Hierarchical Contrastive Inconsistency Learning for Deepfake Video Detection

Zhihao Gu[1,2⋆], Taiping Yao[2⋆], Yang Chen[2], Shouhong Ding[2†], and Lizhuang Ma[1,3†]

[1] DMCV Lab, Shanghai Jiao Tong University, China
[2] Youtu Lab, Tencent, China
[3] MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, China
ellery-holmes@sjtu.edu.cn, ma-lz@cs.sjtu.edu.cn
{taipingyao, wizyangchen, ericshding}@tencent.com

Abstract. With the rapid development of Deepfake techniques, the capacity of generating hyper-realistic faces has aroused public concerns in recent years. The temporal inconsistency which derives from the contrast of facial movements between pristine and forged videos can serve as an efficient cue in identifying Deepfakes. However, most existing approaches tend to impose binary supervision to model it, which restricts them to only focusing on the category-level discrepancies. In this paper, we propose a novel Hierarchical Contrastive Inconsistency Learning framework (HCIL) with a two-level contrastive paradigm. Specially, sampling multiply snippets to form the input, HCIL performs contrastive learning from both local and global perspectives to capture more general and intrinsical temporal inconsistency between real and fake videos. Moreover, we also incorporate a region-adaptive module for intra-snippet inconsistency mining and an inter-snippet fusion module for cross-snippet information fusion, which further facilitates the inconsistency learning. Extensive experiments and visualizations demonstrate the effectiveness of our method against SOTA competitors on four Deepfake video datasets, *i.e.,* FaceForensics++, Celeb-DF, DFDC, and Wild-Deepfake.

Keywords: Deepfake Video Detection, Inconsistency Learning

1 Introduction

As the rapid development of deep learning [22, 19, 20, 23, 21, 44, 43, 59, 58], the resulting privacy and security concerns [24, 56, 7] have received numerous attention in recent years. Face manipulation technique, known as Deepfakes, is one of the most emerging threats. Since the generated faces in videos are too realistic to be identified by humans, they can easily be abused and trigger severe societal or political threats. Thus it is urgent to design effective detectors for Deepfakes.

Recently, the Deepfake detection technique has achieved significant progress

⋆ Equal contributions. †Corresponding authors. This work was done when Zhihao Gu was an intern at Youtu Lab.

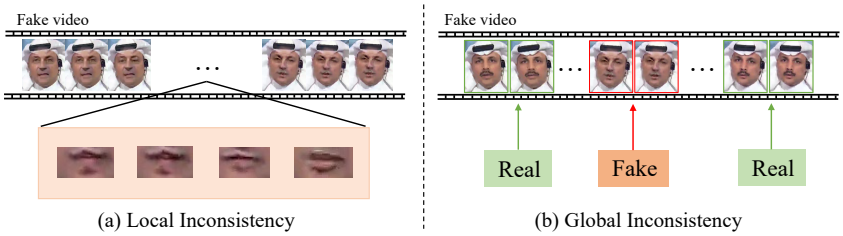

Fig. 1. Illustration of local and global inconsistency. The former one refers to the irregular facial movements within several consecutive frames. The latter one stands for the cases of partial manipulation in the video. Both of them are essential for identifying Deepfakes. Therefore, we construct the hierarchical contrastive learning from both local and global perspectives.

and developed into image-based and video-based methods. Image-based methods [52, 6, 25, 38, 5, 11, 2] aim to exploit various priors for mining discriminative frame-level features, including face blending boundaries [25], forgery signals on frequency spectrum [38, 5] and contrast between augmented pairs [52, 6]. However, the development of manipulation techniques promotes the generation of highly realistic faces, the subtle forgery cues can hardly be identified in the static images and it calls for attention to the temporal information in this area. Therefore, some researches tend to develop video-based approaches. Early works treat it as a video-level classification task and directly adopt the off-the-shelf networks like LSTM [17] and I3D [3] to deal with it, which results in inferior performance and high computational cost. Recent works [26, 14, 12] focus on designing efficient paradigms for modeling inconsistent facial movements between pristine and forged videos, known as temporal inconsistency. S-MIL [26] treats faces and videos as instances and bags for modeling the inconsistency in adjacent frames. LipForensics [14] extracts the high-level semantic irregularities in mouth movements to deal with the temporal inconsistency. STIL [12] exploits the difference over adjacent frames in two directions. Although the overall performance is improved, there are still some limitations. First of all, they heavily rely on the video-level binary supervision without exploring the more general and intrinsical inconsistency. Second, they either exploit the sparse sampling strategy, failing to model the local inconsistency contained in subtle motions or apply dense sampling over consecutive frames, ignoring the long-range inconsistency.

As illustrated in Fig. 1, the temporal inconsistency can be divided from local and global perspectives. Moreover, since it reveals the inconsistent facial movements between real and fake videos, we argue that it should be excavated through comparison, which is significantly neglected by existing works. To address this issue, we aim to introduce contrast into temporal inconsistency learning, and a hierarchical contrastive inconsistency learning framework is proposed. However, there are still several challenges: how to 1) conduct local and global contrast, and 2) extract finer local and global representations for it. To solve the former one, we sample a few snippets from each video and build the local contrastive

inconsistency learning on snippet-level representations. Snippets from pristine videos are viewed as positive samples while ones from fake videos are negative samples (no matter whether the videos are partially manipulated). Then the similarity between an anchor snippet and the one from a fake video is used as the regularizer to adaptively decide whether to repel or not, leading to the weighted NCE loss. And the global contrast is directly established on video-level representations. To solve the latter one, a region-aware inconsistency module and an inter-snippet fusion module are respectively proposed to generate discriminative intra-snippet inconsistency for local contrast and promote the interaction between snippets for the global one. Overall, the proposed framework can capture essential temporal inconsistency for identifying Deepfakes and outperforms the SOTA competitors under both full and partial manipulation settings. Besides, extensive ablations and visualizations further validate its effectiveness. In summary, our main contributions can be summarized as follows:

1. We propose a novel Hierarchical Contrastive Inconsistency Learning (HCIL) framework for Deepfake Video Detection, which performs contrastive learning from both local and global perspectives to capture more general and intrinsical temporal inconsistency between real and fake videos.
2. Considering the partial forgery videos, the weighted NCE loss is specially designed to enable the snippet-level inconsistency contrast. Besides, the region-aware inconsistency module and the inter-snippet fusion module are further proposed to facilitate the inconsistency learning.
3. Extensive experiments and analysis further illustrate the effectiveness of the proposed HCIL against its competitors on several popular benchmarks.

