Masked Relation Learning for DeepFake Detection

Ziming Yang, Jian Liang, *Member, IEEE*, Yuting Xu, Xiao-Yu Zhang, *Senior Member, IEEE*, and Ran He, *Senior Member, IEEE*

***Abstract*— DeepFake detection aims to differentiate falsified faces from real ones. Most approaches formulate it as a binary classification problem by solely mining the local artifacts and inconsistencies of face forgery, which neglect the relation across local regions. Although several recent works explore local relation learning for DeepFake detection, they overlook the propagation of relational information and lead to limited performance gains. To address these issues, this paper provides a new perspective by formulating DeepFake detection as a graph classification problem, in which each facial region corresponds to a vertex. But relational information with large redundancy hinders the expressiveness of graphs. Inspired by the success of masked modeling, we propose Masked Relation Learning which decreases the redundancy to learn informative relational features. Specifically, a spatiotemporal attention module is exploited to learn the attention features of multiple facial regions. A relation learning module masks partial correlations between regions to reduce redundancy and then propagates the relational information across regions to capture the irregularity from a global view of the graph. We empirically discover that a moderate masking rate (*e.g.*, 50%) brings the best performance gain. Experiments verify the effectiveness of Masked Relation Learning and demonstrate that our approach outperforms the state of the art by 2% AUC on the cross-dataset DeepFake video detection. Code will be available at https://github.com/zimyang/MaskRelation.**

***Index Terms*— Multimedia forensics, DeepFake detection, masked learning, relation feature.**

I. Introduction

THE tremendous development in deep generative models [1], [2] have spawned DeepFake techniques to counterfeit faces of images or videos [3], [4], [5], [6]. Large amounts of face forgeries are created by DeepFake techniques for maliciously spreading political rumors and tampered news [7], [8], [9]. It leads to severe security crises and arouses public concern. To diminish these risks, many efforts have been devoted to developing effective methods for DeepFake detection (*a.k.a.* face forgery detection.) [10], [11], [12], [13].

DeepFake detection is typically formulated as a binary classification problem to differentiate between real faces and fake ones. Many DeepFake detectors use Deep Convolution Neural Networks (CNNs) [14], [15], [16], [17] to learn characteristics of face manipulations from large-scale face forgery datasets. But these detectors heavily rely on the type of manipulation methods and the distribution of training data, which results in inferior generalization performance on new facial manipulations and DeepFake data distributions. To discover the intrinsic patterns of face forgery, recent approaches [18], [19], [20] further explore visual artifacts in spatial and frequency domains. The advent of the attention mechanism [21], [22] allows [23], [24], [25], [26], [27], [28] to extract subtle artifacts and inconsistency among local patches to improve the accuracy of DeepFake detection. Moreover, [29] observes that DeepFake techniques individually manipulate every frame in the video and lack temporal coherence, which inspires the research on video-based DeepFake detection.

Typically, video-based DeepFake detectors mainly focus on temporal incoherence between adjacent frames [11], [12], [30] and spatiotemporal inconsistency [31], [32], [33]. They decrease the dependence on spatial artifacts and enhance generalization abilities to unseen forgeries. Previous methods usually pay attention to finding the forged local regions, but few of them consider the interrelationship between local regions. To this end, a series of works [34], [35], [36], [37] attempt to learn local relations between regions for exposing face forgery. The relational features serve as generalized patterns to further improve the generalization abilities of DeepFake detectors. However, it is desirable for deeper research on relation learning for DeepFake detection. In Figure 1, we show the relations between facial regions. A large proportion of low relations are redundant information. Recent advances indicate that masked modeling [38], [39] can learn general representations by decreasing the redundancy of information. It encourages neural networks to reconstruct masked words or image patches through self-supervised learning. Similarly, masked graph modeling [40], [41], [42], [43] guides graph neural networks (GNNs) [44] to reconstruct the masked vertices and edges of the graph. It effectively promotes the expressiveness of GNNs for graph classification and node classification.

Motivated by the above discussion, a research question naturally raises, *can masked modeling boost the relation learning for DeepFake detection?* In this paper, we explore masked relation learning for DeepFake video detection. Specifically, we propose a new framework called Masked Relation Learning

Manuscript received 31 July 2022; revised 6 January 2023; accepted 13 February 2023. Date of publication 27 February 2023; date of current version 7 March 2023. This work was supported in part by the National Natural Science Foundation of China under Grant U2003111, Grant U21B2045, and Grant 62276256; in part by the Beijing Nova Program under Grant Z211100002121108; and in part by the Chinese Association for Artificial Intelligence (CAAI)-Huawei MindSpore Open Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. William R. Schwartz. *(Corresponding author: Xiao-Yu Zhang.)*

Ziming Yang and Xiao-Yu Zhang are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: yangziming@iie.ac.cn; zhangxiaoyu@iie.ac.cn).

Jian Liang, Yuting Xu, and Ran He are with the Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: liangjian92@gmail.com; yuting.xu@cripac.ia.ac.cn; rhe@nlpr.ia.ac.cn).

Digital Object Identifier 10.1109/TIFS.2023.3249566

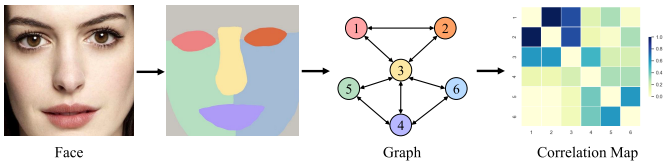

Fig. 1. **Relations between facial regions.** In the graph, vertices represent features of facial regions and edges represent correlations between regions. The correlation map illustrates the values of edges. The darker color indicates the larger value.

for video-based DeepFake detection, which mainly comprises two components: the SpatioTemporal Attention (STA) module and the Masked Relation Learner (MRL) module. STA produces attention maps to extract features from various facial regions, while MRL learns the irregularity in the relation between facial regions from deepfakes. In practice, we construct a graph in which vertices denote features of facial regions and the edge represents feature correlation between two regions. In the training procedure, MRL masks partial edges to reduce the redundancy of relations and learn the critical structure of faces. Note that MRL only masks edges instead of reconstructing the masked edges [43]. First, we consider a random masking strategy by cutting off the edges randomly. Then, we develop an importance-driven strategy that cuts off the edges corresponding to their weights. Interestingly, we find that masking 50% of the *minimal* edges achieves the best performance on DeepFake video detection.

