Exploiting Complementary Dynamic Incoherence for DeepFake Video Detection

Hanyi Wang, Zihan Liu, †Shilin Wang, *Senior Member, IEEE,*

***Abstract*—Recently, manipulated videos based on DeepFake technology have spread widely on social media, causing concerns about the authenticity of video content and personal privacy protection. Although existing DeepFake detection methods achieve remarkable progress in some specific scenarios, their detection performance usually drops drastically when detecting unseen manipulation methods. Compared with static information such as human face, dynamic information depicting the movements of facial features is more difficult to forge without leaving visual or statistical traces. Hence, in order to achieve better generalization ability, we focus on dynamic information analysis to disclose such traces and propose a novel Complementary Dynamic Interaction Network (CDIN). Inspired by the DeepFake detection methods based on mouth region analysis, both the global (entire face) and local (mouth region) dynamics are analyzed with properly designed network branches, respectively, and their feature maps at various levels are communicated with each other using a newly proposed Complementary Cross Dynamics Fusion Module (CCDFM). With CCDFM, the global branch will pay more attention to anomalous mouth movements and the local branch will gain more information about the global context. Finally, a multi-task learning scheme is designed to optimize the network with both the global and local information. Extensive experiments have demonstrated that our approach achieves better detection results compared with several SOTA methods, especially in detecting video forgeries manipulated by unseen methods.**

***Index Terms*—DeepFake video detection, Video forensics.**

I. Introduction

THE rapid development of deep generative models, especially Variational AutoEncoders(VAE) [1] and Generative Adversarial Networks(GANs) [2], [57], permits the use of off-the-shelf models to create visually extremely realistic fake videos with little expertise using open source community applications [3]–[5]. Any amateur user is capable of producing DeepFake videos where human faces, or sometimes only lip regions, are modified in order to simulate the presence of a specific subject in a certain context or to make someone speak coherently with a different and probably compromising speech. The combination of indistinguishable DeepFake products and social media networks further amplify the effect of abusing such technology (e.g. spreading politician propaganda and discrediting individuals). As a visual confrontation technology, DeepFake can deceive human eyes and even modern face recognition models [48].

H. Wang and Z. Liu are with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shang-hai 200240, China, (e-mail: why_820@sjtu.edu.cn; lzh123@sjtu.edu.cn; wsl@sjtu.edu.cn)

†Shilin Wang is the corresponding author.

The work described in this paper was supported by the National Natural Science Foundation of China under Grant 62271307, and Grant 61771310.

Considering its serious social impacts, DeepFake has recently gained significant attention. Various methods have been proposed for DeepFake detection, which can be roughly divided into two categories: image-based and video-based approaches.

Image-based DeepFake Detection. Early attempts targeted at the spatial artifacts on the manipulated faces and various 2D CNN models [12], [17]–[19] were designed to extract high-level information from the spatial domain for forgery detection. [17] designed two compact networks to capture the mesoscopic features and [19] proposed a capsule network. Meanwhile, some researchers [20]–[23] exploited frequency-aware anomalous patterns to mine irregularities with low-level statistic information. [21] and [22] detected the artifacts with anomalous amplitude and phase spectrum, respectively. Besides, some works [6], [10], [11], [24] took advantage of RGB space and frequency domain simultaneously to capture subtle forgery artifacts. [24] leveraged spatial features and steganalysis features. F^3-Net [10] extracted DCT-based frequency features to enhance the detection robustness against compression. [56] proposed a dual-stream network by integrating RGB and YCbCr color spaces to detect post-processed generated faces. [53], [54] simulated the synthetic data generation pipeline where a fake image is generated by blending two pristine images, aiming to learn the artifacts caused by blending. [55] proposed an encoder-decoder generator to track the potential texture traces left in image generation. [59] leveraged a two-stage self-supervised paradigm to learn features of intra-class consistency and inter-class diversity.

Video-based DeepFake Detection. Recent studies tended to leverage temporal inconsistencies as a means to indicate abnormal face movement in a video stream. This inconsistency usually occurs during the synthesis process since many manipulation methods were processed on isolated frames. [25] exploited abnormal eye-blinking frequency for detection. [26] detected irregular binoculars movements based on spontaneous and consistent eye-gaze motion. In [28]–[30], CNNs followed by LSTM modules were introduced to capture spatial-temporal artifacts. [27] adopted optical flow fields to exhibit latent inter-frame dissimilarities. In [52], a two-stream method analyzing frame-level and temporality-level characteristics was designed to improve the detection performance against compression. [31] exploited the intrinsic synchronization patterns between visual and auditory modalities for joint detection.

The above methods have achieved impressive progress benefiting from the releases of large-scale face forgery datasets

[12]–[15]. However, they often tended to overfit their training dataset(s) and usually experienced significant performance degradation under the cross-dataset scenario. Recent works have noticed this problem and attempted to enhance the generalization capability to unseen forgeries. FWA [32] and Face X-ray [9] detected the artifacts produced by blending boundaries based on the assumption that most of the manipulation methods share similar blending processes between the altered face and the background. However, this assumption is not always valid and susceptible to some well-designed post-processing operations. [11] proposed to utilize the image's high-frequency noise features by removing the color texture to reveal forgery traces; however, it is vulnerable to compression and other post-processing procedures. [7] focused on the temporal coherence by reducing the spatial convolution kernel size to 1 and maintained the temporal convolution kernel size unchanged. [8] targeted high-level semantic irregularities in mouth movement using the lipreading network. However, when the most noticeable forgery traces lie outside the mouth region, their method cannot achieve satisfactory results.

In this paper, a novel DeepFake detection network, i.e. Complementary Dynamic Interaction Network (CDIN), is proposed, which leverages both the dynamic information of the most discriminative local region (i.e. the mouth region) and the global context (i.e. the entire face) to achieve more accurate and generalizable detection results. Based on the observation that some discontinuities may occur in non-adjacent frames, for example, moles on the face may appear or disappear constantly [7]. We propose to leverage Transformer [33] to model long-range dependencies as well as capture incoherent artifacts along the temporal dimension for the entire faces. Besides, a 3D Convolutional neural network [34] with well-designed reception fields is adopted to capture dynamic details of mouth region. The two branches interact with each other to learn complementary dynamics using the proposed Complementary Cross Dynamics Fusion Module (CCDFM) at multiple levels. With CCDFM, the global branch will pay more attention to the lip movements, and the local branch will gain more information about the global face region. In this way, both the global dependency and the local fine-grained artifacts can be effectively captured and represented in a spatiotemporal manner.

The major contributions are summarized as follows:

- A two-branch Complementary Dynamic Interaction Network (CDIN) is proposed to exploit both global (entire face) and local (mouth region) anomalous dynamic artifacts for DeepFake detection.
- A new Complementary Cross Dynamics Fusion Module (CCDFM) is proposed for mutual reinforcement between the entire face and the mouth region in an interactive manner.
- Extensive experiments on several benchmarks have demonstrated the superiority of our method, especially in detecting unknown manipulation methods.

The rest of the paper is organized as follows. Section II discusses the major challenges and introduces our motivations. Section III presents the proposed method and elaborates each key module. The experiment results and discussions are given in Section IV. Section V draws the conclusion.

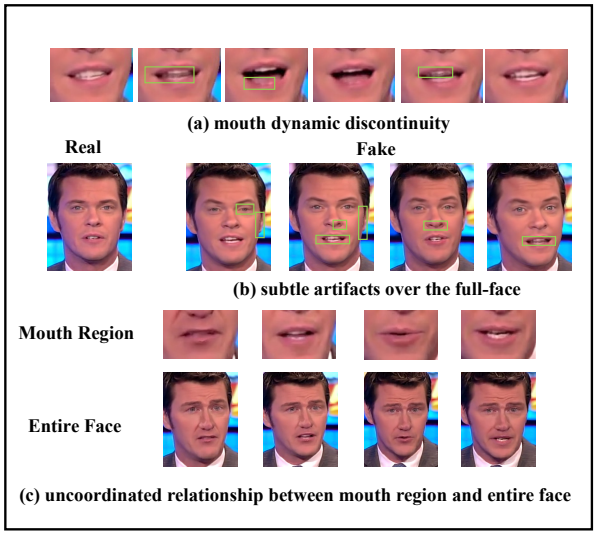

Fig. 1. Illustration of visually noticeable artifacts left by DeepFake. (a): inconsistent mouth movement;(b): mouth dynamic discontinuity(c): uncoordinated relationship between mouth region and entire face.

