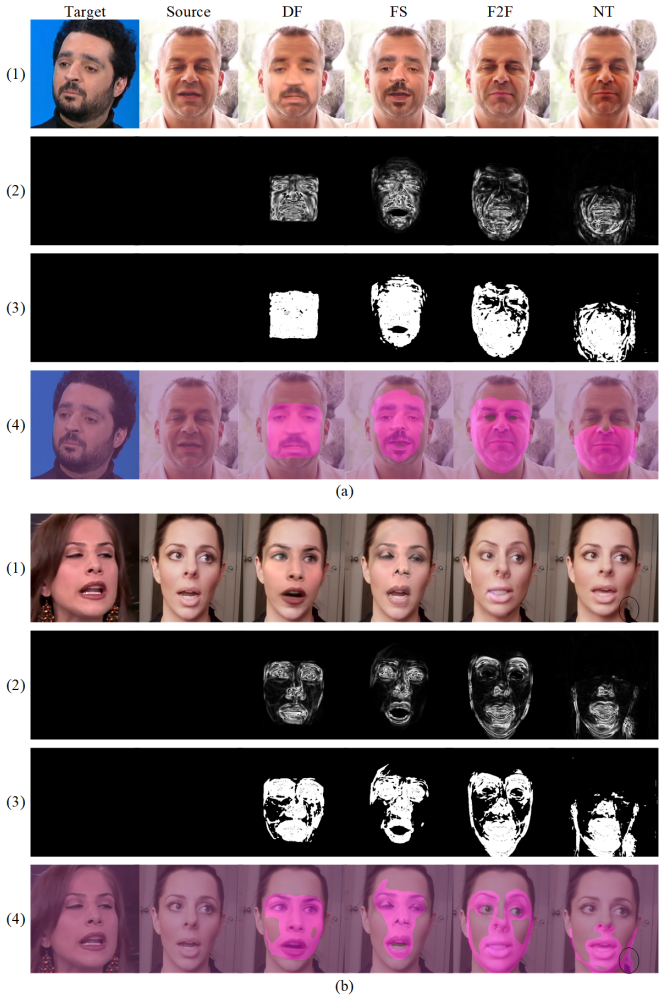

Fig. 6. Examples of the segmentation results obtained from LiSiam on the test set of FF++: (1) Input image, (2) SSIM map, (3) SSIM mask, (4) segmentation map obtained from LiSiam. Left to Right: target image, source image, fake images manipulated by the DF, FS, F2F, and NT method.

G. Visualization

To further illustrate the superiority of our method, we show visualization results of LiSiam on the test set of FF++. As shown in Fig. 6, we compare the segmentation results of LiSiam with ground truth segmentation. We can see that the segmentation map obtained from the proposed LiSiam can effectively localize tampered regions. Moreover, segmentation results can well show tampered regions caused by each forgery method. For example, NeuralTextures (NT) is used to only manipulate facial expressions around the mouth region in FF++. And, as expected, our method can well highlight the forged region around the mouth. It is worth noting that the fake image obtained by NT leaves some slight traces in Fig. 6 (b). To better show these noises, we mark the noise region with a black oval. Our LiSiam highlights this slight noise, which proves the powerful generalization ability of our model.

Fig. 7. Grad-CAM visualization of the feature maps learned by the networks in same-quality evaluation. We show two groups of Grad-CAM visualization from four forgery methods. Top to down: (a) input deepfake image, (b) SSIM image, (c) gray map learned by Xception, (d) gray map learned by LiSiam, (e) grad-CAM visualization for Xception, (f) grad-CAM visualization for LiSiam.

Besides, we compare the feature maps learned by LiSiam with the baseline model (Xception) in the same-quality and cross-quality evaluation, as shown in Fig. 7 and Fig. 8, respectively. Gradient-weighted Class Activation Mapping (Grad-CAM) [59], [60] is an advanced visualization method, which can generate the hot map to enhance visual explanations for the CNN-based networks. Therefore, we apply Grad-CAM to visualize the internal feature representations learned by the networks. For better comparison, we use the SSIM image to show the difference between the real image and the fake image as a reference. Moreover, we utilize the gray map to visualize the intensity of the hot map. From Fig. 7, we can see that the proposed method can locate the forgery region more accurately and display richer details than Xception in same-quality evaluation. To compare our proposed LiSiam with Xception in cross-quality evaluation, both the methods are trained with the raw (i.e., C0) images from the FF++ training set. And then we visualize the feature maps learned by the two methods on the three image-qualities (i.e., C0, C23, and C40) faces from FF++ test set. The visualization results of Fig. 8 clearly show that our LiSiam is robust against quality-degraded forgery owing to localization invariance. Although Xception can roughly locate the forged region on the C0 test image, it fails on some quality-degraded images (e.g., the C23 DF image, the C40 F2F image, etc.). Compared with Xception, our proposed LiSiam can maintain stable localization performance in cross-quality evaluation.