Extensive experiments are conducted to validate the effectiveness of masked relation learning. Our framework significantly outperforms the state-of-the-art methods by 2% AUC scores in the cross-dataset evaluations on Celeb-DF [45] and DFDC [46]. Besides, our method is able to localize the face forgery without additional supervision, which serves as an intuitive interface for the explanation of the model.

The main contributions are summarized as follows:

- We provide a new insight that formulates DeepFake detection as a graph classification problem to leverage the relationship between facial regions.
- We propose a masked relation learning framework that extracts attentional features of multiple facial regions and masks partial relations between regions to learn critical relationships for information aggregation and propagation.
- Extensive experiments demonstrate that the proposed masked relation learning achieves superior generalization ability on DeepFake video detection.

II. Related Work

In this section, we review previous studies on DeepFake detection, relation learning, and masked graph modeling.

A. DeepFake Detection

Early DeepFake detectors attempt to expose fake faces through image forensic patterns [47] and physiological signals [48], [49], [50]. But they are incompetent to detect realistic face forgery. Deep learning [14], [16] becomes a predominant paradigm for DeepFake detection owing to its strong representation capability. CNNs including Xception [15], ResNet-50 [51], and EfficientNet [52] are effective to learn discriminative characteristics of specific face manipulation algorithms. However, it is an endless rivalry between DeepFake and detection. Although the aforementioned methods can accurately detect the known DeepFake techniques, their performances inevitably plummet when facing novel face manipulations. The advanced approaches resort to visual attention mechanism [23], [24], [35], [53], which capture subtle artifacts from local parts for better generalization abilities. Reference [23] proposes a multi-attentional network to extract fine-grained features from multiple facial regions. Frequency domain features [26], [35], [53], [54] assist classifiers to capture fine-grained clues of face forgery. Reference [26] mines fine-grained clues of face forgery by a Progressive Enhancement Learning (PEL) framework. Reference [53] exploits frequency attention distillation and multi-view attention distillation to detect low-quality deepfakes. Reference [35] designs an RGB-Frequency Attention module that collaboratively learns comprehensive features in both spatial and frequency domains to improve the generalization and robustness of DeepFake detection. Single-center loss [54] clusters real faces while separating real faces from fake faces in the adaptive frequency feature space.

Apart from spatial artifacts, temporal artifacts have attracted the attention of researchers in recent years. Discontinuity of eye blinking [48], lip motion [11], and facial landmarks [55] are extracted as effective clues for DeepFake detection. Two-branch recurrent network [30] suppresses high-level face content to amplify artifacts of forged videos. SMIL [31] detects partially manipulated faces in deepfake videos by sharp multiple instance learning. Reference [12] utilizes Vision Transformers (ViTs) [22] to spot the inconsistency between adjacent frames. Furthermore, STIL [32], Intra-SIM [33], and HCIL [56] capture region-aware inconsistency from local patches and fuse snippet-aware features among global snippets.

In short, the existing works mainly attend to visual artifacts and temporal irregularities in local regions. However, there are few studies involving the relationship between features of local regions. To this end, we propose a masked relation learning to further investigate the relationship between facial regions for DeepFake video detection.

B. Relation Learning

Relation learning refers to modeling the interrelation between objects or local parts. It is widely adopted for visual tasks such as face recognition [57], person re-identification [58], video grounding [59], and scene graph generation [60]. Different from convolutional features, relational features can flexibly encode non-Euclidean structure data [61] like chemical molecule [62], [63] and point clouds [64], [65]. They contain comprehensive relation-aware information including compositional relationship [58], [60], [66], geometric structure [57], and temporal context [59]. Graph Convolution Networks (GCNs) [61] exert profound influence in this field, which respectively model object or entity as vertex and connectivity between objects as edge. However, measuring connectivity between vertices is a critical yet difficult entry. A straightforward way is to annotate the adjacency matrix

of graph [67]. Obviously, it is labor-extensive and unsustainable for complex and massive visual data. Self-attention mechanism [68] offers a good solution to adaptively learn the correlations between vertices. The relationships among the facial components effectively mitigate the gap of spectral domains for face recognition [57], [69].

As for DeepFake detection, [34], [35], [36], [37] are the most related studies on relation learning. Pixel-Region relation network (PRRNet) [34] analyzes both pixel-wise and region-wise relations to expose face forgery. Rao et al. [36] design a multi-semantic Conditional Random Field (CRF) based attention module that uses CRF to learn the local spatial correlations among adjacent pixels. Reference [35] proposes a Multi-scale Patch Similarity Module (MPSM) to measure the similarity between local patches with the fusion of features in both RGB and frequency domains. Multimodal Contrastive Classification by Locally Correlated Representation (MC-LCR) [37] captures the frequency discrepancies from local correlations between patch-wise amplitude and phase. Overall, current studies on relation learning for DeepFake detection mainly learn the relations among pixels [34], [36] and patches [35], [37]. Besides, patterns in the frequency domain are effective clues of face forgery [20]. Many approaches [35], [37] attempt to fuse frequency-aware and spatial forgery patterns to improve the expressiveness of relations.

In this paper, we first investigate masked relation learning for DeepFake detection. The masked strategy is devised to disconnect the minimal edges in the learned relations. It facilitates learning the generic discriminative representation from relations instead of visual artifacts. Unlike [35], we only extract features and relational information from spatial space. Moreover, our method only relies on the region-wise relations and reduces the computational overhead [34], [36].

