Dynamic Inconsistency-aware DeepFake Video Detection

Ziheng Hu[1], **Hongtao Xie**[1*], **Yuxin Wang**[1], **Jiahong Li**[2], **Zhongyuan Wang**[2] and **Yongdong Zhang**[1]

[1]University of Science and Technology of China
[2]Kuaishou Technology
{hzh519, wangyx58}@mail.ustc.edu.cn, {htxie, zhyd73}@ustc.edu.cn, {lijiahong, wangzhongyuan}@kuaishou.com

Abstract

The spread of DeepFake videos causes a serious threat to information security, calling for effective detection methods to distinguish them. However, the performance of recent frame-based detection methods become limited due to their ignorance of the inter-frame inconsistency of fake videos. In this paper, we propose a novel Dynamic Inconsistency-aware Network to handle the inconsistent problem, which uses a Cross-Reference module (CRM) to capture both the global and local inter-frame inconsistencies. The CRM contains two parallel branches. The first branch takes faces from adjacent frames as input, and calculates a structure similarity map for a global inconsistency representation. The second branch only focuses on the inter-frame variation of independent critical regions, which captures the local inconsistency. To the best of our knowledge, this is the first work to totally use the inter-frame inconsistency information from the global and local perspectives. Compared with existing methods, our model provides a more accurate and robust detection on *FaceForensics++*, *DFDC-preview* and *Celeb-DFv2* datasets.

1 Introduction

With recent advances of deep learning theories like GANs [Goodfellow *et al.*, 2014], a series of technologies commonly known as DeepFake have emerged, which allow automation of facial expression transformation and face swapping possible. Though these technologies can make some entertainment videos, they could also be used to make misinformation for fraudulent or malicious purposes. Thus, there is an urgent need for automatic detection methods.

As shown in Figure 1, different from the natural variation between real video frames, existing face manipulation tools which operate on single frame destroy the consistency between adjacent frames and lead to pixel jitter in fake videos. In this paper, we define this problem as an Inconsistent Problem, this is an important attribute that distinguishes fake videos from real videos. Early works [Fridrich, 2012] focus on extracting hand-crafted features and employing classifiers (e.g., SVM) to detect fake videos. These traditional methods are less effective with more realistic faces generated. Benefiting from the development of deep learning, some recent works [Yang *et al.*, 2019; Li *et al.*, 2018] use Convolutional Neural Network (CNN) on forgery detection by analysis of physiological characteristics (e.g., headpose, eye blink). Besides, some data-driven CNN-based methods [Chollet, 2017; Afchar *et al.*, 2018] are also popular, which take a large amount of independent frames to train classifiers without specific physiological characteristics. In summary, the Inconsistent Problem of fake videos has not received enough attention in existing detection methods. The performance of these methods become limited without considering inter-frame inconsistency information.

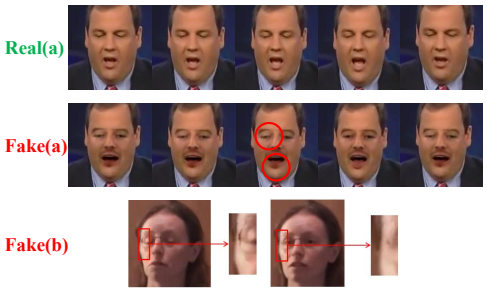

Figure 1: The adjacent frames of the real video change naturally, while the fake video has the inter-frame inconsistency. As shown in example Fake(a), although each frame is quite realistic, there is pixel jitter around eyes and mouth (can be obviously seen by playing these frames continuously). Example Fake(b) shows the inconsistency of artifacts of the same area between adjacent frames.

In this paper, we propose a novel Dynamic Inconsistency-aware Network for DeepFake video detection by utilizing the inconsistency information between adjacent frames. As shown in Figure 2, the DIANet consists of three modules: Feature Extracting module, Cross-Reference module(CRM) and Classification network. Specifically, the DIANet takes a pair of frames as input and obtains their feature representations through the Feature Extracing module. Then the proposed CRM is adopted to capture both the global and local inconsistencies between adjacent frames. The CRM contains two parallel branches which are called the Global Correlation

*Corresponding Author.