2 Related work

Deepfake Detection. Deepfake detection has obtained more and more attention in recent years. Early researches mainly focus on designing hand-crafted features for identification, such as face warping artifacts [28, 49], eye blinking [27] and inconsistent head poses [54]. With the development of deep learning, some image-based methods are proposed to extract discriminative frame-level features for detection. [39] evaluates several well-known 2D neural networks to detect Deepfakes. X-ray [25] identifies Deepfakes by revealing whether the input image can be decomposed into the blending of two images from different sources. F^3-Net [38] exploits the frequency information as a complementary viewpoint for forgery pattern mining. All these approaches perform well in image-level detection. However, with the development of Deepfake techniques, the forgery trace can be hardly found. Recent works [33, 26, 37, 14, 13] tend to consider the temporal inconsistency as the key to distinguishing Deepfakes and propose various methods to model it. Two-branch [33] designs a two-branch network to amplify artifacts and suppress high-level face contents. S-MIL [26] introduces a multi-instance learning framework, treating faces and videos as instances and bags, for modeling the inconsistency in adjacent frames. DeepRhythm [37] conjectures that heartbeat rhythms in fake videos are entirely broken and uses CNNs to

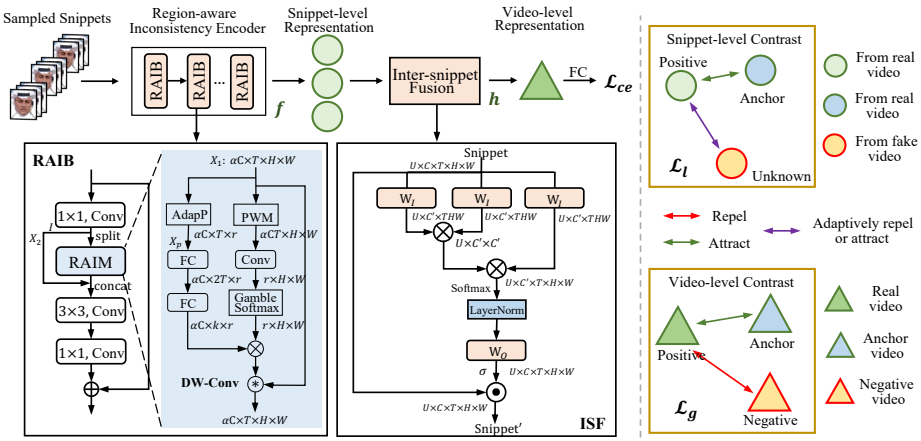

Fig. 2. Overview of the hierarchical contrastive inconsistency learning. The framework is constructed on both snippet and video-level representations. To enable the local contrast, a weighted NCE loss $\mathcal{L}_l$ is designed to adaptively decide whether snippets from fake videos need to be repelled. DW-Conv, $\otimes$, $\oplus$ and $\odot$ stand for depth-wise convolution, matrix multiplication element-wise addition and multiplication, respectively.

monitor them for detection. LipForensics [14] proposes to capture the semantic irregularities of mouth movements in generated videos for classification. Different from them capturing the temporal inconsistency via category-level discrepancy, we conduct two-level contrast to formulate the more general and intrinsical temporal inconsistency.

Contrastive Learning. The main idea behind contrastive learning is to learn visual representations via attracting similar instances while repelling dissimilar ones [52, 6, 15]. Recently, some works [1, 45, 10] attempt to introduce the contrastive learning to detect Deepfakes. DCL [42] specially designs augmentations to generate paired data and performs contrastive learning at different granularities for better generalization. SupCon [53] uses the contrast in the representation space to learn a generalizable detector. Our hierarchical contrastive framework differs from these methods in three aspects. 1) We focus on the temporal inconsistency, specially design the sampling unit called snippet, and establish the local and global contrast paradigm. 2) Not only constructing the contrastive pair is essential, so is extracting the inconsistency. Therefore, we elaborately develop the RAIM and ISF to extract region-aware local temporal inconsistency and refine the global one. 3) Considering the partial forgery in fake videos (snippet-level label unavailable), a novel weighted NCE loss is proposed to enable local contrastive learning.

Video-Analysis. Video-related tasks highly rely on the temporal modeling. Early efforts say I3D [3], exploit the 3D CNNs to capture temporal dependencies. Since they are computationally expensive, various efficient temporal modeling paradigms [30, 47, 32] are then proposed. TSM [30] shifts part of the channels

along the temporal dimension to enable information exchange among adjacent frames. TAM [32] learns video-specific temporal kernels for capturing diverse motion patterns. Although they can be directly applied to detect Deepfakes, considering no task-specific knowledge largely impacts their performance.

3 Proposed Method

In this section, we elaborate on how to generate positive and negative pairs and conduct contrastive inconsistency learning from both local and global perspectives. In Sec. 3.1, we first give the overview of the proposed framework. Then local contrastive inconsistency learning is introduced in Sec. 3.2. Finally, we describe the global contrastive inconsistency learning in Sec. 3.3.

3.1 Overview

As mentioned in Sec. 1, compared to pristine videos, the temporal inconsistency in fake videos can be captured from both local and global perspectives. Therefore, we aim to explicitly model it via simultaneously conducting local and global contrast. Given a real video $V^+ = [S_1^+, \ldots, S_U^+]$ with U sampled snippets of shape $T \times 3 \times H \times W$ from the set $\mathcal{N}^+$ of real videos (T, H and W denote its spatiotemporal dimensions), its anchor videos $V^a = [S_1^a, \ldots, S_U^a] \in \mathcal{N}^a$ are defined as other real videos and the corresponding negative ones are the fake videos $V^- = [S_1^-, \ldots, S_U^-] \in \mathcal{N}^-$, where $\mathcal{N}^a = \mathcal{N}^+ \setminus V^+$ and $\mathcal{N}^-$ represents the set of anchor and fake videos, respectively. Then for a positive snippet $S_i^+ \in V^+$, its anchor snippet, and negative snippets are defined as $S_j^a \in V^a$ and $S_k^- \in V^-$. To enable the local contrast, we mine dynamic snippet-level representations by a region-aware inconsistency encoder and optimize a novel weighted NCE loss [36] on them. It attracts real snippets and adaptively decides whether snippets from fake videos contribute to the loss via measuring their similarity with the anchor snippets. For global contrast, an inter-snippet fusion module is proposed to fuse the cross-snippet information and the InfoNCE loss is optimized based on the video-level features. The overall framework is illustrated in Fig. 2.