II. Major Challenges and Motivations

Previous face forgery detection methods have achieved remarkable successes in detecting specific manipulation methods but experienced a dramatic performance drop under cross-database scenarios. In real-word scenarios, Deepfake producers can choose any type of manipulation method, which may not appear in the detector's training set. Therefore, how to design a detector with good generalization ability becomes an important task. Compared with static artifacts, dynamic information depicting the movements of facial features is much more difficult to forge without leaving visual or statistical traces and recent researches [7]–[9], [11], [32] have demonstrated the dynamic features can help improve the detector's generalization ability. However, most existing methods based on dynamic feature analysis focused on either the entire face region [9] or a specific local region [8] and cannot yet achieve satisfactory results under the cross-dataset scenario due to the following challenges.

Local characteristics analysis alone lacks sufficient discriminative traces. Previous works [8] have shown that DeepFake manipulation methods leave traceable inconsistencies in certain local regions during movement. The mouth region is one of the most discriminative regions in DeepFake detection because it has rich dynamic information and its movement is difficult to synthesize naturally and smoothly. For example, in Fig.1 (a), it is observed that the consistency of tongue, teeth and oral cavity between adjacent frames are often disrupted due to illumination or missing reflection. Besides, the dynamics of mouth shape maybe irregular compared to pristine videos [49]. Although the local information alone is conducive to detecting some anomalous artifacts, from a global perspective, there are numerous forgery methods manipulating face out of such local regions. If the model is forced to learn the differences in the local regions, it tends to learn the manipulation-independent discrepancies such as background

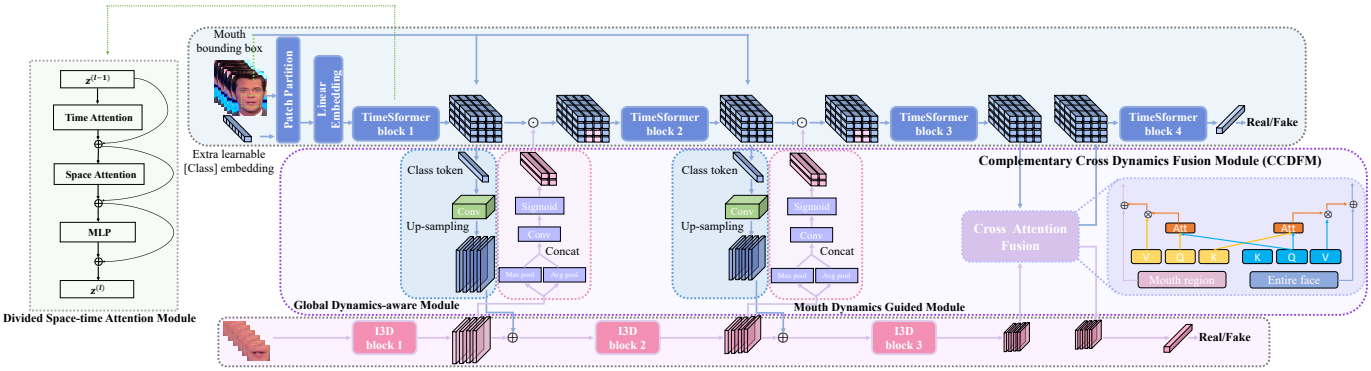

Fig. 2. Network architecture of the proposed CDIN. CDIN consists of two parallel branches: the global fundamental dynamics-aware branch(GAD-branch in short) and the local representative mouth movements-aware branch(LRM-branch in short), which process consecutive frames of entire face and mouth region respectively. CCDFM is carefully designed as a bridge module to fuse local movement details in LRM-branch with global dynamic representations in GAD-branch in an interactive fashion.

or content information. Thus, it easily overfits on the training set and shows limited generalization capability. Besides, with numerous advances in lip-sync manipulation techniques [16], [50], [51], the mouth region and its movements can be generated more realistically, which makes it more difficult to differentiate fake lip movements from genuine ones solely based on mouth region analysis.

Blurry focus caused by full-face analysis alone. With the counterfeits manipulated more realistically, the discrepancies between real and fake videos become more subtle and localized. For example, in Fig.1 (b), it is observed that the artifacts may appear on the nostril, oral cavity, etc., where the differences between real and fake videos are very subtle. To differentiate such anomalies, the full-face representations should focus more on the local manipulation-sensitive regions, which are prone to exposure of anomalies. However, it is difficult to concentrate on these manipulation-sensitive regions only relying on the global supervision from a binary label. Such analysis from a global view solely is hard to reveal the fine-grained artifacts and it is vulnerable to learn some manipulation-independent differences. Therefore, the blurry focus caused by full-face analysis alone also limits the improvement of generalization capability.

In order to address the above challenges, both the global (i.e. entire face) and local (i.e. mouth region) dynamics should be analyzed simultaneously and comprehensively. [60], [61] also utilize both the global and local characteristics to expose the spatiotemporal artifacts for DeepFake detection. [60] takes random selected two frames as inputs, which did not consider the long-term abnormal variations over time. [61] modeled the dynamic incoherence from features acquired by the pooling operation. However, temporal consistency representation from pooling features may result in a loss of fine-grained information. Moreover, a late fusion strategy is applied in both of [60], [61] to simply concatenate the global and local features for classification. In contrast, according to our motivation, local representations should be exploited to refine the global representations to enhance local manipulation-sensitive features. On the other hand, to avoid the local branch being trapped in local decision bias, the local representation should preserve the global perception consistency of entire face dynamics. To this end, both the global and local dynamics are mutually complemented and reinforced in an interactive manner. Specifically, the information interaction and complementarity between the global and local features are progressively learned from different stages of the network (i.e., from low-level textural to high-level semantic features). The global information can help the local feature gain additional reception fields, and thus the incoordination artifacts between the local region and the rest of the face region (Fig. 1 (c)) could be revealed effectively. On the other hand, the local information can help the global feature to emphasize the local manipulation-sensitive regions. Motivated by the above analysis, a novel network for Deepfake detection, i.e. Complementary Dynamic Interaction Network (CDIN), is designed and elaborated in Section III.

III. Proposed Method

A. Overview

The overall architecture of the proposed network is given in Fig. 2. It consists of two parallel branches: the Global Adaptive Dynamics-aware branch (GAD-branch) and the Local Representative Lip Movements-aware branch (LRM-branch), which process consecutive frames of the entire face and the mouth region, respectively. The two branches interact with each other to learn complementary dynamics using the proposed Complementary Cross Dynamics Fusion Module (CCDFM) at multiple levels. Finally, the predictions of these two branches are integrated to derive the final detection result.