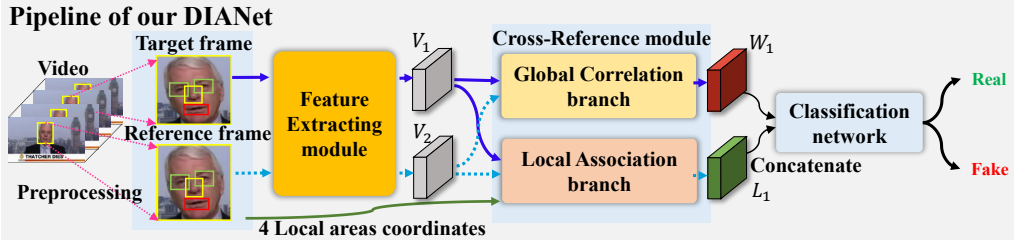

Figure 2: The pipeline of the DIANet.

branch (GCB) and the Local Association branch (LAB). The GCB takes faces from adjacent frames as input, and calculates a structure similarity map for global inconsistency representation. In contrast, the LAB only focuses on inter-frame association of independent critical regions to capture local inconsistency. Finally, the global and local inter-frame inconsistencies are combined together and sent into the Classification network for classification.

In experiments, we quantitatively verify the effectiveness of utilizing inconsistency information between adjacent frames. Benefiting from exploring the global and local inconsistencies, our model outperforms existing methods on *FaceForensics++* (FF++) [Rossler *et al.*, 2019], *DFDC-preview* [Dolhansky *et al.*, 2019] and shows a good rubustness to degradation of video quality and unseen manipulation techniques.

The major contributions of our paper include:

- The DIANet is specially designed to explore the inter-frame inconsistency information, which provides a more powerful approach to model the inter frame relationship.
- A novel Cross-Reference module is proposed to capture the global structure inconsistency and local inconsistency through two parallel branches, which aims to give a complete and robust representation to the inconsistency information between two adjacent frames.
- Our model outperforms existing methods on popular datasets and generalizes well on videos of low quality and unseen manipualtion techniques.

2 Related Work

2.1 DeepFake Video Detection

Current DeepFake detection algorithms roughly fall into two branches: image-based methods and video-based methods.

Image-based Methods

Early image-based methods [Fridrich, 2012] are driven by hand-crafted features or statistical artifacts that occur during image formation. Some classifiers (e.g., SVM) are employed to judge whether a video is real or fake. However, fake faces are more and more realistic with the emergence of new manipulation tools, traditional image-based methods become less effective.

With the development of deep learning[Chollet, 2017; He *et al.*, 2016; Wang *et al.*, 2020a; Wang *et al.*, 2020b], CNN-based methods[Yang *et al.*, 2019; Afchar *et al.*, 2018; Zhou *et al.*, 2017; Rossler *et al.*, 2019; Shang *et al.*, 2021] become popular. Part of CNN-based methods are based on artifacts of physiological characteristics. [Yang *et al.*, 2019] proposed an approach to detect by utilizing incoherent head poses in fake videos. Another part of methods are data-driven without relying on any specific physiological characteristics. Researchers are trying various CNN structures, such as Xception [Chollet, 2017], MesoNet [Afchar *et al.*, 2018] and Two-stream [Zhou *et al.*, 2017]. Some methods [Rahmouni *et al.*, 2017; Bayar and Stamm, 2016] put forward some improvments on the convolutional layer or pooling layer. Many methods use transfer learning and fine-tuning to take advantage of pretrained models [Raja *et al.*, 2017; Cozzolino *et al.*, 2018]. Further more, [Nguyen *et al.*, 2019] segments the tampered region while classifying. However, these methods rely on features within a single frame, without consideration of correlation information across frames.

Video-based Methods

The second category of detection methods are based on a sequence of frames. [Li *et al.*, 2018] developed a network for detecting eye blinking by using a temporal approach. Other video-based methods include using Recurrent Neural Network (RNN) [Sabir *et al.*, 2019]. These methods take a sequence of frames as input to RNN to tell whether the video is real or fake. Though some temporal information is utilized, they ignore the specific inconsistency information between adjacent frames of fake videos. Different from them, our method explore the inconsistency information by a novel cross-reference module from the global and local perspectives.