3.2 Local Contrastive Inconsistency Learning

The local contrastive inconsistency learning is shown in Fig. 2. The core of it are the region-aware inconsistency module for rich local inconsistency representations learning, and the weighted NCE loss for the contrast between snippets from real and fake videos.

Region-aware inconsistency module. Inspired by DRConv [4] that assigns generated spatial filters to corresponding spatial regions, we design the region-aware inconsistency module (RAIM) to mine comprehensive temporal inconsistency features based on different facial regions. As shown in Fig. 2, on the one hand, it adaptively divides the face into r regions according to the motion information by the right branch (PWM-Conv-gamble-softmax). On the other hand, r

region-agnostic temporal filters are learned via the left branch (AdapP-FC-FC). Based on these branches, each region is assigned with its unique temporal filter and the corresponding temporal inconsistency is thus captured through the convolution between each region and its corresponding temporal filter.

Formally, given the input $I \in R^{C\times T\times H\times W}$, we split it along channel dimensional into two parts with the rate α. Then one part X_1 is exploited to extract the region-aware inconsistency while keeping the other part X_2 unprocessed, which is found both effective and efficient [51]. In the left branch, $X_1 \in R^{\alpha C\times T\times H\times W}$ is first spatially pooled by an adaptive average pooling (AdaP) operation, resulting in $X_p \in R^{\alpha C\times T\times r}$. Then two full connected layers FC_1 and FC_2 further deal with the temporal dimension to produce r temporal filters with kernel size k:

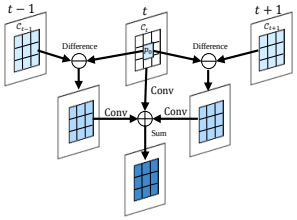

Fig. 3. Illustration of PWM.

$$[W_1, \cdots .W_r] = \mathrm{FC}_2(\mathrm{ReLU}(\mathrm{FC}_1(\mathrm{AdaP}(X_1)))), \tag{1}$$

where $W_i \in R^{\alpha C\times k}$ denotes the learned temporal kernels. We exploit the pixel-wise motion (PWM) as the guidance for adaptive face division:

$$\begin{aligned}\mathrm{PWM}(X_{p_0}) = &\sum_{p\in\mathcal{C}_{t-1}} w_p(X(p_0+p) - X(p_0)) \\ &+ \sum_{p\in\mathcal{C}_{t+1}} w_p(X(p_0+p) - X(p_0)) + \sum_{p\in\mathcal{C}_t} w_p X(p),\end{aligned} \tag{2}$$

where $\mathcal{C}_{t-1}, \mathcal{C}_t$ and $\mathcal{C}_{t+1}$ stands for the 3×3 region with center position p_0. w_p represents the weight at potion $p_0 + p$. Based on the representations, a 1×1 convolution and a gamble softmax operation are conducted on each spatial location to generate a r-dimensional one-hot vector, which is used to select their temporal filters. Positions with the identical one-hot form are viewed to belong to the same facial region. Finally, X_1 is depth-wise convoluted with the temporal filters to give the region-aware inconsistency within each snippet. The RAIM is inserted right before the second convolution in each resnet block, leading to the encoder f. And snippets go through it to form the snippet-level representations.

Local (snippet-level) Contrast. We treat snippets from two real videos as positive pairs. However, fake videos may be partially manipulated and we can't simply treat sampled snippets from them as negative ones. To alleviate this issue, we use the normalized similarity between $S_j^a \in V^a$ and $S_k^- \in V^-$ to adjust the impact of S_k^-. A weighted NCE loss is thus proposed and formulated as:

$$\mathcal{L}_{\mathrm{NCE}}^{w}(q_i, p_j, \{g_l(f(S_w^-))\}_k) = -\log\frac{e^{\phi(q_i,p_j)/\tau}}{e^{\phi(q_i,p_j)/\tau} + \sum_k \left(\frac{1-\phi(p_j,n_k)}{2}\right)^{\beta}\cdot e^{\phi_(q_i,n_k)/\tau}}, \tag{3}$$

where $g_l(\cdot) : R^C \rightarrow R^{128}$ is a projection head, $q_i = g_l(f(S_i^+))$, $p_j = g_l(f(S_j^a))$ and $n_k = g_l(f(S_w^-))_k$. $\phi(x, y)$ denotes the cosine similarity between two l_2-normalized vectors. τ refers to the temperature scalar and β is a tunable factor. The term $(\cdot)^\beta$ dynamically decides whether the snippet from fake videos contributes to the contrastive loss based on its similarity with the anchor. That is, if the snippet is pristine/forged, the term approximates 0/1 and thus it suppresses/activates the contrast with real snippets. Then the local contrastive loss is given by:

$$\mathcal{L}_l = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N-U} \mathcal{L}_{\text{NCE}}^w(g_l(f(S_i^+)), g_l(f(S_j^a)), \{g_l(f(S_w^-))\}_k), \tag{4}$$

where $N = |\mathcal{N}^+| \cdot U$, and $|\mathcal{N}^+|$ denotes the number of videos in $\mathcal{N}^+$. Note that q_i has multiple anchor sample p_j, we average their separate contrastive loss.