V. Conclusion and Future Work

In this paper, a novel Localization Invariance Siamese Network is proposed to enforce localization consistency across images with different image quality degradation. The major advantage of our method is that the proposed localization invariance loss enhances the robustness of the detector against image quality degradation. Moreover, the Mask-guided Transformer is designed to better capture the difference between the forgery region and its surroundings, which allows the model to learn robust and discriminative features. Finally, we utilize a multi-task learning strategy to optimize the network in an end-to-end manner. We conduct extensive experiments to evaluate the performance of the proposed LiSiam. Promising results demonstrate the effectiveness of our approach for deepfake detection.

In future work, we will extend our frame-level method to the video level with a temporal association strategy. Our goal is to utilize temporal information for learning more robust features and discovering temporal forgery traces.

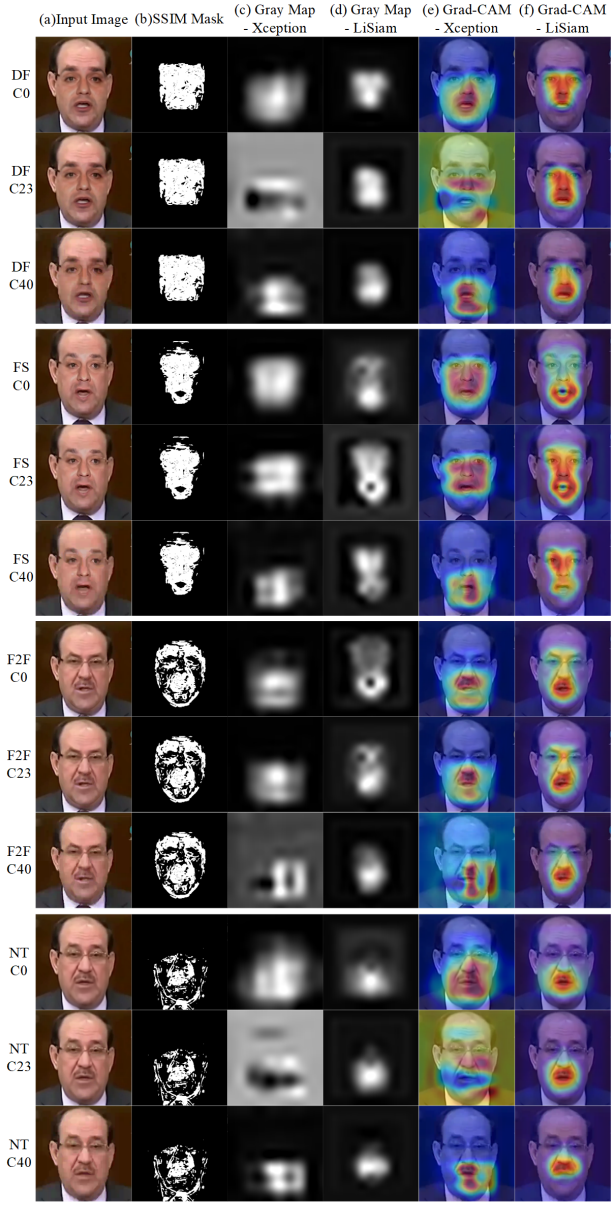

Fig. 8. Grad-CAM visualization of the feature maps learned by the networks in cross-quality evaluation.

ACKNOWLEDGMENT

The authors would like to thank the associate editor and anonymous reviewers for their useful suggestions and significant efforts spent to help them for further improving their article.

REFERENCES

[1] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, "Local relation learning for face forgery detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1081–1088.

[2] 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.

[3] A. Rossler, 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.

[4] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot…For now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8695–8704.

[5] 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," 2021, *arXiv:2103.09096*.

[6] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 667–684.

[7] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7890–7899.

[8] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? Understanding properties that generalize," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 103–120.

[9] 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. 5001–5010.

[10] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5781–5790.

[11] Z. Shang, H. Xie, Z. Zha, L. Yu, Y. Li, and Y. Zhang, "PRRNet: Pixel-region relation network for face forgery detection," *Pattern Recognit.*, vol. 116, Aug. 2021, Art. no. 107950.

[12] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1325–1334.

[13] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6023–6032.