2.2 Attention Mechanism in Neural Network

Various attention mechanisms which are inspired by human perception [Denil *et al.*, 2012] have been widely studied. The attention mechanism enables neural network to pay attention to a specific subset of inputs. In addition, many works propose spatial and channel-wise mechanism to pay attention to part of image dynamically. In this way, the trained models can focus on regions or segments selectively. Different from these works, our cross reference mechanism is used to capture the global and local relationship across different frames.

3 Proposed Method

We propose the DIANet for effective DeepFake video detection, by utilizing inconsistency information between adjacent

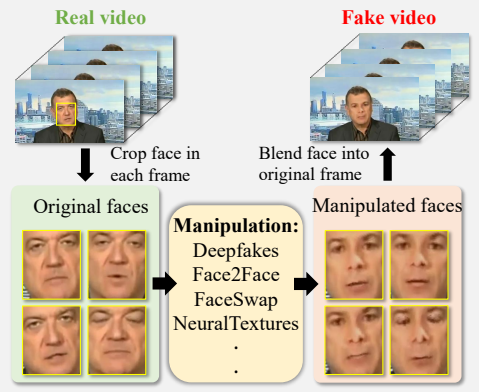

Figure 3: Overview of a typical video manipulation process.

frames with account of the whole face and the local critical areas. In this section, we first introduce the preliminary knowledge which helps to understand our motivation. Then we describe the overall pipeline of the DIANet. And then we will introduce the proposed CRM in more detail. Finally, we introduce how the Classification network gives the classification result.

3.1 Preliminary Knowledge

With recent advancement of techniques[DeepFakes, 2019; Thies *et al.*, 2016; FaceSwap, 2019; Yu *et al.*, 2019; Yu *et al.*, 2020], it is possible for people to conduct facial expression transformation and face swapping automatically. The most popular methods are Deepfakes [DeepFakes, 2019], Face2Face [Thies *et al.*, 2016] and FaceSwap [FaceSwap, 2019]. A typical video face manipulation process consists of three steps: (1) Decompose the original video into frames. Detect face in each frame and crop it. (2)Manipulate face images by the pretrained manipulation model. (3) Blend each manipulated face back into the corresponding original frame.

Because the manipulation is operated on independent frames, the natural variation of face in the original video is destroyed. This results in micro dynamic jitters and inconsistency across adjacent frames which are ignored by existing detection methods.

3.2 Pipeline of DIANet

As shown in Figure 2, the proposed DIANet consists of three modules: (1) Feature Extracting module. (2) Cross-Reference module (CRM). (3) Classification network. We call the frame to be distinguished as the target frame. Firstly, we select one adjacent frame as the reference frame and preprocess these two frames. The preprocessing includes croping the whole face region and detecting the coordinates of critical areas. All the following "frames" refer to the face region image in the corresponding frame. Secondly, the Feature Extracting module extracts feature representations of two frames respectively. The Feature Extracting module is based on ResNet [He *et al.*, 2016]. There are five Bottleneck layers, designated as $layer1$ to $layer5$. Each Bottleneck layer consists of three convolutional layers with *BatchNormalization* and *ReLU*. We take the feature map after Bottleneck $layer5$ as the output of the Feature Extracting module. The feature representations of the target and the reference frames are marked as V_1 and V_2 respectively.

Thirdly, we calculate the inconsistency map X_1 based on feature representations V_1 and V_2 through the CRM. Finally, the Classification network gives the classification result Y of the target frame according to the inconsistency map X_1. The specific implementation of the CRM and the Classification network will be described in more detail as follows.

3.3 Cross-Reference Module (CRM)

As shown in Figure 2, the proposed CRM uses the feature representations V_1 and $V_2 \in R^{C\times H\times W}$ obtained by the Feature Extracting module and the coordinates of four local critical areas as input. It includes two branches: the Global Correlation branch (GCB) and the Local Association branch (LAB). The GCB uses features of whole faces in two frames to generate the structure similarity map W_1. In contrast, the LAB first implements RoIPooling to obtain four local critical areas features (i.e. two eyes, nose, mouth) with the same size, and then calculate the local inconsistence features L_1 which reflects regional inconsistency of adjacent frames. In the end, L_1 are concatenated with W_1 to generate the inconsistency map X_1, as the output of the CRM:

$$X_1 = concatenate(W_1, L_1) \tag{1}$$

Global Correlation Branch (GCB)

To explore the global long-range inconsistency information between the input two frames, The GCB takes feature representations of whole faces V_1 and V_2 as input. The procedure of the GCB is illustrated in Figure 4(a). More specifically, we first compute the affininty matrix A_{12} between V_1 and V_2:

$$A_{12} = V_1^T Q V_2 \tag{2}$$

where V_1 and V_2 are flattened to $C\times(WH)$. $Q \in R^{C\times C}$ is the weight matrix and can be implemented by a linear layer. As a result, each entry of A_{12} reflects the similarity between each row of V_1^T and each column of V_2.