Analysis of the weighted NCE Loss. We now analyze the relationship between the proposed loss and the InfoNCE loss [36]. First of all, the NCE loss can be viewed as a special case ($\beta = 0$) of the proposed weighted NCE loss. Furthermore, we derive the relation between their derivatives:

$$\frac{\partial \mathcal{L}_{\text{NCE}}^w}{\partial w} = \frac{\partial (\frac{b^w}{b})'}{\partial w} \cdot (e^{\mathcal{L}_{\text{NCE}}} - 1) e^{-\mathcal{L}_{NCE}^w} + \frac{b^w}{b} \cdot \frac{e^{\mathcal{L}_{\text{NCE}}}}{e^{\mathcal{L}_{\text{NCE}}^w}} \cdot \frac{\partial \mathcal{L}_{\text{NCE}}}{\partial w}. \tag{5}$$

where $b = \sum_k e^{\phi(q,n_k)/\tau}$ and $b^w = \sum_k (\frac{1-\phi(p,n_k)}{2})^2 e^{\phi(q,n_k)/\tau}$. If all the n_k from fake videos are forged, we then have $\frac{\partial \mathcal{L}_{NCE}^w}{\partial w} \rightarrow \frac{\partial \mathcal{L}_{NCE}}{\partial w}$. On the contrary, assume there exists a snippet $n_j \in \mathcal{N}^-$ is in-manipulated, then the learning process is less affected by n_j. Surprisingly, our solution performs on part with or even better than the InfoNCE loss as shown in Table. 6.

3.3 Global Contrastive Inconsistency Learning

The global contrastive inconsistency learning is illustrated in Fig. 2. The main components are the inter-snippet fusion module (ISF) for forming the video-level representation and the NCE loss for the contrast between real and fake videos.

Inter-snippet Fusion. A common way to generate video-level representation is averaging snippet-level features along the U dimension. Inspired by [50] using the modified non-local [48] to enhance the short-term features, we instead propose to enhance the channels that reveal the intrinsical inconsistency in a similar way. To achieve this, we design an inter-snippet fusion module upon the encoder f to promote information fusion between $f(S_i) \in R^{U \times C' \times T \times H' \times W'}$ in $f(V)$, where $V = [S_1, \ldots, S_U]$. Specially, the cross-snippet interaction is defined as the self-attention operation between snippets:

$$\text{Atten} = \text{softmax}(\frac{(f(V)W_I)(W_I^\top f(V)^\top)}{\sqrt{C'}}) f(V) W_I, \tag{6}$$

where W_I is learnable parameter of projection for dimension reduction. Then Atten is used to re-weight channels of $f(V)$ by:

$$\text{ISF(f(V))} = \sigma(\text{Norm}(\text{Atten}) W_O) \odot f(V), \tag{7}$$

where W_O is the learnable parameters of projection for dimension retrieval. Norm$(\cdot)$ is the layer-norm and σ refers to the sigmoid function.

Global (video-level) Contrast. We build the video-level contrast on the outputs of the inter-snippet fusion module. Different from the snippet-level contrast, labels of videos are provided and the InfoNCE loss [36] can be directly exploited:

$$\mathcal{L}_{\text{NCE}}(u_i, v_j, \{g_g(f(V_w^-))\}_k) = -\log \frac{e^{\phi(u_i, v_j)/\tau}}{e^{\phi(u_i, v_j)/\tau} + \sum_k e^{\phi(u_i, m_k)/\tau}}, \tag{8}$$

where $g_g(\cdot) : R^C \rightarrow R^{128}$ is the projection head, $u_i = g_g(f(V_i^+))$, $v_j = g_g(f(V_j^a))$, $n_k \in \mathcal{N}^-$ and $m_k = g_g(f(V_w^-))_k$. The video-level contrast can be written as:

$$\mathcal{L}_g = \frac{1}{|\mathcal{N}^+|} \sum_{i=1}^{|\mathcal{N}^+|} \mathcal{L}_{\text{NCE}}(g_g(h(f(V_i^+))), g_g(h(f(\mathcal{N}^a))), g_g(h(f(\mathcal{N}^-)))), \tag{9}$$

where $h(f(V_i^+)) = h(f(S_1^+), \ldots, f(S_U^+))$, $h(f(\mathcal{N}^a)) = \{h(f(V_i^a))\}_j$ and $h(f(\mathcal{N}^-) = \{h(f(V_w^-))\}_k$. Note that one video u_i usually contains multiple anchor videos, we simply compute the contrastive loss separately and take their average.

3.4 Loss Function

Apart from using the contrastive loss mentioned above, we also adopt the binary cross-entropy loss $\mathcal{L}_{ce}$ to supervise the category-level discrepancies. The final loss function for training is formulated as the weighted sum of them:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_1 \mathcal{L}_l + \lambda_2 \mathcal{L}_g, \tag{10}$$

where λ_1 and λ_2 are two balance factors for balancing different terms. All the projection heads for contrastive learning are discarded during inference.

4 Experiments

4.1 Datasets

We conduct all the experiments based on four popular datasets, *i.e.,* FaceForensics++ (FF++) [39], Celeb-DF [29], DFDC-Preview [9] and WildDeepfake [60].

FaceForensics++ is comprised of 1000 original and 4000 forged videos with several visual quality, *i.e.,* high quality (HQ) and low quality (LQ). Four manipulation methods are exploited for forgery, that is DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). It provides only video-level labels and nearly every frame is manipulated.

Celeb-DF contains 590 real and 5639 fake videos and all the videos are post-processed for better visual quality. Only video-level labels are provided in it and nearly every frame is manipulated.

WildDeepfake owns 7314 face sequences of different duration from 707 Deepfake videos. Video-level labels are provided only and whether videos in it are

partially forged is not mentioned. Therefore, it is relatively more difficult.
DFDC-Preview is the preview version of the DFDC dataset and consists of around 5000 videos. These videos are *partially manipulated* videos by two unknown manipulations, making it more challenging. And no frame-level labels are provided.

4.2 Experimental Settings

Implemental Details. For each Deepfake video dataset, we perform face detection with a similar strategy in [26]. λ_1, λ_2, β and τ are empirically set as 0.1, 0.01, 2 and 5. For simplification, we select the split ratio $\alpha = 0.5$ and the region number $r = 8$. The squeeze ratio is $\frac{1}{8}$ by default. ResNet-50 [16] pre-trained on the ImageNet [8] is exploited as our backbone network. During training, we equally divide a video into $U = 4$ segments and randomly sample consecutive $T = 4$ frames within them to form a snippet. Each frame is then resized into 224×224 as input. We use Adam as the optimizer and set the batch size to 12. The network is trained for 60 epochs and the initial learning rate is 10^{-4}, which is divided by 10 for every 20 epochs. Only random horizontal flip is employed as data augmentation. During testing, we centrally sample $U = 8$ snippets with $T = 4$ and also resize them into the shape of 224×224. The projection heads $g_l(\cdot)$ and $g_g(\cdot)$ are implemented as two MLP layers, transforming features to a 128-dimension space for computing similarity.
Baselines. In order to demonstrate the superiority of our HCIL, we select several video-based detectors for comparison, including Xception [39], VA-LogReg [34], D-FWA [28], FaceNetLSTM [41], Capsule [35], Co-motion [46], DoubleRNN [33], S-MIL [26], DeepRhythm [37], ADDNet-3d [60], STIL [12] , DIANet [18] and TD-3DCNN [55]. Besides, for more comprehensive validation, we also re-implement some representative works in video analysis to detect Deepfakes, *i.e.,* LSTM [17], I3D [3], TEI [31], TDN [47], V4D [57], TAM [32] and DSANet [51].