C. Masked Graph Modeling

The masked language modeling [39], [70], [71] (MLM) proves that masking partial tokens can boost the performance of language models. Witnessing the success of MLM, recent studies [38], [72] indicate that masked modeling is also effective for computer vision tasks. Masked Autoencoders (MAEs) [38] are the pioneering methods that learn visual representation by reconstructing masked image patches. The rationale behind masked modeling is information redundancy. Through predicting the masked words or patches, models tend to understand the high-level semantic representations with few inductive biases [38]. It is favorable to the generalization of neural networks.

Several studies [40], [41], [42], [62] have explored masked graph modeling. Since the connected vertices tend to have similar attributes, the masked graph convolution network (Masked GCN) [41] only propagates partial vertices' attributes to their neighbors. Reference [42] designs three self-supervised tasks, *i.e.*, autoencoding, corrupted input reconstruction, and corrupted embedding reconstruction to assist the training of GNNs. Furthermore, [43] discovers that pretext tasks have limited improvement and develops advanced pretext tasks to exploit global self-supervised information. Reference [40] points out that graph autoencoders (GAEs) [43], [73], [74] suffer from four challenges including over-emphasized objective, vulnerable feature reconstruction, unstable criterion, and little expressive decoder. Masked graph autoencoder (GraphMAE) [40] utilizes masked feature reconstruction, re-mask decoding, and scaled cosine error to address these challenges.

Most works on masked graph modeling adopt self-supervised learning to predict the masked vertices and edges of a graph. We observe that the performance of the graph-based DeepFake detector can be improved just by masking edges during training. It encourages the model to depict the global structure of the face and learn task-specific representations. It is interesting to explore the self-supervised masked relation learning for DeepFake detection, which will be our research in the future.

III. Method

The framework of masked relation learning is shown in Figure 2. It consists of two main components: a spatiotemporal attention (STA) module and a masked relation Learner (MRL). STA simultaneously extracts attention features of multiple facial regions from a video snippet. MRL models the relations among regions and corrupts partial relations in the training procedure. For better understanding, we describe a brief overview of our framework before introducing details of the architecture.

A. Overview

We conduct DeepFake detection at the video level. The workflow of our method is divided into two stages: attention representation and masked relation learning. Given a video, frames are divided into a sequence of snippets I^t, where each snippet has the same length D. The index t ranges from 0 to $T-1$. In the first stage, a 3D-CNN [75] is used as the backbone to extract feature map F_0^t. Then STA module produces N attention maps F_A^t. The attention features F^t are the products of F_0^t and F_A^t. Each attention map corresponds to a specific facial region.

In the second stage, the attention features F^t are flattened into vertices $\mathcal{V}^t$, where $\mathcal{V}^t = \{v_1^t, v_2^t, \ldots, v_N^t\}$. The edges $\mathcal{E}$ among vertices are learnable parameters $\mathbb{R}^{N\times N}$. $\mathcal{E}_{i,j}$ represents the correlation between vertex v_i and vertex v_j. MRL module discards redundant edges to learn the important relational features. As shown in Eq. (1), it distributes a mask indicator $\eta_{i,j}$ to each edge $\mathcal{E}_{i,j}$, $\eta_{i,j} \in \{0, 1\}$. The edge $\mathcal{E}_{i,j}$ is set to zero when $\eta_{i,j} = 0$. Otherwise, the edge $\mathcal{E}_{i,j}$ remains unchanged when $\eta_{i,j} = 1$.

$$\mathcal{E}_{i,j} = \eta \cdot \mathcal{E}_{i,j}. \tag{1}$$

The graph is constructed as $\mathcal{G}^t =< \mathcal{V}^t, \mathcal{E} >$. We adopt a Temporal Convolution Network (TGCN) [67] to learn relational features from a sequence of video snippets. At the timestep t, the TGCN performs an interaction between the input graph $\mathcal{G}^t$ and the hidden graph $\mathcal{H}^t$ to exchange the relational information. It updates the hidden graph as follows:

$$\mathcal{H}^{t+1} = \psi(\mathcal{G}^t, \mathcal{H}^t). \tag{2}$$

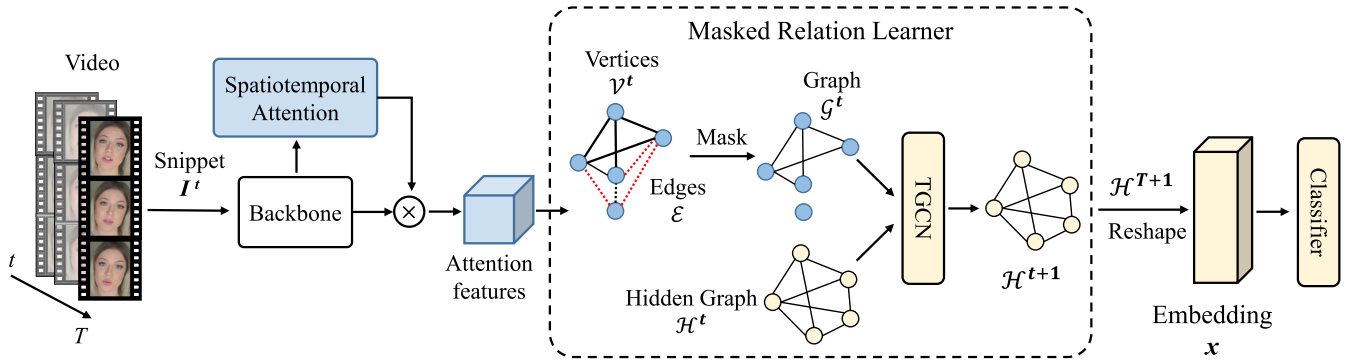

Fig. 2. **The architecture of masked relation learning for DeepFake video detection.** It comprises spatiotemporal attention (STA) and masked relation learner (MRL). STA extracts attentional features from multiple facial regions as vertices. MRL represents the relationship between vertices and then exploits masked graph modeling to reduce redundant relational information. Framework detects the temporal inconsistency in relational information with a TGCN. Finally, the hidden graph from the last video snippet is fed to a graph classifier for binary classification.