After obtaining the affininty matrix A_{12}, we compute attention summaries Z_{12} for V_1:

$$Z_{12} = V_1 A_{12} \tag{3}$$

Then a convolution operation with N filters $\{\omega_G^i, b_G^i\}_{i=1}^N$ is performed for V_1. For the generation of correspongding N heatmaps, another group of filters $\{\omega_Z^i, b_Z^i\}_{i=1}^N$ is applied for Z_{12}. They are calculated as follows:

$$G_1^i = V_1 \omega_G^i + b_G^i, \quad for \quad i = 1, 2, ..., N \tag{4}$$

$$I_{12}^i = Softmax(Z_{12}\omega_Z^i + b_Z^i), \quad for \quad i = 1, 2, ..., N \tag{5}$$

The heatmaps are utilized to perform channel-wise selection as follows:

$$I_1 = concatenate(G_1^1 \odot I_{12}^1, ..., G_1^N \odot I_{12}^N) \tag{6}$$

where $'\odot'$ denotes the element-wise multiply. In order to keep consistent with the output size of the LAB, a maximum pooling is performed:

$$W_1 = MaxPool(I_1) \tag{7}$$

W_1 represents the structure similarity map which reflects the global inconsistency. The dimension of W_1 is $C\times H'\times W'$.

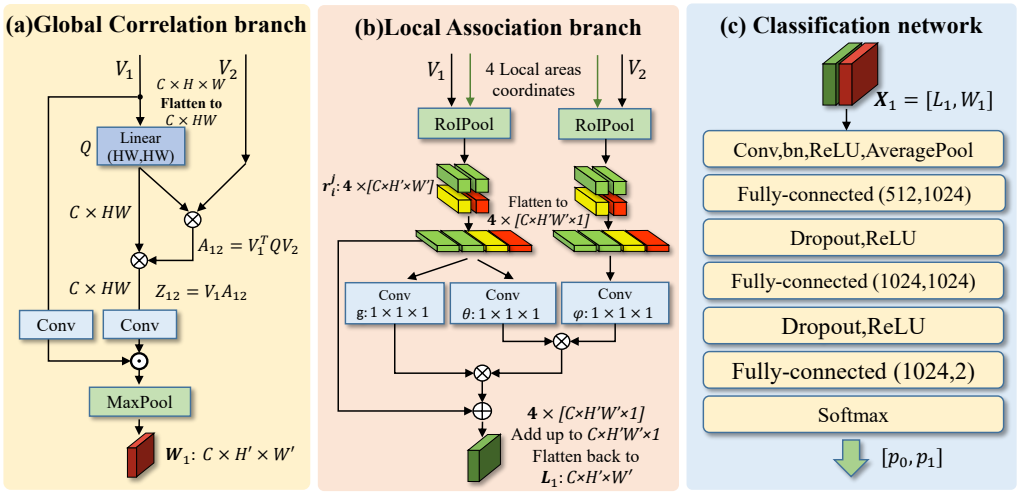

Figure 4: The details of Global Correlation branch(GCB), Local Association branch(LAB) and Classification network. $'\otimes'$ represents matrix multiply. $'\oplus'$ represents element-wise add. $'\odot'$ denotes the element-wise multiply.

Local Association Branch (LAB)

Face manipulation tools often produce different fake textures at the same area in adjacent frames. As a result, fake textures of local areas are inconsistent. Motivated by this, we design the LAB to mine inconsistency of independent critical areas across frames. The local semantic areas which the LAB focuses on are two eyes, nose and mouth whose positions are extracted by RetinaFace [Deng *et al.*, 2020] in the preprocessing.