4.3 State-of-the-art Comparisons

Following [37, 12], we conduct both intra-dataset evaluation and cross-dataset generalization to demonstrate the effectiveness of the proposed framework. The accuracy and the Area Under Curve (AUC) metrics are reported, respectively.
Comparison on FF++ dataset. We first conduct the comprehensive experiments on four subsets of the FF++ dataset under both HQ and LQ settings. Table 1 illustrates the corresponding results, from which we have several observations. Firstly, since no temporal information is considered, the frame-based detectors achieve inferior accuracy compared to video analysis methods. Besides, V4D and DSANet can also achieve a comparable result to the SOTA method STIL. However, they usually employ the sparse sampling strategy and no local motion information is involved, which is also important for this task. Secondly, The proposed HCIL outperforms nearly all the competitors in all settings. Specifically, in challenging LQ NT setting, HCIL owns 94.64% accuracy, exceeding the best action recognition model V4D and the SOTA Deepfake video detector

Table 1. Comparisons on FF++ dataset under both HQ and LQ settings. All subsets are measured and accuracy is reported. † indicates re-implementation.

Methods	FaceForensics++ HQ				FaceForensics++ LQ			
	DF	F2F	FS	NT	DF	F2F	FS	NT
ResNet-50†	0.9893	0.9857	0.9964	0.9500	0.9536	0.8893	0.9464	0.8750
Xception	0.9893	0.9893	0.9964	0.9500	0.9678	0.9107	0.9464	0.8714
LSTM	0.9964	0.9929	0.9821	0.9393	0.9643	0.8821	0.9429	0.8821
I3D†	0.9286	0.9286	0.9643	0.9036	0.9107	0.8643	0.9143	0.7857
TEI†	0.9786	0.9714	0.9750	0.9429	0.9500	0.9107	0.9464	0.9036
TAM†	0.9929	0.9857	0.9964	0.9536	0.9714	0.9214	0.9571	0.9286
DSANet†	0.9929	0.9929	0.9964	0.9571	0.9679	0.9321	0.9536	0.9178
V4D†	0.9964	0.9929	0.9964	0.9607	0.9786	0.9357	0.9536	0.9250
TDN†	0.9821	0.9714	0.9857	0.9464	0.9571	0.9178	0.9500	0.9107
FaceNetLSTM	0.8900	0.8700	0.9000	-	-	-	-	-
Co-motion-70	0.9910	0.9325	0.9830	0.9045	-	-	-	-
DeepRhythm	0.9870	0.9890	0.9780	-	-	-	-	-
ADDNet-3d†	0.9214	0.8393	0.9250	0.7821	0.9036	0.7821	0.8000	0.6929
S-MIL	0.9857	0.9929	0.9929	0.9571	0.9679	0.9143	0.9464	0.8857
S-MIL-T	0.9964	**0.9964**	1.0	0.9429	0.9714	0.9107	0.9607	0.8679
STIL	0.9964	0.9928	1.0	0.9536	0.9821	0.9214	0.9714	0.9178
HCIL	**1.0**	0.9928	**1.0**	**0.9676**	**0.9928**	**0.9571**	**0.9750**	**0.9464**

STIL by 2.14% and 2.86%, respectively. All these improvements validate that the well-designed contrastive framework prompts learning general while intrinsical inconsistency between pristine and forged videos.

Comparison on other datasets. We also evaluate the proposed method on Celeb-DF, WildDeepfake, and DFDC datasets, as listed in Table 2. On Celeb-DF and WildDeepfake datasets, our framework consistently performs better than SOTAs (0.2% ↑ on Celeb-DF and 1.24% ↑ on Wild-DF). This is mainly because we extract the rich local inconsistency features and generate the global ones, leading to more comprehensive representations. The overall performance on Wild-DF is still low since the duration of videos varies a lot, making them difficult to deal with. On the more challenging DFDC dataset containing vast partially forged videos, HCIL still outperforms the compared works. A large performance margin can be observed (up to 6.0%). Several reasons may account for this. First, different from compared works exploiting either sparse or dense sampling strategy, we sample several snippets from each video to form the input. This strategy enables to capture partially forged parts and both local and global temporal information are covered. Second, the constructed local contrast allows the network to perform fine-grained learning. Therefore, the intrinsical inconsistencies are obtained from contrast, not only from the category-level differences.

Table 2. Comparison on Celeb-DF, DFDC, and WildDeepfake datasets. Accuracy is reported.

Methods	Celeb-DF	Wild-DF	DFDC
Xception	0.9944	0.8325	0.8458
I3D†	0.9923	0.6269	0.8082
TEI†	0.9912	0.8164	0.8697
TAM†	0.9923	0.8251	0.8932
V4D†	0.9942	0.8375	0.8739
DSANet†	0.9942	0.8474	0.8867
D-FWA	0.9858	-	0.8511
DIANet	-	-	0.8583
ADDNet-3D†	0.9516	0.6550	0.7966
S-MIL	0.9923	-	0.8378
S-MIL-T	0.9884	-	0.8511
STIL†	0.9961	0.8462	0.8980
HCIL	**0.9981**	**0.8586**	**0.9511**

Table 3. Cross-dataset generalization in terms of AUC. † implies re-implementation.