The function ψ is a gated recurrent unit (GRU) [76] to determine which relational information to be memorized and which information to be forgotten. The TGCN recursively updates the hidden graph and learns relational features from the consecutive snippets. Finally, the last hidden graph $\mathcal{H}^T$ is fed to a graph classifier to predict whether the input video is real or fake.

B. SpatioTemporal Attention

The spatiotemporal attention (STA) module aims to attach high weights to various face regions. Extended from [23], [77], we take the temporal features into consideration. It fosters the model's perception in both spatial artifacts [23] and temporal incoherence [12]. STA module comprises two 3D convolution blocks. Each block contains a $3\times3\times3$ convolution layer, a 3D batch normalization layer, and an activation layer ReLU. Firstly, we use the backbone to extract the initial feature map $F_0^t \in \mathbb{R}^{C\times D\times H\times W}$, where C is the number of channels, D, H, and W indicate length, height, and width, respectively. Then STA module transforms the feature maps F_0^t into attention maps $F_A^t \in \mathbb{R}^{N\times H\times W}$, where N is the number of attention heads. Attention maps focus on facial regions such as mouth, nose, and eye. Each attention map is supposed to consistently coordinate with one specific region. Then we use the normalized average pooling [23] to flatten the attention features into embeddings $\mathcal{V}^t$. The normalized average pooling is defined in Eq. (3). Compared with average pooling, max pooling [78] discards partial features of smooth regions and enhances facial textures. Recent works [26], [79] show the effectiveness of max pooling to amplify the blending boundary of forged faces. However, relation learning requires the complete information of each facial region to construct a vertex. Average pooling aggregates the attention features of facial regions, which are essential for relation learning. Therefore, average pooling is adopted in STA module. The difference between average pooling and max pooling is further analyzed in section IV-E.1.

$$F^t = F_0^t \bigodot F_A^t,$$
$$\mathcal{V}^t = \frac{\sum_i^H \sum_j^W F_{i,j}^t}{\| \sum_i^H \sum_j^W F_{i,j}^t \|_2}, \tag{3}$$

where $\mathcal{V}^t \in \mathbb{R}^{N\times C}$ and $\bigodot$ denotes the Hadamard product.

Moreover, [80] suggests that the multi-head attention mechanism is inclined to generate similar attention features. It leads to an overlap of attention features and low expressiveness of the model. In STA module, different attention heads are required to concentrate on different facial regions. Inspired by orthogonal regularizers [80], [81], [82], we use an orthogonal diversity loss $\mathcal{L}_{od}$ to decrease the correlation of attention features. Since the correlation of attention is unrelated to time, the orthogonal diversity loss $\mathcal{L}_{od}$ is formulated as:

$$\mathcal{L}_{od} = \frac{1}{T}\sum_{t=1}^{T} \| \mathcal{V}^t \mathcal{V}^{t\prime} - I \|_F^2, \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\mathcal{V}^{t\prime}$ is the transpose of $\mathcal{V}^t$, $I \in \mathbb{R}^{N\times N}$ is an identity matrix.

In the time dimension, each attention head is supposed to track its facial region in the whole video. Hence, we propose a temporal consistency loss $\mathcal{L}_{tc}$ to shorten the distance between attention features of adjacent frames. The temporal consistency loss $\mathcal{L}_{tc}$ is formulated as:

$$\mathcal{L}_{tc} = \frac{1}{N}\sum_{t=1}^{T} \| \mathcal{V}^t - \mathcal{V}^{t-1} \|_2, \tag{5}$$

where t ranges from 1 to $T-1$.

C. Masked Relation Learner

After getting features of various facial regions by STA, MRL is devised to learn the relational information among facial regions. Concretely, a graph $\mathcal{G}$ is constructed to model the relationship, whose vertices $\mathcal{V}$ are the features of facial regions while edges $\mathcal{E}$ are the relations between regions. The edges $\mathcal{E}$ are learnable parameters of MRL. They are initialized with a Gaussian random matrix $\mathbb{R}^{N\times N}$ before training the model.

We provide an example to introduce the motivation of MRL in Fig. 3. According to the relational matrix, the correlation value between the left eye and nose is high. It implies that the left eye has a strong relationship with the nose. However, the correlations from the left eye and nose to the right eyebrow are low. These two regions have weak relation with the right eyebrow. This phenomenon indicates relational information among facial regions exists redundancy. Inspired

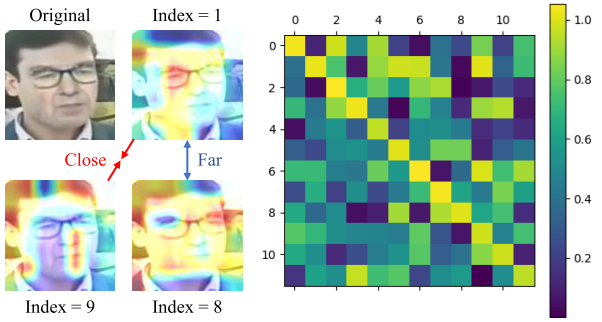

Fig. 3. **Facial regions (left) and relational matrix (right)**. The left eye (region 1) has a strong relationship with the nose (region 9). But the left eye and nose have weak relation with the right eyebrow (region 8). Best viewed in color.

by observation [38], [72] that low redundant representations learn the intrinsic features, we decrease the redundancy of relational information to learn compact and generic facial features. Specifically, MRL masks partial edges during training as shown in Eq. (1). There are two masking strategies for MRL module: *minimal* and *random*. The *minimal* strategy sets the minimal edges to zero. The proportion p of masked edges is predefined, $p \in [0, 1]$. Concretely, we measure the p-th quantile of the edges. The edge $\mathcal{E}_{i,j}$ is set to zero if its value is less than the quantile q. The *minimal* strategy is formulated as follows:

$$q = \mathcal{E}_{min} + (\mathcal{E}_{max} - \mathcal{E}_{min}) \cdot p,$$
$$\mathcal{E}_{i,j} = \begin{cases} 0 & \mathcal{E}_{i,j} < q \\ \mathcal{E}_{i,j} & \mathcal{E}_{i,j} \geq q \end{cases}, \tag{6}$$

where $\mathcal{E}_{min}$ and $\mathcal{E}_{max}$ are the minimal and maximal values of the edges, respectively.