The procedure of the LAB is illustrated in the Figure 4(b). For V_1 and V_2 whose dimension is $C \times H \times W$, we first implement RoIPooling of four local semantic areas based on their positions to obtain eight local represention r_i^j with the same dimension:

$$r_i^j = RoIPool(V_i, l_j) \quad for \quad i = 1, 2, \quad j = 1, 2, 3, 4 \quad (8)$$

where l_j is a four-dimension vector which represents coordinates of area j. The dimension of r_i^j is set as $C \times H' \times W'$. Then we flatten r_i^j to $C \times H'W' \times 1$ and concatenate these four RoI features in axis-1 to get $r_1, r_2 \in R^{C \times 4H'W' \times 1}$ respectively.

Next, the relationship matrix between r_1 and r_2 is calculated:

$$v = Softmax(r_1 \Omega_\theta^T \Omega_\phi r_2^T) \quad (9)$$

where Ω_θ and Ω_ϕ represents that r_1 and r_2 pass through a $1 \times 1 \times 1$ convolutional layer respectively and the dimension of v is $C \times 4H'W' \times 4H'W'$.

Then the local discriminating features can be obtained:

$$r_1' = r_1 + \Omega_z(g(r_1)v) \quad (10)$$

Here the dimension of r_1' is $C \times 4H'W' \times 1$. we split r_1' into four $C \times H'W' \times 1$ matrixs and add them up to a $C \times H'W' \times 1$ matrix. Finally, we flatten this $C \times H'W' \times 1$ matrix back to $C \times H' \times W'$ as the local inconsistency information L_1.

In summary, the GCB takes the whole face features as input to calculate the structure similarity map W_1, while the LAB only focuses on the regional discriminating features to capture local inconsistency information L_1.

3.4 Classification Network

W_1 and L_1 are concatenated together and sent into the Classification network for classification. As shown in Figure 4(c), the Classification network consists of one convolutional layer and three fully connected layers, equipped with *BatchNormalization*, *average pooling*, *ReLU*, *Dropout* and *Softmax*. The specific size and parameter settings of every layer are detailed in the figure. The output of the Classification network is a two dimensional vector $[p_0, p_1]$ and the final result $Y \in [0, 1]$is calculated by $p_0/(p_0 + p_1)$. The closer Y is to 1, the more likely the target frame has been manipulated.

In the training phase, we use cross-entropy (CE) loss:

$$Loss_{CE}(p, y) = -\frac{1}{2} \sum_{i=0}^{1} [y_i log(p_i) + (1 - y_i) log(1 - p_i)] \quad (11)$$

Where $[p_0, p_1]$ is the output of the Classification network. $[y_0, y_1]$ is the ground truth of the input frames. If the input frames are sampled from a real video, $[y_0, y_1] = [0, 1]$. In contrast, a fake video is labeled as $[y_0, y_1] = [1, 0]$.

4 Experiments

4.1 Experiment Setup

Dataset Setting

To verify the effectiveness and generalization of the proposed DIANet, we conduct experiments on multiple datasets: FaceForensics++ (FF++) [Rossler *et al.*, 2019], Celeb-DeepFake v2 (Celeb-DFv2) [Li *et al.*, 2020b] and DeepFake Detection Challenge preview (DFDC-preview) [Dolhansky *et al.*, 2019]. FF++ is a large scale dataset consisting of 1000 real videos that have been manipulated with four popular methods: Deepfakes (DF) [DeepFakes, 2019], Face2Face (F2F) [Thies *et al.*, 2016], FaceSwap (FS) [FaceSwap, 2019] and NeuralTextures (NT) [Thies *et al.*, 2019]. Celeb-DFv2 is another large scale DeepFake video dataset with many different subjects (e.g., ages, ethic groups, gender), including 590 real videos and 5639 fake videos with reduced visual arifacts. DFDC-preview is a dataset for Facebook DeepFake Detection

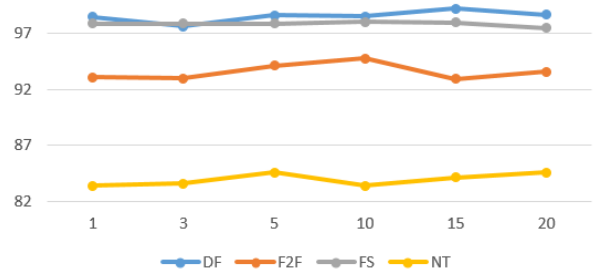

Figure 5: Evaluation of different reference frame intervals. The metric standard is frame-level AUC.

	w/o eyes	w/o nose	w/o mouth	all
AUC	94.54	94.58	94.90	**95.23**

Table 1: Evaluation of using four critical regions of the LAB on the FF++ (LQ) dataset. The metric standard is frame-level AUC.