B. Network Structure

1) Global Adaptive Dynamics-aware Branch

Given the entire face sequence as input, the Global Adaptive Dynamics-aware branch (GAD-branch) aims to characterize the global dynamics and model long-range dependencies for the entire face. In the GAD-branch, TimeSformer [33] is employed as the backbone because it does not perform any spatial or temporal downsampling, which facilitates learning the fine-grained dynamics for each patch in the face region. The GAD-branch is divided into 4 stages and each stage contains 3 "Divided Space-time Attention" modules as shown in Fig. 2, where temporal and spatial attention are separately applied one after the other. The temporal attention is computed from all the patches at the same spatial location in the other frames and the spatial attention is computed from all the patches in the same

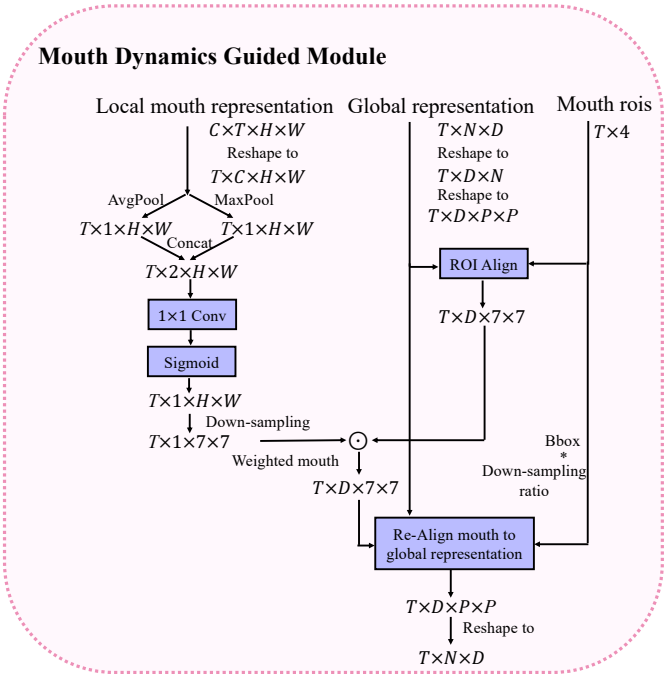

Fig. 3. Detailed operation of Mouth Dynamics Guided Module.

frame. The semantic information is gradually enhanced with the stack of stages with the non-downsampling scheme and the divided spatiotemporal self-attention mechanism. Therefore, GAD-branch is able to capture the characteristics of the global dynamics effectively.

2) Local Representative Lip Movements-aware branch

Local anomalous dynamics are crucial for face forgery detection. In this work, we employ the I3D structure as the backbone of the LRM-branch. The whole branch can be divided into 3 stages, each stage is composed of multiple convolution blocks. Each block consists of several Inception modules, which increase the width of the network with multi-scale reception fields to obtain high-level semantic representations. The convolution kernels slide over feature maps with overlap, which aims to extract fine-grained local features. LRM-branch takes the lip sequence as input and adopts the feature pyramid structure to further enlarge the reception field for exploring dynamic details. The resolution of feature maps decreases while the number of channels increases from the lower to upper layers. With the occasional max-pooling layers to halve the resolution of the grid, the local motion artifacts will be further amplified. Finally, the global average pooling is adopted to integrate all the features and then fed it to an FC layer for classification. The above 3D CNN with well-designed multi-scale reception fields is sensitive to subtle movement details, especially for lip movements, which contain rich dynamic information.

3) Complementary Cross Dynamics Fusion Module

To explore irregular mouth movements as well as preserve the global perception of entire face dynamics, we devise a novel Complementary Cross Dynamics Fusion Module (CCDFM) for mutual reinforcement between the two branches in an interactive manner, whose structure is shown in Fig. 2. CCDFM bridges the two branches at multiple levels. Specifically, after the first and second stage of both branches, it

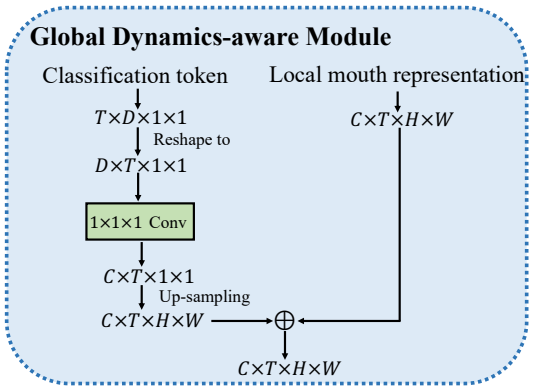

Fig. 4. Detailed operation of Global Dynamics-aware Module.

is applied to enable both branches to learn from each other. CCDFM contains a Mouth Dynamics Guided module and a Global Dynamics-aware module in the first two stages, a Cross Attention Fusion module in the last stage. Considering that the low-level features mainly contain local details, while high-level features are strongly related to semantic information, thus we apply the mutual early fusion strategy in the first two stages for low-level spatial content interaction, the cross-attention mechanism for high-level semantic communication. The Mouth Dynamics Guided module helps the global branch focus more on the local manipulation-sensitive regions and the Global Dynamics-aware module helps the local branch to gain addition global context information. The Cross Attention Fusion module further reinforces the high-level semantic interaction between the two branches.

Mouth Dynamics Guided Module. Inspired by CBAM [39], we adopt spatial attention from the local branch to highlight the mouth dynamics and guide the global branch to focus more on irregular lip movements. It is worth noting that the spatial attention weights are derived from the local mouth movements, and are applied to the mouth region of the global feature maps. The detailed operations are shown in Fig. 3. We firstly obtained the mouth movement representations with feature size $C \times T \times H \times W$ from the local branch, and global dynamic representations with feature size $T \times N \times D$ from the global branch, respectively. T, N, and D separately represent the number of RGB frames, the total number of image patches, and patch embedding dim. C, H, and W separately represent the channel number, height, and width of the current feature map. As in [39], the attention weights are firstly produced from the local mouth movements through spatial attention mechanism. Besides, Sigmoid is applied to confine the weight values to 0-1. We use the ROI Align proposed in [62] to acquire the feature maps of the mouth region from the global representations. A down-sampling operation is performed on the attention weights to align the spatial dimension of the ROI features. And then we acquire the weighted mouth feature maps through the element-wise product. Finally, we use the down-sampled bounding box to re-align the mouth feature maps to the global positions, and then the weighted features of the mouth region are added to the mouth region of global representations.

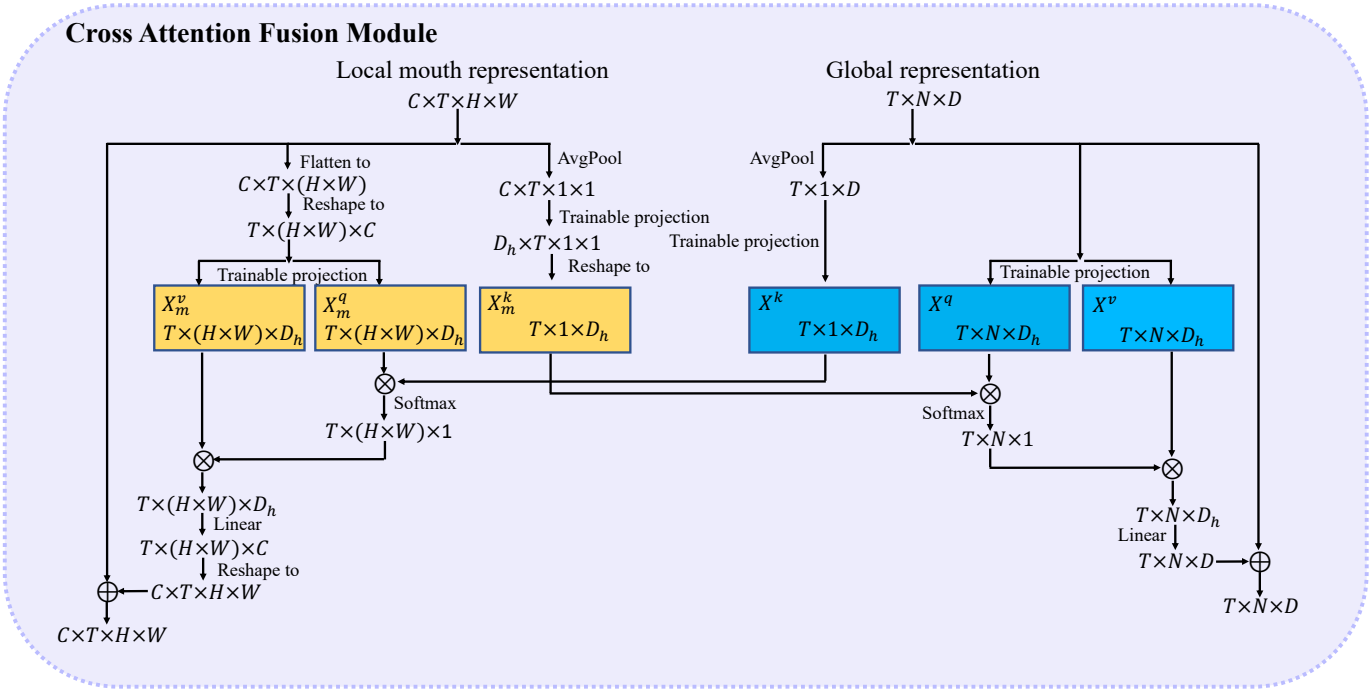

Fig. 5. Detailed operation of Cross Attention Fusion Module.