Methods	FF++ DF	Celeb-DF	DFDC
Xception	0.9550	0.6550	0.5939
I3D†	0.9541	0.7411	0.6687
TEI†	0.9654	0.7466	0.6742
TAM†	0.9704	0.6796	0.6714
V4D†	0.9674	0.7008	0.6734
DSANet†	0.9688	0.7371	0.6808
Capsule	0.9660	0.5750	-
DIANet	0.9040	0.7040	-
TD-3DCNN	-	0.5732	0.5502
DoubleRNN	0.9318	0.7341	-
ADDNet-3D†	0.9622	0.6085	0.6589
STIL†	0.9712	0.7558	0.6788
HCIL	**0.9832**	**0.7900**	**0.6921**

Cross-dataset generalization. Following previous work [37], we first train the network on the FF++ dataset to discern the pristine videos against four manipulation types under the LQ setting. Then evaluating on FF++ DF, Celeb-DF, and DFDC datasets to measure its generalization capacity, as studied in Table 3. We achieve 98.32% AUC on FF++ DF and 69.21% on the DFDC dataset, improving the state-of-the-art competitors by about 1% on average. Larger performance gains of 3.42% are obtained on Celeb-DF. Since FF++ DF and Celeb-DF datasets are manipulated through similar face forgery techniques frame-by-frame, the detector presents relatively better generalization compared to the DFDC dataset. Note that videos in DFDC are partially forged and contain varied lighting conditions, head poses, and background. Therefore, it is harder to generalize. However, benefiting from the mechanism that learning representations from not only label-level discrepancies but also the local and global contrast, the network owns robustness to a certain degree and consequently exceeds the previous SOTA STIL by 1.3%.

4.4 Ablation Study

We conduct comprehensive ablation studies to further explore the effectiveness of the proposed modules and contrastive framework from Table 4 to Table 6.

Study on key components. The core components of HCIL include the RAI module and ISF module and the specially designed hierarchical contrastive learning paradigm. We perform the ablation under both the intra and inter-dataset settings. And the corresponding results are shown in Table 4 and 5, respectively. In the FF++ dataset, only extracting the frame-level features, the baseline model has the lowest accuracy and AUC. Surprisingly, based on these frame-level representations, directly constructing the contrastive inconsistency learning from multi-hierarchy improves the performance a lot (88.93% $\rightarrow$ 91.43% on F2F and 87.50% $\rightarrow$ 90.36% on NT). Besides equipping the baseline with

Table 4. Ablation study on key components under intra-dataset evaluation. ResNet-50 is used as baseline.

(a) RAI and CSI on FF++.

model	DF	F2F	FS	NT
Baseline	0.9536	0.8893	0.9464	0.8750
+ RAI	0.9821	0.9321	0.9607	0.9214
+ ISF	0.9785	0.9286	0.9607	0.9178
+ RAI + ISF	0.9857	0.9428	0.9643	0.9250
+ All	**0.9928**	**0.9571**	**0.9750**	**0.9464**

(b) Contrastive framework on FF++.

model	DF	F2F	FS	NT
Baseline	0.9536	0.8893	0.9464	0.8750
+ $\mathcal{L}_l$ + $\mathcal{L}_g$	0.9750	0.9143	0.9571	0.9036
+ RAI + $\mathcal{L}_l$	0.9857	0.9393	0.9678	0.9250
+ ISF + $\mathcal{L}_g$	0.9857	0.9325	0.9642	0.9214
+ All	**0.9928**	**0.9571**	**0.9750**	**0.9464**

Table 5. Ablation study on key components under cross-dataset generalization. ResNet-50 is used as baseline.

(c) Generalization for RAI and ISF.

model	FF++ DF	Celeb-DF	DFDC
Baseline	0.9232	0.6956	0.6323
+ RAI	0.9673	0.7478	0.6575
+ ISF	0.9603	0.7392	0.6601
+ RAI + ISF	0.9743	0.7578	0.6665
+ All	**0.9832**	**0.7900**	**0.6921**

(d) Generalization for contrastive framework.

model	FF++ DF	Celeb-DF	DFDC
Baseline	0.9232	0.6956	0.6323
+ $\mathcal{L}_l$ + $\mathcal{L}_g$	0.9572	0.7323	0.6518
+ RAI + $\mathcal{L}_l$	0.9764	0.7569	0.6898
+ ISF + $\mathcal{L}_g$	0.9711	0.7668	0.6792
+ All	**0.9832**	**0.7900**	**0.6921**

RAI to extract diverse region-related inconsistency within snippets, large performance gains are observed (95.36% → 98.21% on DF, 88.93% → 93.21% on F2F, 94.64% → 96.06% on FS and 87.50% → 92.14% on NT). This is reasonable that without mining the temporal relation within snippets, the general inconsistency can not be captured well via contrast. Similarly, the ISF module also improves the baseline but achieves an inferior result to the RAI module, which implies the importance of local temporal information. If constructing contrastive framework with the corresponding representations, *i.e.,* $\mathcal{L}_s$+ RAI and $\mathcal{L}_v$+ CSF, the accuracy are further boosted (0.36% ↑ on DF, 0.72% ↑ on F2F, 0.71% ↑ on FS and 0.36% ↑ on NT). Combining RAI and ISF modules is better than contrasting based on one of them, indicating both local and global temporal features are essential for the task. No doubt, combining them all obtains the best results. Similar conclusions can be observed in cross-dataset generalization settings.

Study on parameter β. Eq. 3 plays an important role in our contrastive inconsistency framework. β in it indeed adaptively adjusts the importance extent of each negative pair based on the similarity between anchor and snippets from fake videos. Table 6 studies the impacts of it under intra-dataset evaluation settings. In full manipulation settings, *i.e.,* FF++, Celeb-DF and Wild-DF datasets, compared to the baseline, constructing the snippet-level contrast boosts the accuracy (0.35% ↑ on DF, 1.07% ↑ on F2F, 0.72% ↑ on FS and 1.39% ↑ on NT). And NCE loss with adaptive weights presents on part with or even slightly better performance than the vanilla NCE loss. This implies that adjusting the weights in full manipulation settings is helpful. However, the gains from it are limited (0.% ↑ on DF, 0.35% ↑ on F2F, 0.36% ↑ on FS and 0.16% ↑ on NT). In the DFDC dataset, using the NCE loss leads to slight accuracy gains (0.64%). On

Table 6. Ablation study on β under intra-dataset evaluation settings. The proposed detector without $\mathcal{L}_s$ is set as baseline.

(e) FF++ dataset.