Challenge, including 855 real videos and 3618 fake videos. Training and testing sets are divided according to the official video lists respectively. All the characters in our illustrations come from these public datasets. In the preprocessing, we use OpenCV to extract each frame in the video. For each frame, we detect the face region and four key points (two eyes, nose and mouth) by RetinaFace [Deng *et al.*, 2020].

Implementation Detail

The input images are resized to $3\times224\times224$. The dimension of the feature representation V_1 and V_2 is $256\times29\times29$. The output dimension of RoIPooling (i.e. the size of W_1 and L_1) is $256\times14\times14$. The GCB and LAB are trained with SGD. The SGD optimizer is used with an initial learning rate of 1×10^{-3} with momentum of 0.9 and weight decay of 1×10^{-4}. During training, the batchsize is set to 32.

4.2 Ablation Study

Ablation Study on Selection of Reference Frame

The DIANet takes two frames as input. One frame is called the target frame, and another is called the reference frame. After the target frame is determined, the selection of reference frame will affect the prediction result. The frame rate of videos in datasets is 30 frames per second. We test different selection intervals, from 1 frame to 30 frames. As shown in Figure 5, in general, the performance is not very sensitive to different selection intervals. The best effect is got in the interval between 5 and 15 frames for different manipulation types. On the one hand, if the interval between these two frames is too close, the pixel change is too small to capture the inconsistency information. On the other hand, if the interval is long, the face may have a change in action. In order to consider different kinds of manipulation types and get the best balanced effect, we randomly select the reference frame between 5 and 15, corresponding to 1/6 second and 1/2 second respectively.

GCB	LAB	DF	F2F	FS	NT
-	-	98.60	94.02	97.46	83.51
✓	-	98.71	94.35	97.71	84.76
-	✓	97.48	91.22	96.29	82.01
✓	✓	99.27	95.91	98.79	87.57

Table 2: Evaluation of the effectiveness of the GCB and the LAB on the FF++ (LQ) dataset. The metric standard is frame-level AUC.

Method	FF++(LQ) Acc(%)	FF++(LQ) AUC(%)	FF++(HQ) Acc(%)	FF++(HQ) AUC(%)
Steg.Features	55.98	-	70.97	-
LD-CNN	58.69	-	78.45	-
Constrained Conv	66.84	-	82.97	-
MesoNet	70.47	-	83.10	-
Xception	82.71	89.3	95.04	96.3
DSP-FWA	-	59.2	-	56.9
Face X-ray	-	61.6	-	87.4
F3-Net	86.89	93.3	**97.31**	98.1
Two-branch	-	86.6	-	98.7
Ours	**89.77**	**94.5**	96.37	**98.8**

Table 3: Performance on the FF++ HQ and LQ dataset.

Ablation Study on Critical Regions

Since the manipulated textures are inconsistent at the same area in adjacent frames, we design the LAB to explore the association of several independent critical regions. The facial expressions of characters are mainly determined by various organs. Speaking needs movement of mouth and blinking needs movement of eyes. The artifacts of mouth and eyes in the manipulated video often spread to the nose area connected with them. Thus we conduct experiments with these four regions of the face (i.e. two eyes, nose, mouth). As shown in Table 1, when the LAB uses all these four regions, AUC is the highest. When we reduce any of them, the effect decreases in varying degrees. This reflects that all the four regions have a positive effect.

Ablation Study on Each Branch

As shown in Table 2, to verify the effectiveness of the CRM with the GCB and the LAB, we compare four structures. When both two branches are not added, the feature representation of the target frame is directly used for classification. In this situation, it only gains average AUC of 93.40%, similar to common structures like MesoNet and Xception. After adding the GCB, the performance shows a considerable improvement. The AUC of each kind tampered video is increased by 0.49% on average. Only using the LAB gets the worst performance. This is because only features of eyes, nose and mouth are used while the rest of the features are discarded. However, when the local inconsistency is combined with the global inconsistency, best improvement can be achieved. The best results are obtained by adding all the two branches. This demonstrates the effectiveness of the CRM and mutual promotion of global and local inconsistencies.