Global Dynamics-aware Module. Feeding the global information to local branch is implemented by the classification token. Similar to the settings in the vanilla ViT [35], after patch partition and linear embedding, a learnable vector $Z \in \mathbb{R}^{T\times D}$ is prepended to the embedded patches of clip sequence in the first position. T, and D separately represent the number of RGB frames, and patch embedding dim. The learnable vector serves as the classification token which integrates all the features through Divided space-time blocks and represents the global information. The detailed operations are shown in Fig. 4. Specifically, in the early interaction, we fetch it from the first position of the global features and then feed it to the Global Dynamics-aware Module for fusion. The classification token embedding is further up-sampled by bi-linear interpolation to align the spatial scale. The channel dimension is then aligned by a 1×1 convolution. Finally, it is reshaped and added to the local feature maps.

Cross Attention Fusion Module. A cross-attention fusion module is proposed to further enhance the interaction between the global and local branches. As shown in Fig. 5, denoting input features obtained from the GAD-branch and LRM-branch as $X \in \mathbb{R}^{T\times N\times D}$ and $X_m \in \mathbb{R}^{T\times C\times H\times W}$, respectively. We first convert the mouth dynamic representation $X_m \in \mathbb{R}^{C\times T\times H\times W}$ using an adaptive average pooling along spatial scales. Then a trainable projection followed by shape transformation is applied to acquire key $X_m^k \in \mathbb{R}^{T\times 1\times\ D_h}$. Besides, we flatten X_m and then reshape it to $T \times (H \times W) \times C$. Query $X_m^q \in \mathbb{R}^{T\times(H\times W)\times\ D_h}$ and Value $X_m^v \in \mathbb{R}^{T\times(H\times W)\times\ D_h}$ are obtained through a trainable projection. Similarly, X is transformed to obtain $X^k \in \mathbb{R}^{T\times 1\times D_h}$, $X^q \in \mathbb{R}^{T\times N\times D_h}$ and $X^v \in \mathbb{R}^{T\times N\times D_h}$. Then we compute the correlation between the two branches.

For the entire face, we generate the attention map by

$$A = Softmax(X^q \otimes X_m^k) \tag{1}$$

and then obtain the refined feature

$$X^{out} = A \otimes X^v \tag{2}$$

according to its correlation with local dynamics, where $X^{out} \in \mathbb{R}^{T\times N\times D_h}$. We then acquire $X^{out} \in \mathbb{R}^{T\times N\times D}$ using a Linear projection to re-align the original embedding dim. Finally, the global representation is obtained by

$$X^{out} = X^{out} + X \tag{3}$$

The same operation is executed on mouth movements, where

$$X_m^{out} = Softmax(X_m^q \otimes X^k) \otimes X_m^v \tag{4}$$

and $X_m^{out} \in \mathbb{R}^{T\times(H\times W)\times\ D_h}$. A Linear projection followed by shape transformation is then applied to obtain $X_m^{out} \in \mathbb{R}^{C\times T\times H\times W}$. Finally, the refined local representation is obtained by

$$X_m^{out} = X_m^{out} + X_m \tag{5}$$

X^{out} and X_m^{out} embody the interaction of complementary information and mutually promote the dynamic feature communication.

IV. Experiments and Discussions

A. Settings

Datasets. Following the evaluation protocol of recent works [9], [12], we adopted the most commonly-used benchmark datasets FaceForensics++(FF++) [12] for training, Celeb-DF(v2) [13], DeepFake Detection Challenge(DFDC) [14] and ForgeryNet [63] for evalution. FF++ is the most widely used face forgery dataset consisting of 1,000 original videos and

4,000 fake videos. Fake videos are generated by four state-of-the-art face manipulation methods: Deepfakes [3], Face2Face [36], FaceSwap [4] and NeuralTextures [37]. There are three versions of FF++ in terms of compression level, i.e., raw, high quality(HQ) and low quality(LQ). Generally speaking, the videos in the LQ version are usually more difficult to distinguish between real and fake owing to high compression loss. The HQ version was adopted for training and the official split was adopted (720 videos for training, 140 videos for validation, and 140 videos for testing). The corresponding four types of manipulated videos were adopted to generate negative samples. During training, each video was divided into non-overlapping clips by a sliding window mechanism. Each video clip contained 16 frames. We randomly sampled 4 clips per fake video as negative samples. As there are four types of fake videos, to preserve the class balance, an equal amount 4 $\times$ 4=16 clips were sampled from the corresponding real video as positive samples. Celeb-DF(v2) is a high-quality face swapping dataset which comprises 5,639 videos from YouTube. The fake videos are generated by an improved manipulation method thus leaving few noticeable tampering traces. DFDC is a large-scale dataset where subjects in complex scenes are manipulated using various unknown methods. ForgeryNet is a large face forgery dataset with unified annotations in image-level and video-level data. It is worth noting that the ForgeryNet dataset contains diverse manipulation types where 15 forgery approaches are applied to image-forgery construction and 8 of them are applied to the video-forgery construction. During inference, each video was divided into overlapping clips with a step of 1 by the sliding window mechanism where the window size was 16 frames. The final video-level prediction result was obtained by averaging all the clips' prediction scores. The threshold was set to 0.5 to decide the final prediction label.

Evaluation Scenarios. To comprehensively evaluate the detection performance and generalization capability of our method, we considered the following two scenarios: Intra-dataset and Cross-dataset. Intra-dataset: The experiments were conducted on FF++(HQ) which consists of forgery videos tampered by four different manipulation methods from the same source videos. Cross-dataset: The experiments are performed under the cross-dataset scenario where training and testing samples are collected from different datasets. All the models are trained on FF++(HQ) [12] and evaluated on Celeb-DF-v2 [13], DFDC [14], respectively. This setting is more challenging than the previous one due to the variations in manipulation methods, video content, etc.

Baselines. To comprehensively evaluate our method, several state-of-the-art DeepFake detection methods were adopted for comparison. Image-based detection methods: **Xception** [12], **Face X-ray** [9], **F^3 Net [10]** and Video-based detection methods: **FTCN** [7], **LipForensics** [8]. For a fair comparison, we re-implemented all the baselines according to the training protocol in their original papers.[1]

[1]For Xception, FTCN, and LipForensics, we only wrote scripts for dataset organization and training, which were not available in their open-source projects.

TABLE I
EFFECTIVENESS OF LRM-BRANCH

Model	Intra-dataset					Cross-dataset			
	DF	F2F	FS	NT	**Avg**	FF++	Celeb-DF	DFDC	**Avg**
LSTM [42]	0.991	0.964	0.981	0.962	**0.974**	**0.985**	0.684	0.658	0.776
C3D [43]	0.977	0.911	0.919	0.902	0.927	0.938	0.656	0.681	0.759
P3D [44]	0.982	0.944	0.958	0.913	0.949	0.968	0.703	0.692	0.788
R(2+1)D [45]	0.980	0.922	0.945	0.918	0.941	0.961	0.714	0.686	0.787
I3D [34]	0.988	0.972	0.979	0.951	0.972	0.979	**0.749**	**0.701**	**0.810**

TABLE II
EFFECTIVENESS OF GAD-BRANCH

Model	Intra-dataset					Cross-dataset			
	DF	F2F	FS	NT	**Avg**	FF++	Celeb-DF	DFDC	**Avg**
ViViT [46]	0.966	0.953	0.969	0.920	0.952	0.961	0.713	0.683	0.786
Swin Base [47]	0.986	0.971	0.988	0.967	**0.978**	**0.988**	0.755	0.703	0.815
TimeSformer [33]	0.980	0.971	0.970	0.974	0.974	0.977	**0.858**	**0.712**	**0.849**

B. Implementation Details

For each video frame, DLIB [38] was adopted to extract and align face regions and the aligned faces were resized to 224×224. The corresponding mouth regions were also cropped and resized to 64×128. The face crops and the corresponding mouth regions were fed to the network in parallel. The proposed framework was implemented via open-source PyTorch [40]. TimeSformer [33] and I3D [34] were adopted as our backbone and the weights were initialized with the pre-trained model on the Kinetics dataset [34]. The number of DSTA blocks L was set to 12 and its embedding dim D was set to 768. The cross attention hidden dim D_h was set to 256. The Adam optimizer was used to train the framework with a learning rate of 1e-5 and a weight decay of 1e-4. The batch size was 8 and 30 epochs were trained. During training, the overall loss function is composed of two equally weighted cross entropy losses supervising both the global and local features. During inference, the final prediction is calculated by a weighted summation over the prediction scores of both the GAD-branch and LRM-branch. The weights for GAD and LRM branches are empirically set as 0.6 and 0.4, respectively.