Moreover, the *random* strategy adopts the Bernoulli trials to randomly sample the masked indicators η. Given a masking rate p, we individually sample the $\eta \sim \text{Bernoulli}(p)$ for every edge. As a Bernoulli random variable, η becomes 0 with a probability of p. The *random* strategy is formulated as follows:

$$\mathcal{E}_{i,j} = \eta * \mathcal{E}_{i,j},$$
$$s.t.\ \eta \sim \text{Bernoulli}(p). \tag{7}$$

In practice, the *random* strategy can be easily implemented with the Dropout [78], [83]. But it is different from the Dropout in that the value of edge is not scaled up by a factor of $\frac{1}{1-p}$. The masked edge can not propagate relational information to the linked vertices. For better understanding, we describe these two masking strategies in Algorithm 1.

Furthermore, MRL exploits a TGCN to capture irregularity of relations to expose DeepFake videos. As shown in Fig. 4, the TGCN contains a graph convolution layer and a GRU. It adaptively stores relational information in the hidden graph and forgets unrelated information. Noted that an initial hidden graph $\mathcal{H}^0$ is an $N \times 3C'$ zero matrix. Given the t-th snippet, embeddings of facial regions are vertices $\mathcal{V}^t \in \mathbb{R}^{N \times C}$. Each vertex is a vector that represents features of a specific facial region. Different from CNNs [51], we exploit graph convolution [61] for every vertex to aggregate information from its neighboring vertices. The graph $\mathcal{G}^t =< \mathcal{V}^t, \mathcal{E} >$ is constructed

Algorithm 1 Masking Strategies

Input: Edges E; Masking rate p; Strategy stg.
Output: The masked edges E.

```
if stg=='minimal' then
    # Minimal masking strategy.
    low = min(E); high = max(E)
    quantile = low + (high - low) * p
    q_mat = ops.gt(E, quantile)
else if stg=='random' then
    # Random masking strategy.
    q = ops.dropout(E, p)
    q_mat = ops.gt(E, q)
end if
E = ops.mul(q_mat, E)
return E
```

by a graph convolution layer as follows:

$$\begin{aligned} \mathcal{G}^t &= \text{GConv}(\mathcal{V}^t, \mathcal{E}) \\ &= \mathcal{E}\ \mathcal{V}^t\ W_g \\ &= (\mathcal{G}_r^t,\ \mathcal{G}_z^t,\ \mathcal{G}_h^t), \end{aligned} \tag{8}$$

where $W_g \in \mathbb{R}^{C \times 3C'}$ is a weight matrix for the input graph, C' indicates the number of channels in hidden graph. We split the graph $\mathcal{G}^t$ into three latent matrices $\mathcal{G}_r^t, \mathcal{G}_z^t, \mathcal{G}_h^t$. The shape of each latent matrix is $\mathbb{R}^{N \times C'}$. Similarly, the hidden graph $\mathcal{H}^t$ is composed of vertices $\mathcal{V}_{\mathcal{H}}^t$ and edges $\mathcal{E}$. The edges $\mathcal{E}$ are the same as the edges of graph $\mathcal{G}^t$. The initial vertices of hidden graph $\mathcal{V}_{\mathcal{H}}^0 \in \mathbf{0}^{N \times C'}$ are zero vectors. A hidden graph $\mathcal{H}^t =< \mathcal{V}_{\mathcal{H}}^t, \mathcal{E} >$ is constructed by a graph convolution layer and then split into three hidden matrices $\mathcal{H}_r^t, \mathcal{H}_z^t, \mathcal{H}_h^t$:

$$\begin{aligned} \mathcal{H}^t &= \text{GConv}(\mathcal{V}_{\mathcal{H}}^t, \mathcal{E}) \\ &= \mathcal{E}\ \mathcal{V}_{\mathcal{H}}^t\ W_h \\ &= (\mathcal{H}_r^t,\ \mathcal{H}_z^t,\ \mathcal{H}_h^t), \end{aligned} \tag{9}$$

where $W_h \in \mathbb{R}^{C' \times 3C'}$ is a weight matrix for the hidden graph.

Subsequently, the TGCN performs gate operations to update the hidden graph through the following four steps:

$$\begin{aligned} r^t &= \sigma(\mathcal{G}_r^t + \mathcal{H}_r^t + b_r), \\ z^t &= \sigma(\mathcal{G}_z^t + \mathcal{H}_z^t + b_z), \\ \tilde{\mathcal{H}}^t &= \phi(\mathcal{G}_h^t + r^t \odot \mathcal{H}_h^t + b_h), \\ \mathcal{H}^{t+1} &= z^t \odot \mathcal{H}^t + (1 - z^t) \odot \tilde{\mathcal{H}}^t, \end{aligned} \tag{10}$$

where r^t and z^t denote a reset gate and an update gate of GRU, respectively. $\tilde{\mathcal{H}}^t$ is a candidate hidden state that memorizes the temporal information. b_r, b_z, b_h are the biases. σ and ϕ are sigmoid and tanh functions, respectively. The hidden graph $\mathcal{H}^{t+1}$ is recursively updated until the last snippet I^T is processed.