β	DF	F2F	FS	NT
Baseline	0.9893	0.9464	0.9678	0.9321
$\mathcal{L}_{\text{NCE}}$	**0.9928**	0.9536	0.9750	0.9428
$\mathcal{L}^{w}_{\text{NCE}}$	**0.9928**	**0.9571**	**0.9750**	**0.9464**

(f) Other three datasets.

β	Celeb-DF	Wild-DF	DFDC
Baseline	0.9942	0.8472	0.9331
$\mathcal{L}_{\text{NCE}}$	**0.9981**	0.8573	0.9395
$\mathcal{L}^{w}_{\text{NCE}}$	**0.9981**	**0.8586**	**0.9511**

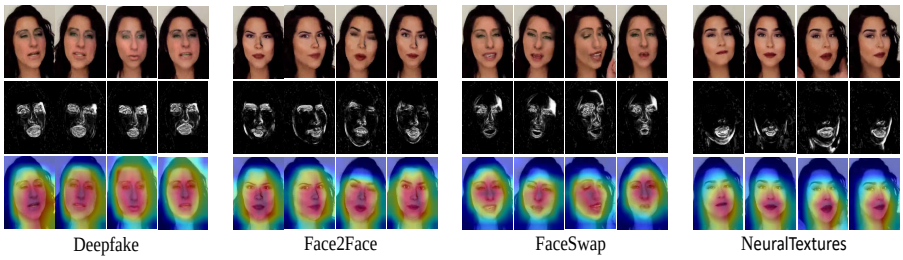

Fig. 4. Visualization of activation maps with CAM. First row: RGB images, Second row: forgery masks, Third row: activation maps.

the contrary, our weighted NCE loss surprisingly gives a larger improvement (from 0.9331% to 0.9511%).

4.5 Visualization Analysis

In this section, we visualize the region-of-interest via Grad-CAM [40], as shown in Fig. 4. Similar analyses of the RAI module and the weighted NCE loss are presented in Fig. 5.

Class activation maps. In Fig. 4, we visualize the class activation maps against four manipulations to verify which regions the model focuses on. The forgery masks derive from the difference between the manipulated videos and the correspondingly pristine videos. The activation maps almost cover the whole faces for Deepfake that are generated from deep learning tools. Similarly, the detector also notices the swapped facial region for FaceSwap. On more challenging Face2Face and NeuralTextures whose transferred expression and mouth regions are difficult to identify, our model still successfully locates the forged areas.

Impacts of RAI module. The RAI module aims to adaptively extract dynamic local representations for snippet-level contrastive inconsistency learning. We compare it with the vanilla temporal convolution and visualize the corresponding cam maps in the third and second row of Fig. 5 (a). With the content-agnostic weights, the vanilla temporal convolution treats all the facial regions equally and is easy to focus on incomplete forgery regions (second row of left part in Fig. 5 (a)) or wrong areas (second row of right part in Fig. 5 (a)). On the contrary, the RAI module generates the region-specific temporal kernels to

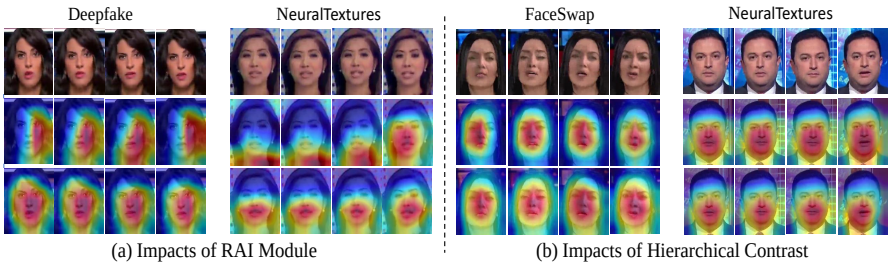

Fig. 5. Visualization on impacts of (a) snippet-level feature extraction (second row: without RAIM, third row: with RAIM) and (b) hierarchical contrast (second row: without contrast, third row: with contrast).

extract dynamic temporal information and, therefore, more complete temporal inconsistency representations can be captured for identification.

Impacts of hierarchical contrast. The hierarchical contrast attracts positive samples while repelling negative ones from both snippet and video levels. To intuitively illustrate impacts of it, we directly visualize the activation in Fig. 5 (b). From the figure, we can observe that for the manipulation performed on small facial areas, the hierarchical contrastive loss acts as a regularizer, regularizing the attention on more accurate regions. Besides, for the manipulation performed on large facial regions, the loss instead guides the detector to focus on more comprehensive locations.

5 Conclusions

In this paper, we introduce the hierarchical contrastive inconsistency learning framework for Deepfake video detection from local and global perspectives. For local contrast, we design a region-aware inconsistency module for dynamic snippet-level representations and a novel weighted NCE loss to enable the snippet-level contrast. For global contrast, an inter-snippet fusion module is introduced for fusing cross-snippet information. The proposed framework presents superior performance and generalization on several benchmarks, and extensive visualizations also illustrate its effectiveness.

Acknowledgements

This research is supported in part by the National Key Research and Development Program of China (No. 2019YFC1521104), National Natural Science Foundation of China (No. 61972157 and No. 72192821), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Science and Technology Commission (21511101200 and 21511101200) and Art major project of National Social Science Fund (I8ZD22). We also thank Shen Chen for the proof-read of our manuscript.