C. Ablation Study

Effectiveness of the Backbone Networks. To comprehensively evaluate the effectiveness of our network design, several experiments were conducted for both the global and local branches. For the LRM-branch, various lip movement analysis networks such as LSTM [42], C3D [43], P3D [44], R(2+1)D [45], I3D [34] were adopted for comparison. The comparison results are illustrated in Table I. It is observed that LSTM exhibits the best detection results within dataset, but is more likely to suffer from the overfitting problem. I3D outperforms all the competitors in terms of the generalization ability as well as shows a good detection performance under within-dataset scenario, which has demonstrated that it can learn a more robust dynamic representation.

For the GAD-branch, several video-level Transformer architectures were experimented to validate their effectiveness in charactering global dynamics. Popular models used in video understanding and analysis such as ViViT [46], TimeSformer [33] and Swin Transformer [47] were considered for comparison. It can be observed from Table II that Swin Transformer achieves the best intra-dataset classification results but it does not generalize well to unseen DeepFake methods. TimeS-

TABLE V
CROSS-MANIPULATION GENERALIZATION EVALUTION

Training Set	Methods	Intra-dataset					Cross-dataset								
		DF	F2F	FS	NT	Avg	FaceShifter	FS-GAN	DeepFakes	BlendFace	MMRepla	DF-S-S	T-H-V	ATVG-Net	**Avg**
DF	Xception [12]	0.999	0.718	0.421	0.877	0.754	0.480	0.499	0.552	0.472	0.476	0.593	0.501	0.547	0.515
	FTCN [7]	0.993	0.760	0.535	0.874	0.791	0.502	0.635	0.595	0.567	0.556	0.616	0.528	0.531	0.566
	LipForensics [8]	0.997	0.757	0.366	0.908	0.757	0.573	0.703	0.561	0.578	0.478	0.675	0.566	0.862	0.662
	Ours	0.998	0.799	0.542	0.881	**0.805**	**0.574**	0.612	0.571	**0.596**	0.512	**0.707**	**0.592**	0.745	**0.692**
F2F	Xception [12]	0.514	0.993	0.266	0.463	0.559	0.322	0.271	0.349	0.308	0.361	0.334	0.303	0.271	0.315
	FTCN [7]	0.817	0.989	0.764	0.859	0.857	0.608	0.589	0.679	0.602	0.610	0.622	0.538	0.613	0.608
	LipForensics [8]	0.884	0.992	0.723	0.918	**0.879**	0.625	0.757	0.624	0.613	0.616	0.665	0.507	0.623	0.629
	Ours	0.845	0.996	0.705	0.919	0.866	0.618	**0.670**	**0.681**	0.600	**0.672**	**0.692**	0.479	**0.641**	**0.632**
FS	Xception [12]	0.722	0.566	0.999	0.756	0.761	0.640	0.512	0.516	0.430	0.330	0.492	0.514	0.338	0.472
	FTCN [7]	0.881	0.697	0.989	0.767	**0.834**	0.518	0.531	0.580	0.573	0.632	0.578	0.593	0.599	0.576
	LipForensics [8]	0.540	0.686	0.995	0.384	0.651	0.660	0.618	0.589	0.646	0.560	0.679	0.650	0.380	0.598
	Ours	0.781	0.677	0.998	0.772	0.807	**0.670**	0.607	0.556	**0.675**	**0.636**	0.640	**0.655**	0.542	**0.622**
NT	Xception [12]	0.845	0.774	0.391	0.987	0.749	0.533	0.466	0.540	0.517	0.547	0.455	0.554	0.639	0.531
	FTCN [7]	0.900	0.881	0.624	0.980	0.846	0.566	0.537	0.550	0.541	0.447	0.550	0.572	0.601	0.546
	LipForensics [8]	0.955	0.890	0.520	0.965	0.833	0.466	0.677	0.484	0.454	0.449	0.501	0.481	0.696	0.526
	Ours	0.929	0.901	0.622	0.998	**0.863**	**0.591**	0.563	0.540	**0.558**	0.499	**0.578**	0.562	**0.721**	**0.577**

TABLE III
STUDY ON MODULE EFFECTIVENESS

Model	Intra-dataset					Cross-dataset			
	DF	F2F	FS	NT	**Avg**	FF++	Celeb-DF	DFDC	**Avg**
GAD	0.986	0.917	0.961	0.906	0.942	0.935	0.828	0.712	0.825
LRM	0.979	0.922	0.940	0.883	0.931	0.937	0.759	0.701	0.799
GAD+LRM+MG(Stg .1)	0.989	0.931	0.965	0.908	0.948	0.955	0.857	0.759	0.841
GAD+LRM+MG(Stg .1, 2)	0.993	0.935	0.969	0.914	0.953	0.961	0.873	0.770	0.868
GAD+LRM+MG(Stg .1, 2)+CAM	0.998	0.980	0.979	0.923	**0.970**	**0.985**	**0.891**	**0.784**	**0.887**

TABLE IV
STUDY ON INFORMATION COMPLEMENTARITY

branch	Model	Intra-dataset					Cross-dataset			
		DF	F2F	FS	NT	**Avg**	FF++	Celeb-DF	DFDC	**Avg**
GAD	-	0.986	0.917	0.961	0.906	0.942	0.935	0.828	0.712	0.825
	LRM→GAD	0.994	0.929	0.968	0.919	0.953	0.943	0.847	0.744	0.845
	GAD↔LRM	0.997	0.919	0.972	0.923	0.953	**0.972**	**0.880**	**0.775**	**0.876**
LRM	-	0.979	0.922	0.940	0.883	0.931	0.937	0.759	0.701	0.799
	GAD→LRM	0.987	0.925	0.951	0.899	0.941	0.960	0.794	0.730	0.818
	GAD↔LRM	0.992	0.926	0.955	0.909	**0.946**	0.966	0.846	0.767	**0**.860

former exhibits high detection accuracies in the FF++ dataset and also has a much better generalization ability.

Study on module effectiveness. We also conducted experiments to evaluate the effectiveness of each interactive module in our method. The comparison results are presented in Table III, where GAD represents the GAD-branch, LRM represents the LRM-branch, MG denotes the Mouth Dynamics Guided module and the Global Dynamics-aware module. CAF represents the Cross Attention Fusion module. Stg. 1, 2 denotes the fusion stage.

From the table, it can be observed that: i) the two-branch model outperforms either GAD or LRM on all of the datasets, which demonstrates the global and local information promote each other effectively; ii) the model's performance and generalization capability are gradually improved with each module activated, which demonstrates the effectiveness of each interactive component.

Moreover, with all modules activated, experiments are conducted on the following variations from the perspective of complementarity between global representation and local dynamics to explore the effect of information flows from different branch : 1) isolated branch: GAD. Only the entire face sequences were adopted to train the GAD-branch and the LRM-branch is deactivated; LRM. Only the mouth region sequences were employed for training the LRM-branch and GAD-branch is deactivated. 2) Unidirectional information transmission: GAD→LRM. Unidirectional information transmission from global to local branch; LRM→GAD. Unidirectional information transmission from local to global branch. 3) Bi-directional information interaction: GAD↔LRM. Bi-directional information interaction between global and local branch.