In the end, the last hidden graph $\mathcal{H}^T$ is reshaped into a feature vector x with a length of $N \times C'$. We feed it into a graph classifier to predict whether the input video is real or fake. The classifier is a multi-layer perceptron trained by minimizing the binary cross entropy loss $\mathcal{L}_{ce}$:

$$\mathcal{L}_{ce} = y \log f(x) + (1 - y) \log(1 - f(x)), \tag{11}$$

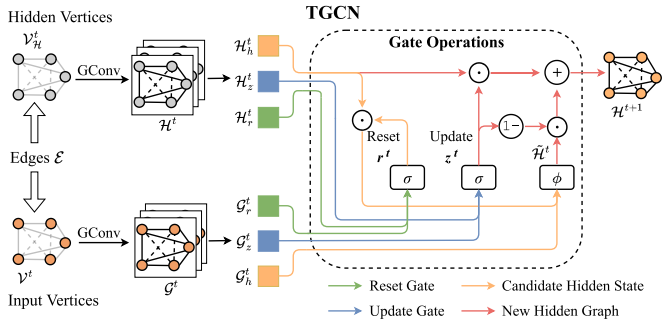

Fig. 4. **The architecture of TGCN.** The edges are shared for graph convolution layers.

where f denotes the classifier and y is the label of the input video.

Overall: The loss function of the whole framework is combined with the binary cross entropy loss $\mathcal{L}_{ce}$, the orthogonal diversity loss $\mathcal{L}_{od}$, and the temporal consistency loss $\mathcal{L}_{tc}$.

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{od}\mathcal{L}_{od} + \lambda_{tc}\mathcal{L}_{tc}, \tag{12}$$

where λ_{od} and λ_{tc} are hyper-parameters to adjust the constraints of attention in terms of spatial diversity and temporal consistency, respectively.

In conclusion, we introduce the pseudocode of training our framework in Algorithm 2.

IV. Experiments

First and foremost, We conduct experiments to validate the effectiveness of masked relation learning for DeepFake detection. We compare the performance and generalization ability of our proposed framework with the state-of-the-art methods on three public benchmarks. Besides, the contributions of the proposed spatiotemporal attention, orthogonal diversity loss, temporal consistency loss, and masked relational learning are assessed. More importantly, we further analyze our method to answer three concerned research questions as follows:

- **RQ1:** How do masking strategies and masking rates affect the performance of DeepFake detection?
- **RQ2:** Is it beneficial to mask partial edges during model inference?
- **RQ3:** Can our method detect the authenticity of a single image?

Algorithm 2 The Training Procedure of Masked Relation Learning for DeepFake Detection

Input: An input video; Length of sequence T; Masking rate p; Masking strategy; 3D CNN f; STA module M; Edges $\mathcal{E}$; GCN Cell G; Hidden graph $\mathcal{H}$; Graph classifier C.
Output: The prediction $\hat{y}$.

```
Initialization: E ∈ R^{N×N}, H^0 ∈ 0^{N×d}.
Divide the input video into T snippets, each snippet I^t ∈ R^{3×D×H×W}.
for epoch to Total Epoch do
    Mask edges E with Algorithm 1;
    for t = 0 to T − 1 do
        F_t^0 = f(I_t);
        Attention maps F_t^A = M(F_t^0);
        Calculate vertices V^t based on F_t^A and F_t^0; Eq. (3)
        Graph G^t = GConv(V^t, E); Eq. (8)
        Update hidden graph H^{t+1} = G(G^t, H^t); Eq. (10)
    end for
    x = H^T.flatten();
    ŷ = C(x);
    Update parameters of E, f, M, G, and C. Eq. (12)
end for
return parameters of E, f, M, G, and C.
```

A. Experimental Setup

1) Datasets: In the experiments, the intra-dataset evaluation and cross-dataset evaluation are performed on three public benchmarks, including FaceForensics++ (FF++) [15], Celeb-DF [45], and DeepFake Detection Challenge (DFDC) [46].

- **FaceForensics++ (FF++)** [15]: a standardized dataset for DeepFake detection. It consists of 1,000 pristine videos and 4,000 fake videos. Four manipulation techniques are used to generate fake videos, including DeepFakes[1] (DF), Face2Face (F2F) [84], FaceSwap[2] (FS), and NeuralTextures (NT) [85]. To simulate the setting of social networks, FF++ has high-quality (HQ) and low-quality (LQ) copies created by light compression and heavy compression, respectively.
- **Celeb-DF** [45]: a large-scale deepfakes dataset. It contains 590 real videos and 5,639 fake videos of celebrities. An undisclosed improved synthesis algorithm is devised to produce face forgeries. The realistic forgeries make it difficult for DeepFake detection.
- **DeepFake Detection Challenge (DFDC)** [46]: a public faceswap video dataset. It contains 1,131 real videos and 4,119 fake videos. Six advanced faceswap algorithms are used to craft fake videos. The real videos are filmed in a variety of real-world scenes. Many distractors such as dark lighting, extreme pose, and occlusion lead to challenging forgery detection.

2) Implementation Details: For data pre-processing, we follow [32], [33] to use Dlib [86] as the face detector for FF++ dataset and use MTCNN [87] for other datasets. The cropped face images are resized to 112×112. Only random horizontal flip is used for data augmentation during training. The 3D CNN, Mixed Convolution network (MC3-18) [75] pre-trained on Kinetics-400 [88] is exploited as the backbone of our framework. For each video, we sample $T = 5$ snippets and each snippet has $D = 20$ frames. The number of attention heads in SpatioTemporal Attention is $N = 12$. We adopt an Adam [89] optimizer to train the model. The initial learning rate is 1×10^{-4}. We decrease the learning rate by a factor of 2 after every 5 epochs. We train the model for 30 epochs with a batch size of 30. The hyper-parameters $\mathcal{L}_{od}$ and $\mathcal{L}_{tc}$ are both 0.75.

3) Evaluation Metrics: We follow the previous studies [12], [15], [30], [32], [33], [35] to apply Accuracy score (ACC) and Area Under the Receiver Operating Characteristic Curve

[1]https://github.com/deepfakes/faceswap
[2]https://github.com/MarekKowalski/FaceSwap

(AUC) as evaluation metrics. Since our method predicts the authenticity of a video rather than a single image, we follow [11], [30], [33] to measure video-level AUC for a fair comparison. For each video, all the predictions of frames or clips are averaged as the final result. The number of sampled frames is the same.