Method	Acc(%)	AUC(%)
RNN	78.74	82.41
Xception	82.87	86.96
Ours	**85.83**	**90.54**

Table 4: Performance on the DFDC-preview dataset.

Method	Test set FF++(LQ)	Test set Celeb-DFv2
Two-stream	-	53.8
MesoNet	-	54.8
HeadPose	-	54.6
Multi-task	62.0	54.3
Xception-RAW	-	48.2
Xception-HQ	87.3	65.3
Xception-LQ	-	65.5
DSP-FWA	62.0	64.6
Face X-ray	72.8	-
Ours	**90.4**	**70.4**

Table 5: Experiments of cross-quality and cross-dataset experiments. The second column are models trained on FF++ (HQ) and tested on FF++ (LQ). The third column are models trained on FF++ and tested on Celeb-DFv2. The metric standard is frame-level AUC.

4.3 Comparison with Recent Works

We first conduct in-dataset experiments on FF++ and compare with prior famous works such as MesoNet [Afchar *et al.*, 2018], Xception [Rossler *et al.*, 2019] and Face X-ray [Li *et al.*, 2020a]. To get a comprehensive evaluation, we report both Accuracy (Acc) and Area Under Curve (AUC) of each method. FF++ contains three grades of video quality: raw qualtiy (RAW), high quality (HQ) and low quality (LQ). On the RAW videos, almost all CNN-based methods including ours can achieve quite high results (> 0.99), therefore it can only have a slight improvment. However, on the HQ and LQ videos, our method has a obvious improvment. As shown in Table 3, our method gains the highest AUC of 98.8% on HQ and 94.5% on LQ videos, even competitive with the latest algorithms F3-Net[Qian *et al.*, 2020] and Two-branch[Masi *et al.*, 2020].

Performance on DFDC-preview

We also conduct experiments on the latest DFDC-preview dataset. As shown in Table 4. We reproduce the RNN structure according to the paper [Sabir *et al.*, 2019]. Compared to Xception and RNN, the DIANet gains the highest Accuracy and AUC. This further proves the applicability of the DIANet.

Robustness to Video Quality

In real internet scenes, fake videos will be uploaded and downloaded many times in the process of spreading, video quality will be gradually reduced. Many manipulated artifacts are blurred and some noises are introduced with the decrease of image quality. This will misleads the extracted features and leads to wrong judgment. As shown in Table 3, the DIANet performs better than previous methods on the LQ videos, which demonstartes our method has stronger detection capability for heavily low quality videos. In addition, we use the models trained on a certain video quality to detect videos of other qualities. As shown in the second column of Table 5, part of methods like DSP-FWA and Face X-ray, when they are trained on high quality videos and tested on low quality videos, their performance decreased seriously. Their AUC drops below 75%, and the accuracy will be close to random judgment. These results reflect that the features concerned by these existing methods are easily to be destroyed by the degradation of image quality. Compared with them, the DIANet still maintains considerable detection ability in the cross-quality experiments. This shows that the global and local inconsistencies are more robust to video quality.

Robustness to Unknown Techniques

We use Celeb-DFv2 dataset for testing generalization ability of the proposed DIANet and do a more comprehensive comparison with existing works. We use four types of manipulation videos in FF++ to train models and test them on Celeb-DFv2 directly. AUC score is used as the metric for evaluating our approach. The third column of Table 5 shows the performance of existing methods and the DIANet in spotting fake videos on Celeb-DFv2. Results show that the DIANet reaches an AUC score of 70.4% on the testing set provided in Celeb-DFv2 and outperforms all the existing works including Xception, MesoNet and DSP-FWA. According to the results in Table 5, when current models trained on a specific dataset detect some fake videos utilizing various unknown techniques, their effect will be greatly reduced. This reflects cross-dataset detection is still a challenging task.