From Table IV, it is observed that the detection performance and generalization capability of both the GAD and LRM are gradually improved as the Unidirectional and Bi-directional information interaction, which demonstrate that with CCDFM, the global branch and local branch promote the discriminative power of each other. During training, the local LRM-branch tends to provide useful features to the global view of entire face from the local anomalous regions. In other words, the global branch pays more attention to irregular mouth movements through information transmission. On the other hand, the global information integrated to mouth regions improves the capability of local perception to entire faces. These results have demonstrated that the global dynamic representation and local mouth movements are complementary to each other.

D. Comparison with SOTA methods

Generalization to unseen manipulations. To exhibit the generalization capability of our method, we first conducted cross-manipulation evaluation under both intra-dataset and cross-dataset scenarios. It is worth noting that the Validation part of Video Forgery Classification datasets in ForgeryNet [63] were adopted for the cross-manipulation evaluation under cross-dataset. ForgeryNet [63] contains diverse manipulation types where 8 forgery approaches are applied to video-forgery construction: FaceShifter, FS-GAN, DeepFakes, BlendFace, MMReplacement(MMRepla), DeepFakes-StarGAN-Stack(DF-S-S), Talking Head Video(T-H-V), and ATVG-Net. We chose all of the forgery types as negative samples, respectively, and real face dataset RAVDESS [64] used in ForgeryNet as the positive samples in our experiments. From table V, it can be observed that our method achieves excellent generalization ability to unseen manipulations under both intra-dataset and cross-dataset scenarios. In intra-dataset evaluation, all the manipulation types are from the same source

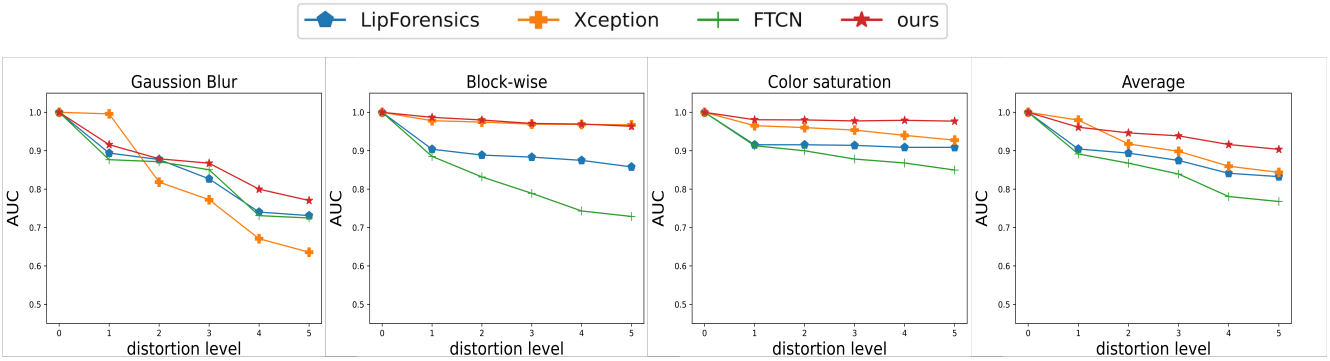

Fig. 6. Robustness to unseen perturbations.

TABLE VI
CROSS-DATASET GENERALIZATION

Training Set	Methods	FF++	Celeb-DF	DFDC	**Avg**
FF++	Xception [12]	0.997	0.659	0.690	0.782
	F^3-Net [10]	0.986	0.732	0.701	0.806
	Face X-ray [9]	**0.998**	0.795	0.655	0.816
	FTCN [7]	0.979	0.869	0.740	0.863
	LipForensics [8]	0.973	0.824	0.735	0.844
	Ours	0.985	**0.891**	**0.784**	**0.887**
Celeb-DF	Xception [12]	0.410	0.994	0.668	0.690
	F^3-Net [10]	0.628	0.990	0.710	0.776
	Face X-ray [9]	0.513	0.991	0.691	0.732
	FTCN [7]	0.721	0.993	0.747	0.820
	LipForensics [8]	0.718	0.994	0.798	0.837
	Ours	**0.723**	**0.997**	**0.802**	**0.840**

videos, thus the four compared detection methods achieve good generalization results. However, in the cross-dataset evaluation, our method outperforms the baseline Xception [12] and recent state-of-the-art LipForensics [8] on most of the manipulation types. LipForensics only analyze the mouth region while ignoring the artifacts lying outside the mouth region of the entire face. In addition, LipForensics is required to be pretrained on a large corpus of lipreading videos which introduces prior knowledge; however, in our LRM-branch, such a requirement is not necessary and thus the computational cost is reduced. Moreover, both the global fundamental representation and the mouth movements are exploited to capture spatial and temporal artifacts, thus showing better generalization capability to unseen manipulations.

Generalization to unseen datasets. To further demonstrate the generalization ability, we performed cross-dataset evaluations by training and testing model on different datasets. Table VI shows that compared with other methods investigated, our method can achieve much better generalization capability under cross-dataset scenarios. It can also be observed that the detection performance of most image-based methods, such as Xception [12], F^3-Net [10], drops much more drastically under cross-dataset scenarios compared with that of the video-based methods. It is because the static texture artifacts leaved by specific manipulation methods are relatively uniform and thus the frame-based methods based on these artifacts are vulnerable against overfitting. FTCN [7] learns feature representation from a global view, and thus is less sensitive to some forgery traces localized in critical facial regions such as mouth. On the contrary, our model aims to explore both global dynamic artifacts and local mouth movements irregularities, and exhibits better generalization performance.

TABLE VII
ROBUSTNESS TO COMPRESSION

Methods	Video-level AUC (%)		
	Raw	HQ	LQ
Xception [12]	0.998	0.993	0.920
Face X-ray [9]	0.998	0.978	0.773
F^3 Net [10]	**0.999**	**0.994**	0.958
FTCN [7]	0.997	0.979	0.963
LipForensics [8]	**0.999**	0.973	0.961
Ours	**0.999**	0.985	**0.968**

Robustness against compression. Due to the wide spreading of compressed videos on social media networks, we further conducted experiments following the setting in [8] on FF++ at different compression levels to validate the robustness of our method.

From the results of Table VII, we can observe that all models perform almost flawlessly on raw videos, but their robustness varies when trained on different compression levels. Frame-based detection algorithms tend to suffer more under compression scenarios due to the destruction of intra-frame artifacts, such as Face X-ray which relies on detecting blending boundaries and its performance drops dramatically on LQ videos. Compared with the state-of-the-art methods, our model achieves comparable performance on raw and LQ versions, which demonstrates that the proposed features are not sensitive to resolution and image noise.

Robustness to unseen perturbations. Following [7], [8], we conducted experiments to validate the robustness of our method to unseen perturbations that may be encountered in real-world scenarios. Three types of perturbations were considered: 1) color saturation change distortion, 2) local block-wise distortion, and 3) Gaussian blur distortion. Each of these distortions was divided into five intensity levels as described in [58]. We trained the models on FF++ raw datasets and evaluated them on FF++ test datasets processed by each of the perturbations. Video-level AUC scores as a function of the perturbation level for various distortions are

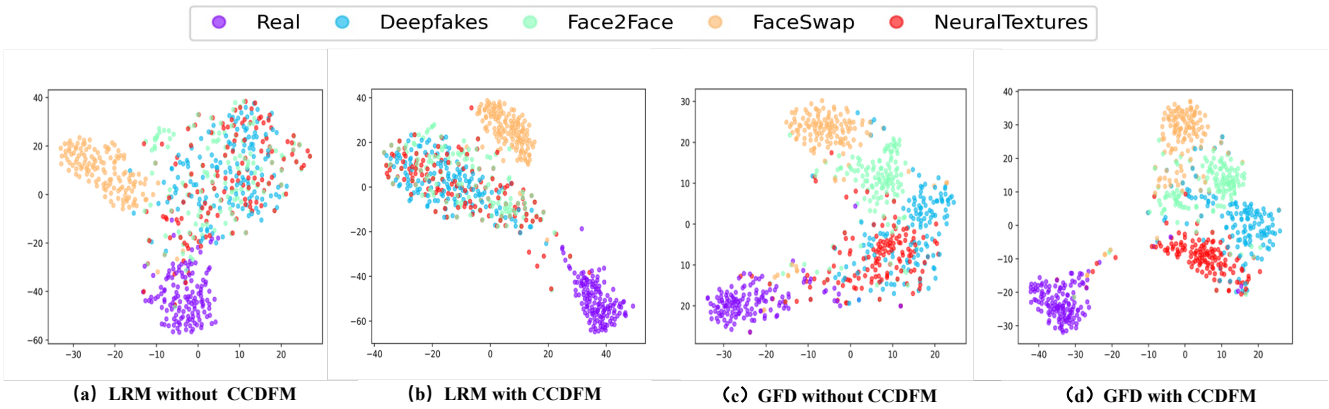

Fig. 7. Visulization of features extracted from LRM branch and GAD branch respectively.