References

1. Beuve, N., Hamidouche, W., Deforges, O.: Dmyt: Dummy triplet loss for deepfake detection. In: WSMMADGD (2021)
2. Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., Yang, X.: End-to-end reconstruction-classification learning for face forgery detection. In: CVPR (2022)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
4. Chen, J., Wang, X., Guo, Z., Zhang, X., Sun, J.: Dynamic region-aware convolution. In: CVPR (2021)
5. Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., Ji, R.: Local relation learning for face forgery detection. In: AAAI (2021)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
7. Chen, Z., Li, B., Xu, J., Wu, S., Ding, S., Zhang, W.: Towards practical certifiable patch defense with vision transformer. In: CVPR (2022)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
9. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C.: The deepfake detection challenge (dfdc) preview dataset. In: arXiv (2019)
10. Fung, S., Lu, X., Zhang, C., Li, C.T.: Deepfakeucl: Deepfake detection via unsupervised contrastive learning. In: IJCNN (2021)
11. Gu, Q., Chen, S., Yao, T., Chen, Y., Ding, S., Yi, R.: Exploiting fine-grained face forgery clues via progressive enhancement learning. In: AAAI (2021)
12. Gu, Z., Chen, Y., Yao, T., Ding, S., Li, J., Huang, F., Ma, L.: Spatiotemporal inconsistency learning for deepfake video detection. In: ACM MM (2021)
13. Gu, Z., Chen, Y., Yao, T., Ding, S., Li, J., Ma, L.: Delving into the local: Dynamic inconsistency learning for deepfake video detection. In: AAAI (2022)
14. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: CVPR (2021)
15. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. In: NC (1997)
18. Hu, Z., Xie, H., Wang, Y., Li, J., Wang, Z., Zhang, Y.: Dynamic inconsistency-aware deepfake video detection. In: IJCAI (2021)
19. Li, B., Sun, Z., Guo, Y.: Supervae: Superpixelwise variational autoencoder for salient object detection. In: AAAI (2019)
20. Li, B., Sun, Z., Li, Q., Wu, Y., Hu, A.: Group-wise deep object co-segmentation with co-attention recurrent neural network. In: ICCV (2019)
21. Li, B., Sun, Z., Tang, L., Hu, A.: Two-b-real net: Two-branch network for real-time salient object detection. In: ICASSP (2019)
22. Li, B., Sun, Z., Tang, L., Sun, Y., Shi, J.: Detecting robust co-saliency with recurrent co-attention neural network. In: IJCAI (2019)
23. Li, B., Sun, Z., Wang, Q., Li, Q.: Co-saliency detection based on hierarchical consistency. In: ACM MM (2019)
24. Li, B., Xu, J., Wu, S., Ding, S., Li, J., Huang, F.: Detecting adversarial patch attacks through global-local consistency. In: CoRR (2021)

25. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: CVPR (2020)
26. Li, X., Lang, Y., Chen, Y., Mao, X., He, Y., Wang, S., Xue, H., Lu, Q.: Sharp multiple instance learning for deepfake video detection. In: ACM MM (2020)
27. Li, Y., Chang, M.C., Lyu, S.: In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. In: arXiv (2018)
28. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. In: arXiv (2018)
29. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: CVPR (2020)
30. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: ICCV (2019)
31. Liu, Z., Luo, D., Wang, Y., Wang, L., Tai, Y., Wang, C., Li, J., Huang, F., Lu, T.: Teinet: Towards an efficient architecture for video recognition. In: AAAI (2020)
32. Liu, Z., Wang, L., Wu, W., Qian, C., Lu, T.: Tam: Temporal adaptive module for video recognition. In: CVPR (2021)
33. Masi, I., Killekar, A., Mascarenhas, R.M., Gurudatt, S.P., AbdAlmageed, W.: Two-branch recurrent network for isolating deepfakes in videos. In: ECCV (2020)
34. Matern, F., Riess, C., Stamminger, M.: Exploiting visual artifacts to expose deepfakes and face manipulations. In: CVPRW (2019)
35. Nguyen, H.H., Yamagishi, J., Echizen, I.: Capsule-forensics: Using capsule networks to detect forged images and videos. In: ICASSP (2019)
36. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. In: arXiv (2018)
37. Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., Liu, Y., Zhao, J.: Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In: ACM MM (2020)
38. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: ECCV (2020)
39. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: ICCV (2019)
40. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
41. Sohrawardi, S.J., Chintha, A., Thai, B., Seng, S., Hickerson, A., Ptucha, R., Wright, M.: Poster: Towards robust open-world detection of deepfakes. In: ACM CCCS (2019)
42. Sun, K., Yao, T., Chen, S., Ding, S., Ji, R., et al.: Dual contrastive learning for general face forgery detection. In: AAAI (2021)
43. Tang, L., Li, B.: CLASS: cross-level attention and supervision for salient objects detection. In: Ishikawa, H., Liu, C., Pajdla, T., Shi, J. (eds.) ACCV (2020)
44. Tang, L., Li, B., Zhong, Y., Ding, S., Song, M.: Disentangled high quality salient object detection. In: ICCV (2021)
45. Wang, G., Jiang, Q., Jin, X., Li, W., Cui, X.: Mc-lcr: Multi-modal contrastive classification by locally correlated representations for effective face forgery detection. In: arXiv (2021)
46. Wang, G., Zhou, J., Wu, Y.: Exposing deep-faked videos by anomalous co-motion pattern detection. In: arXiv (2020)
47. Wang, L., Tong, Z., Ji, B., Wu, G.: Tdn: Temporal difference networks for efficient action recognition. In: CVPR (2021)

48. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
49. Wang, X., Yao, T., Ding, S., Ma, L.: Face manipulation detection via auxiliary supervision. In: ICONIP (2020)
50. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: CVPR (2019)
51. Wu, W., Zhao, Y., Xu, Y., Tan, X., He, D., Zou, Z., Ye, J., Li, Y., Yao, M., Dong, Z., et al.: Dsanet: Dynamic segment aggregation network for video-level representation learning. In: ACM MM (2021)
52. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018)
53. Xu, Y., Raja, K., Pedersen, M.: Supervised contrastive learning for generalizable and explainable deepfakes detection. In: WCACV (2022)
54. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP (2019)
55. Zhang, D., Li, C., Lin, F., Zeng, D., Ge, S.: Detecting deepfake videos with temporal dropout 3dcnn. In: AAAI (2021)
56. Zhang, J., Li, B., Xu, J., Wu, S., Ding, S., Zhang, L., Wu, C.: Towards efficient data free black-box adversarial attack. In: CVPR (2022)
57. Zhang, S., Guo, S., Huang, W., Scott, M.R., Wang, L.: V4d: 4d convolutional neural networks for video-level representation learning. In: arXiv (2020)
58. Zhong, Y., Li, B., Tang, L., Kuang, S., Wu, S., Ding, S.: Detecting camouflaged object in frequency domain. In: CVPR (2022)
59. Zhong, Y., Li, B., Tang, L., Tang, H., Ding, S.: Highly efficient natural image matting. CoRR (2021)
60. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: Wilddeepfake: A challenging real-world dataset for deepfake detection. In: ACM MM (2020)