4) Baselines: We compare our method with representative competitors. The 3D ResNet MC3-18 [75] is selected as our baseline. The most related methods include Local Relation Learning (LRL) [35], MC-LCR [37], PRRNet [34], and Multi-attentional DeepFake Detection (MADD) [23]. In addition, 3D CNNs C3D [90] and I3D [88] are selected for comparisons. The advanced video-based DeepFake detectors HCIL [56], Intra-SIM [33], STIL [32], SMIL [31], ADDNet-3D [91], and Two-branch [30] are chosen. We also make comparisons with the image-based methods SIA [28], PEL [26], F3-Net [20], Face X-Ray [18], and Xception [15].

B. Intra-Dataset Evaluation

In the intra-dataset evaluation, the training set and testing set are sampled from the same dataset. It evaluates the performance of models on detection accuracy. We conduct the intra-dataset evaluations on FF++, Celeb-DF, and DFDC.

1) FF++: The results are shown in TABLE I. We discover that the relation learning based methods LRL [35], MC-LCR [37], and our approach outperform the advanced image-based and video-based methods MADD [23], ADDNet-3D [91], and Two-branch [30] on FF++ (LQ). It reveals the efficacy of relation learning for exposing deepfakes. The interaction across various facial regions promotes networks to learn the global structure of the face. The relational features have stronger inductive biases [92] than visual features. Hence, relational learning enhances the networks' robustness against video compression [93] and obtains satisfactory results on FF++ (LQ).

In addition, the video-based methods perform better than the image-based methods. The rationale behind this phenomenon is that subtle artifacts are smoothed by video compression and become hard to be captured [32]. The video-based methods effectively address this issue by detecting temporal incoherence. In the intra-dataset evaluation on FF++ (LQ), our method exceeds LRL [35] by 0.97% in terms of AUC. The outstanding performance derives from two properties. On one hand, masked relation learning reduces the redundancy of relations and encourages the network to model general face geometry. On the other hand, relation propagation exchanges relational information across facial regions. It allows the network to detect face forgery from a global view of interaction. Although our method has sub-optimal performance on FF++ (HQ), it outperforms state-of-the-art approaches on detecting unseen deepfake datasets. Strong generalization ability is one of the pursuits of deepfake detection. Our method effectively promotes the generalization ability of detectors.

2) Celeb-DF & DFDC: We also perform intra-dataset evaluations on Celeb-DF [45] and DFDC [46] datasets. The results evaluations are shown in TABLE II. Our method achieves the best results among state-of-the-art competitors. The AUC scores are 99.96% and 99.11% on Celeb-DF and DFDC, respectively. It surpasses HCIL [56] by 4% on DFDC. SIA [28] has the same result as ours on Celeb-DF. The remarkable performance validates the effectiveness of masked relation learning for DeepFake detection.

TABLE I

INTRA-DATASET EVALUATIONS ON FF++. HIGH-QUALITY (HQ) AND LOW-QUALITY (LQ) SETTINGS OF FF++ ARE USED

Method	FF++ (HQ)		FF++ (LQ)	
	ACC	AUC	ACC	AUC
Face X-Ray [18] (CVPR 2020)	95.97	-	82.94	-
Xception [15] (ICCV 2019)	95.73	96.30	86.86	89.30
F^3-Net [20] (ECCV 2020)	97.52	98.10	90.43	93.30
MADD [23] (CVPR 2021)	97.60	99.29	88.69	90.40
PEL [26] (AAAI 2022)	97.63	99.32	90.52	94.28
SIA [28] (ECCV 2022)	97.64	99.35	90.23	93.45
ADDNet-3D [91] (MM 2020)	96.78	97.74	87.50	91.01
Two-branch [30] (ECCV 2020)	96.43	98.70	86.34	86.59
C3D [90] (CVPR 2015)	90.72	-	86.79	-
I3D [88] (CVPR 2017)	93.13	-	86.88	-
MC3 [75] (CVPR 2018)	94.91	98.31	83.87	92.98
S-MIL-T [31] (MM 2020)	98.39	-	92.77	-
STIL [32] (MM 2021)	98.57	-	94.82	-
Intra-SIM [33] (AAAI 2022)	98.93	-	96.78	-
HCIL [56] (ECCV 2022)	**99.01**	**-**	**96.78**	**-**
PRRNet [34] (PR 2021)	96.15	-	86.13	-
LRL [35] (AAAI 2021)	97.59	99.46	91.47	95.21
MC-LCR [37] (KBS 2022)	97.89	**99.65**	88.07	90.28
MRL	93.82	98.27	91.81	**96.18**

TABLE II

INTRA-DATASET EVALUATIONS ON CELEB-DF AND DFDC. THE AUC SCORES (%) ARE REPORTED

Method	Celeb-DF	DFDC
Xception [15] (ICCV 2019)	99.44	84.58
SIA [28] (ECCV 2022)	**99.96**	90.96
ADDNet-3D [91] (MM 2020)	95.16	79.66
I3D [88] (CVPR 2017)	99.23	80.82
MC3 [75] (CVPR 2018)	95.90	93.60
S-MIL-T [31] (MM 2020)	98.84	85.11
STIL [32] (MM 2021)	99.78	89.80
Intra-SIM [33] (AAAI 2022)	99.61	92.79
HCIL [56] (ECCV 2022)	99.81	95.11
MRL	**99.96**	**99.11**

C. Cross-Dataset Evaluation

The cross-dataset evaluation means that the training set and testing set are derived from different datasets. It tests the generalization ability to detect unseen DeepFake techniques. Following [15], [30], [35] for a fair comparison, we train models on FF++ (LQ) and test them on Celeb-DF and DFDC. TABLE III shows the results of cross-dataset evaluations. Although the image-based methods perform well on FF++ (LQ), their weak generalization abilities lead to unsatisfactory results on unseen datasets. In contrast, the video-based methods significantly surpass the image-based methods owing to their strong generalization abilities. Nevertheless, the relation-aware deepfake detectors MC-LCR [37], LRL [35], and our method perform better than others. Even if the results on FF++ (LQ) are inconspicuous, relation-aware methods favorably encode relational representations that