5 Conclusion

Motivated by that the inconsistent problem of fake videos has not received enough attention in existing face manipulation video detection methods, in this work we propose the DIANet to explore the inter-frame inconsistency information. The GCB and the LAB in the CRM are adopted to integrate both the global and local inconsistencies. We conducted extensive experiments on popular datasets (i.e., *FaceForensics++*, *DFDC-preview* and *Celeb-DFv2*). The results demonstrate the effectiveness of the CRM. Compared with existing methods, the DIANet obtains competitive performance on various qualities of fake videos and have strong robustness to degradation of video quality and unseen manipulation techniques. The inconsistency is an import characteristic of fake videos, which still has a lot of mining space in the future to improve the detection performance.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (2018YFB0804203), the National Nature Science Foundation of China (62022076, U1936210, 62032006), the China Postdoctoral Science Foundation (2020M682035),the Anhui Postdoctoral Research Activities Foundation (2020B436), the Youth Innovation Promotion AssociationChinese Academy of Sciences (2017209), and the Fundamental Research Funds for the Central Universities under Grant WK3480000011.

References

[Afchar *et al.*, 2018] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *WIFS*, pages 1–7. IEEE, 2018.

[Bayar and Stamm, 2016] Belhassen Bayar and Matthew 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*, pages 5–10, 2016.

[Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017.

[Cozzolino *et al.*, 2018] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018.

[DeepFakes, 2019] DeepFakes. http://www.github.com/deepfakes/faceswap, 2019. Accessed:September 18, 2019.

[Deng *et al.*, 2020] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, pages 5203–5212, 2020.

[Denil *et al.*, 2012] Misha Denil, Loris Bazzani, Hugo Larochelle, and Nando de Freitas. Learning where to attend with deep architectures for image tracking. *Neural computation*, 24(8):2151–2184, 2012.

[Dolhansky *et al.*, 2019] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

[FaceSwap, 2019] FaceSwap. www.github.com/MarekKowalski/FaceSwap, 2019. Accessed:September 30, 2019.

[Fridrich, 2012] Jessica Fridrich. Rich models for steganalysis of digital images. *IEEE TIFS*, 7(3):868–882, 2012.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *In NIPS*, 27:2672–2680, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Li *et al.*, 2018] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *WIFS*, pages 1–7. IEEE, 2018.

[Li *et al.*, 2020a] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *CVPR*, pages 5001–5010, 2020.

[Li *et al.*, 2020b] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *CVPR*, pages 3207–3216, 2020.

[Masi *et al.*, 2020] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *ECCV*, pages 667–684. Springer, 2020.

[Nguyen *et al.*, 2019] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019.

[Qian *et al.*, 2020] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, pages 86–103. Springer, 2020.

[Rahmouni *et al.*, 2017] Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *WIFS*, pages 1–6. IEEE, 2017.

[Raja *et al.*, 2017] Kiran Raja, Sushma Venkatesh, RB Christoph Busch, et al. Transferable deep-cnn features for detecting digital and print-scanned morphed face images. In *CVPR Workshops*, pages 10–18, 2017.

[Rossler *et al.*, 2019] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019.

[Sabir *et al.*, 2019] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3(1), 2019.

[Shang *et al.*, 2021] Zhihua Shang, Hongtao Xie, Zhengjun Zha, Lingyun Yu, Yan Li, and Yongdong Zhang. Prrnet: Pixel-region relation network for face forgery detection. *Pattern Recognition*, 116:107950, 2021.

[Thies *et al.*, 2016] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, pages 2387–2395, 2016.

[Thies *et al.*, 2019] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM TOG*, 38(4):1–12, 2019.

[Wang *et al.*, 2020a] Yuxin Wang, Hongtao Xie, Zheng-Jun Zha, Youliang Tian, Zilong Fu, and Yongdong Zhang. R-net: A relationship network for efficient and accurate scene text detection. *IEEE TMM*, 2020.

[Wang *et al.*, 2020b] Yuxin Wang, Hongtao Xie, Zheng-Jun Zha, Mengting Xing, Zilong Fu, and Yongdong Zhang. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In *CVPR*, pages 11753–11762, 2020.

[Yang *et al.*, 2019] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP*, pages 8261–8265. IEEE, 2019.

[Yu *et al.*, 2019] Lingyun Yu, Jun Yu, and Qiang Ling. Mining audio, text and visual information for talking face generation. In *ICDM*, pages 787–795. IEEE, 2019.

[Yu *et al.*, 2020] Lingyun Yu, Jun Yu, Mengyan Li, and Qiang Ling. Multimodal inputs driven talking face generation with spatial-temporal dependency. *IEEE TCSVT*, 2020.

[Zhou *et al.*, 2017] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *CVPR Workshops*, pages 1831–1839. IEEE, 2017.