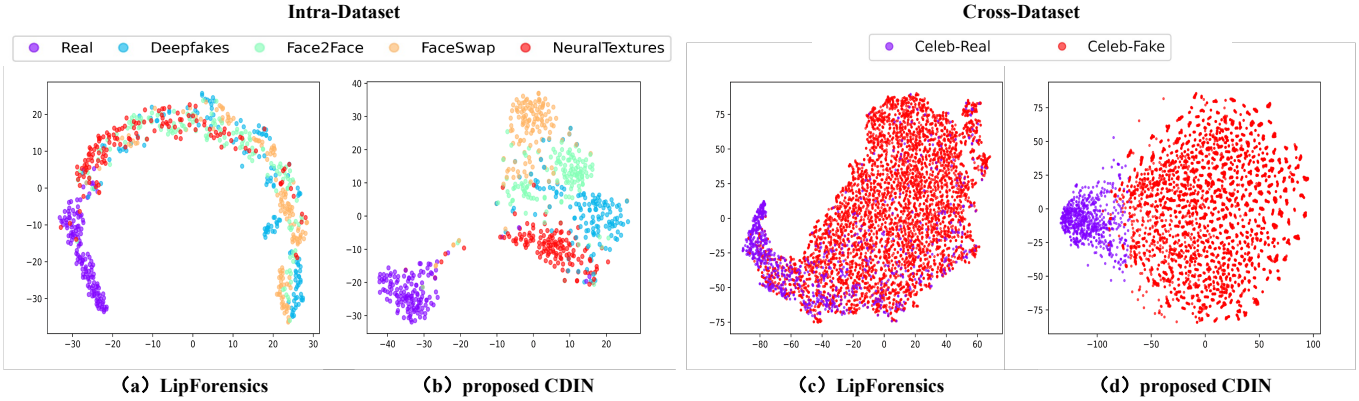

Fig. 8. The t-SNE visualization of features extracted in LipForensics and our proposed CDIN under intra-dataset and cross-dataset respectively.

illustrated in Fig. 6. Average denotes the mean AUC score across all perturbations at each distortion level. We can observe that Xception, a frame-based detection method, is vulnerable against intra-frame content destruction such as gaussian blur, while the other methods all exploit temporal information to capture dynamic artifacts thus showing better robustness against high-frequency content perturbations. Besides, LipForensics and FTCN all experience significant performance drop under block-wise distortion where our method is almost unaffected. It suggests that our method learns more robust feature representations against some common perturbations.

E. Visualization

Visualization of feature distribution. To better illustrate the effectiveness of the proposed framework, feature distributions using t-SNE were visualized.

We firstly visualized features extracted from GAD-branch, LRM-branch with and without CCDFM respectively on FF++ test dataset. The visualization results were given in Fig. 7. From the figure, it can be observed that compared with exploiting full-face dynamics or mouth movements separately, bidirectional information interaction promotes the discriminative power of single branch each other. With CCDFM, both the LRM branch and GAD branch learn better representations where the clusters for real videos and the four manipulations are separated by an obvious margin.

Moreover, we further visualized the learned feature distribution of our proposed method and LipForensics [39] trained on FF++ raw datasets and tested under within-dataset and cross-dataset scenarios respectively. The features of our method were extracted from the layer right before the FC layer in GAD-branch. In particular, a total of 700 FF++ test videos were selected for within-dataset testing, and all 5,639 videos of Celeb-DF were selected for cross-dataset testing. The visualization results are shown in Fig. 8, it is observed that our method embeds the face videos into a relatively compact feature space compared with LipForensics under both within and cross sets. Due to the missing of global perception, mouth region analysis alone struggles to learn common representations over the whole face, and thus the clusters of real and fake may be indistinguishable. On the contrary, our method captures more generalizable features as the clusters of real and fake are separated by an obvious margin under within datasets. Moreover, the features exhibit a more aggregated form compared with mouth region analysis alone under cross-dataset scenario, which exposes the discrepancy between pristine and manipulated videos. The visualization results further verify the effectiveness and generalization capability of our method to exploit both global and local dynamics in a complementary manner.

Visualization of anomalous regions. To intuitively understand the decision-making of the model, we visualize the spatial anomalous regions on which the model depends for

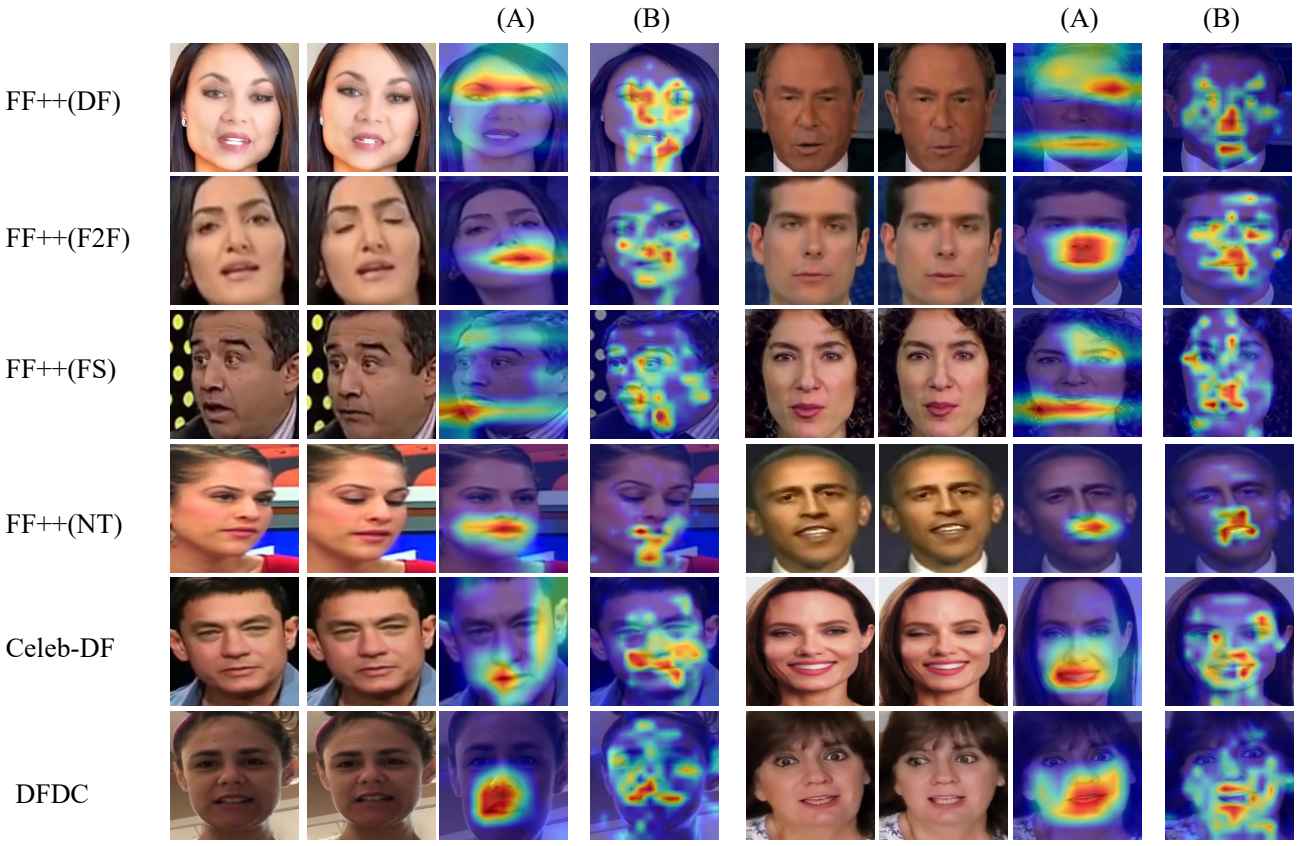

Fig. 9. Visualization of anomalous regions on different datasets. Each row shows two examples. For each example, the first two columns are consecutive frames in a video clip. (A) visualizes anomalous regions using Grad-CAM [41], (B) visualizes temporal defect using the localization method in FTCN [61]