[14] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "DropBlock: A regularization method for convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[15] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3544–3553.

[16] 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.

[17] 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.

[18] T. Jung, S. Kim, and K. Kim, "DeepVision: Deepfakes detection using human eye blinking pattern," *IEEE Access*, vol. 8, pp. 83144–83154, 2020.

[19] 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.

[20] 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.

[21] C.-Z. Yang, J. Ma, S. Wang, and A. W.-C. Liew, "Preventing DeepFake attacks on speaker authentication by dynamic lip movement analysis," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1841–1854, 2021.

[22] D. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and vision transformers for video DeepFake detection," 2021, *arXiv:2107.02612*.

[23] S. A. Khan and H. Dai, "Video transformer for DeepFake detection with incremental learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1821–1828.

[24] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. Int. Conf. Mach. Learn. Workshops*, vol. 2, 2015, pp. 1–30.

[25] A. He, C. Luo, X. Tian, and W. Zeng, "A twofold Siamese network for real-time object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4834–4843.

[26] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12275–12284.

[27] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.

[28] O. Mayer and M. C. Stamm, "Forensic similarity for digital images," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1331–1346, 2020.

[29] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and B. S. Manjunath, "Exploiting spatial structure for localizing manipulated image regions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4970–4979.
[30] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1053–1061.
[31] J. H. Bappy, C. Simons, L. Nataraj, B. S. Manjunath, and A. K. Roy-Chowdhury, "Hybrid LSTM and encoder–decoder architecture for detection of image forgeries," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3286–3300, Jul. 2019.
[32] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[33] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[34] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 5754–5764.
[35] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.
[36] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[38] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 31, 2021, doi: 10.1109/TKDE.2021.3070203.
[39] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 173–190.
[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[41] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," 2019, *arXiv:1909.12962*.
[42] 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. 3207–3216.
[43] (2018). *Deepfakes*. [Online]. Available: https://github.com/deepfakes/faceswap
[44] (2018). Kowalski. *Faceswap*. [Online]. Available: https://github.com/MarekKowalski/FaceSwap/
[45] J. Thies, M. Zollhofer, 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.
[46] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 66:1–66:12, 2019.
[47] Y. Xu, W. Yan, G. Yang, J. Luo, T. Li, and J. He, "CenterFace: Joint face detection and alignment using face as point," *Sci. Program.*, vol. 2020, pp. 1–8, Jul. 2020.
[48] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, Jun. 2012.
[49] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection," in *Proc. 5th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2017, pp. 159–164.
[50] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proc. 4th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2016, pp. 5–10.
[51] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in *Proc. IEEE Workshop Inf. Forensics Security (WIFS)*, Dec. 2017, pp. 1–6.
[52] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, Dec. 2018, pp. 1–7.
[53] H. Liu *et al.*, "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 772–781.
[54] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 46–52.
[55] R. Wang *et al.*, "FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces," 2019, *arXiv:1909.06122*.
[56] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16317–16326.
[57] J. Yang, A. Li, S. Xiao, W. Lu, and X. Gao, "MTD-Net: Learning to detect deepfakes images by multi-scale texture difference," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4234–4245, 2021.
[58] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[59] Z. Liu, X. Qi, and P. H. S. Torr, "Global texture enhancement for fake face detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8060–8069.
[60] U. Ozbulak. (2019). *PyTorch CNN Visualizations*. [Online]. Available: https://github.com/utkuozbulak/pytorch-cnn-visualizations

Jian Wang received the B.E. degree from the Xuzhou University of Technology, Xuzhou, Jiangsu, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include deepfake detection, machine learning, and computer vision.

Yunlian Sun received the M.E. degree in computer science and technology from the Harbin Institute of Technology, China, in 2010, and the Ph.D. degree in ingegneria elettronica, informatica e delle telecomunicazioni from the University of Bologna, Italy, in 2014. After the Ph.D. study, she worked as a Post-Doctoral Researcher with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences. She is currently an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. Her research interests include biometrics, pattern recognition, and computer vision.

Jinhui Tang (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the University of Science and Technology of China in 2003 and 2008, respectively. He is currently a Professor with the Nanjing University of Science and Technology. He has authored more than 150 papers in top-tier journals and conferences. His research interests include multimedia analysis and computer vision. He was a recipient of the Best Paper Award in ACM MM 2007, PCM 2011, and ICIMCS 2011; the Best Paper Runner-Up in ACM MM 2015; and the Best Student Paper Award in MMM 2016 and ICIMCS 2017. He has served as an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, and the IEEE TRANSACTIONS ON MULTIMEDIA.