decision-making according to the classification results of an entire face. To this end, Grad-CAM [41] is utilized, which uses the gradients flowing into the final layer to produce the attention map highlighting the decision regions. Besides, we also employed the visualization method in FTCN [61]. It localize temporally incoherent regions through discriminating whether each of the sliding window area is anomalous or not based on the sliding window mechanism across the entire face. The visualization results derived from the global branch where full-face was fed to extract global features and interact with mouth information through devised fusion module CCDFM. The visualization results against different datasets are illustrated in Fig. 9. It is observed that for different manipulations, the global features pay attention to different manipulated areas, such as eyes, eyebrow, forehead, nose, blending boundary and so on. It can also be observed that anomalous mouth dynamics are captured, which is in accordance with the motivation of our method that the global branch of full-face analysis preserve the ability to perceive the global dynamic artifacts while paying more attention to the anomalous mouth movements. Besides, when the manipulated artifacts are not obvious in mouth regions, our method will still detect DeepFakes depending on other tampered traces appeared in nose, eye regions instead of forcing the global feature to focus on mouth region only. Moreover, Fig. 9 (B) shows that our method could localized the anomalous dynamics even with subtle artifacts. The visualization results further demonstrate the effectiveness and generalization ability of our method.

V. CONCLUSION

Most recent video-based DeepFake detection methods have performed good detection results, however, they cannot yet achieve satisfactory generalization performance under cross-dataset scenario. They focus on either the entire face or a specific local regions, while we argue that both of them should be integrated comprehensively. In this paper, a novel two-branch Complementary Dynamic Interaction Network(CDIN) is proposed. Both the global (i.e. entire face) and local (i.e. mouth region) dynamics are analyzed simultaneously and comprehensively. A Cross Dynamic Fusion Module (CCDFM) is carefully designed to enhance the interation of global and local information. With CCDFM, the global branch will pay more attention on anomalous mouth movements and the local branch will gain more information from the global context. The experiment results demonstrate that our method achieves excellent detection performance compared with several SOTA methods, especially exhibits superior generalization capability to unseen manipulations.

VI. ACKNOWLEDGEMENTS

The work described in this paper was supported by the National Natural Science Foundation of China (Grant 62271307, and Grant 61771310). Our gratitude goes to the anonymous reviewers for their efforts. We especially thank Yinglin Zheng, the author of FTCN, for the help of codes reproduction.

REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[3] Deepfakes, 2017. [Online]. Available: https://github.com/deepfakes/faceswap

[4] Faceswap, 2017. [Online]. Available: https://github.com/MarekKowalski/FaceSwap

[5] Faceapp, 2017. [Online]. Available: https://www.faceapp.com

[6] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 317–16 326.

[7] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 044–15 054.

[8] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5039–5049.

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

[10] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *European Conference on Computer Vision*. Springer, 2020, pp. 86–103.

[11] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *European Conference on Computer Vision*. Springer, 2020, pp. 667–684.

[12] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," *arXiv preprint arXiv:1803.09179*, 2018.

[13] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A new dataset for deepfake forensics," 2019.

[14] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[15] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2382–2390.

[16] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 484–492.

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

[18] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proceedings of the 4th ACM workshop on information hiding and multimedia security*, 2016, pp. 5–10.

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

[20] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking deepfakes with simple features," *arXiv preprint arXiv:1911.00686*, 2019.

[21] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7890–7899.

[22] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 772–781.

[23] K. Chandrasegaran, N.-T. Tran, and N.-M. Cheung, "A closer look at fourier spectrum discrepancies for cnn-generated images detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7200–7209.

[24] X. Wu, Z. Xie, Y. Gao, and Y. Xiao, "Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2952–2956.

[25] C. M. LIY and L. InIctuOculi, "Exposingaicreated fakevideosbydetectingeyeblinking," in *2018IEEEInterG national Workshop on Information Forensics and Security (WIFS). IEEE*, 2018.

[26] M. Li, B. Liu, Y. Hu, and Y. Wang, "Exposing deepfake videos by tracking eye movements," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5184–5189.

[27] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based cnn," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[28] I. Amerini and R. Caldelli, "Exploiting prediction error inconsistencies through lstm-based classifiers to detect deepfake videos," in *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*, 2020, pp. 97–102.

[29] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, no. 1, pp. 80–87, 2019.

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

[31] Y. Zhou and S.-N. Lim, "Joint audio-visual deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 800–14 809.

[32] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656*, 2018.

[33] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding," *arXiv preprint arXiv:2102.05095*, vol. 2, no. 3, p. 4, 2021.

[34] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

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

[37] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.

[38] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and vision computing*, vol. 47, pp. 3–18, 2016.

[39] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[40] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

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

[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[43] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[44] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[45] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[46] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.

[47] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," *arXiv preprint arXiv:2106.13230*, 2021.

[48] N. Clova, "Clova face recognition (cfr)-celebrity face recognition api," *Retrieved October-14-2020 from https://apidocs. ncloud. com/en/ai-naver/clova_face_recognition*, 2020.

[49] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting deep-fake videos from phoneme-viseme mismatches," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 660–661.

[50] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, and X. Guo, "Facial: Synthesizing dynamic talking face with implicit attribute learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3867–3876.

[51] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang, "Ad-nerf: Audio driven neural radiance fields for talking head synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5784–5794.

[52] J. Hu, X. Liao, W. Wang, and Z. Qin, "Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1089–1102, 2021.

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

[54] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 710–18 719.

[55] J. Yang, S. Xiao, A. Li, W. Lu, X. Gao, and Y. Li, "Msta-net: forgery detection by generating manipulation trace based on multi-scale self-texture attention," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[56] B. Chen, X. Liu, Y. Zheng, G. Zhao, and Y.-Q. Shi, "A robust gan-generated face detection method based on dual-color spaces and an improved xception," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[57] H. Xie, J. Ni, and Y.-Q. Shi, "Dual-domain generative adversarial network for digital image operation anti-forensics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1701–1706, 2021.

[58] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2889–2898.

[59] H. Chen, Y. Lin, B. Li, and S. Tan, "Learning features of intra-consistency and inter-diversity: Keys towards generalizable deepfake detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[60] Z. Hu, H. Xie, Y. Wang, J. Li, Z. Wang, and Y. Zhang, "Dynamic inconsistency-aware deepfake video detection," in *IJCAI*, 2021.

[61] X. Zhao, Y. Yu, R. Ni, and Y. Zhao, "Exploring complementarity of global and local spatiotemporal information for fake face video detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2884–2888.

[62] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[63] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu, "Forgerynet: A versatile benchmark for comprehensive forgery analysis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4360–4369.

[64] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

Hanyi Wang received the B.Eng. degree from Xidian University, China, in 2021. She is currently pursuing the Ph.D. degree with the School of Electric Information and Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. Her research interests include image forensics and computer vision.

Zihan Liu received the B.Eng. degree from Shanghai Jiao Tong University, Shanghai, China, in 2022. He is currently pursuing the M.Eng. degree with the School of Electric Information and Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include computer vision and pattern recognition.

Shilin Wang (Senior Member, IEEE) received the B.Eng. degree in electrical and electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2001, and the Ph.D. degree from the Department of Computer Engineering and Information Technology, City University of Hong Kong, in 2004. Since 2004, he has been with the School of Electric Information and Electronic Engineering, Shanghai Jiao Tong University, where he is currently a Professor. His research interests include image processing and pattern recognition.