TABLE III
CROSS-DATASET EVALUATIONS. THE AUC SCORES (%) ARE REPORTED

Method	FF++ (LQ)	Celeb-DF	DFDC
Xception [15] (ICCV 2019)	95.50	65.50	59.39
F^3-Net [20] (ECCV 2020)	93.30	67.95	57.87
MADD [23] (CVPR 2021)	90.40	68.64	63.02
PEL [26] (AAAI 2022)	94.28	69.18	63.31
SIA [28] (ECCV 2022)	**96.94**	77.35	-
ADDNet-3D [91] (MM 2020)	91.01	57.83	51.60
MC3 [75] (CVPR 2018)	92.98	70.83	66.19
STIL [32] (MM 2021)	-	75.58	67.88
Intra-SIM [33] (AAAI 2022)	-	77.65	68.43
HCIL [56] (ECCV 2022)	-	79.00	69.21
MC-LCR [37] (KBS 2022)	90.28	71.61	71.34
LRL [35] (AAAI 2021)	95.21	78.26	**76.53**
MRL	96.18	**83.58**	71.53

aggregate embeddings of various facial regions. Relational representations are common across different face datasets and DeepFake techniques. It helps the models discern unseen face forgeries from real faces. In the cross-dataset evaluation of Celeb-DF, our method outperforms the state-of-the-art methods HCIL [56] and LRL [35] over 4.58% and 5.32% in terms of AUC, respectively. We observe that our method trails behind LRL [35] in the cross-dataset evaluation of DFDC. It outperforms the video-based competitor HCIL by 2.32%.

D. Analyses of Maksed Relation Learning

Apart from performance comparisons, we further analyze the properties of masking relation learning to explore the aforementioned research questions.

1) RQ1: Masking Strategy and Rate: There are two masking strategies for masked relation learning, *i.e.*, *minimal* and *random*. It is desirable to investigate which masking strategy and how a large masking rate can achieve optimal performance. Concretely, we construct two series of variants. One series of variants mask the minimal edges, while the other ones mask random edges. Each series have 11 variants in which masking rates range from $[0.0, 0.1, \ldots, 1.0]$. We train these two series of variants on FF++ (LQ).

The quantitative results of cross-dataset evaluations are shown in Fig. 5. We observe that both the *minimal* and *random* masking strategies achieve optimal results at moderate masking rates. They suffer from severe degradation of performance since the masking rate is larger than 80%. The random masking strategy has its best results at the 60% masking rate. The minimal masking strategy performs best when the masking rate is 50%. It implies that the information density of relation is at the medium level [38]. Different from words and image patches, edges with large values are more important for the relationship between facial regions than those with small values. Some important edges are possibly destroyed by random masking. That is the rationale for the inferior performance of the random masking strategy.

2) RQ2: Mask During Inference: We are concerned about whether masking partial edges during inference can improve performance. Therefore, we test the variant namely 'Mask in inference' that performs the edge masking during inference. As a controlled experiment, the masking rate and masking strategy are the same as those of training for comparison. TABLE IV shows the result of this experiment. The variant has a worse performance than the original model. Because the model has learned a compact structure of the face by the edge masking. It is unnecessary to further mask edges during model inference. Otherwise, it harms the aggregation and propagation of relational information.

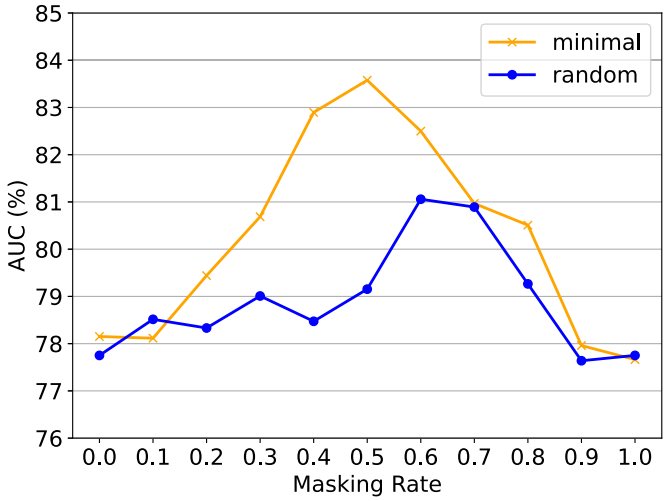

Fig. 5. **Quantitative analyses of minimal masking strategy (orange) and random masking strategy (blue).** The AUC scores of cross-dataset evaluation on Celeb-DF are reported.

3) RQ3: Image-Based DeepFake Detection: Obviously, the video-based methods have an intrinsic limitation in that they are unavailable to detect deepfake images. To address this limitation, we construct an image-based framework of masked relation learning for DeepFake detection. Specifically, we use a 2D CNN ResNet-18 [51] as the backbone. A spatial attention module is designed by replacing 3D layers of STA with 2D layers. We discard the GRU of MRL and only use graph convolution layers for relation learning. The image-based framework is trained on frames instead of snippets. For a fair comparison, we report the video-level accuracy and AUC score. TABLE IV shows unsatisfactory results of the image-based framework. The superiority of the video-based framework derives from TGCN and spatiotemporal attention. Compared with the image-based variant, the variant with spatial attention replaces the GCN with a TGCN. Its significant progress underlines the contribution of TGCN. TGCN helps the model capture temporal inconsistency in the relationship between facial regions. Moreover, the model with spatiotemporal attention attends to temporal inconsistency between multiple frames within a snippet and further boosts performance.

E. Ablation Study

Apart from comparisons and analyses, we conduct ablation studies to validate the efficacy of the proposed components. The ablation studies involve spatiotemporal attention, masked relation learning, orthogonal diversity loss, and temporal consistency loss. As shown in TABLE V, we build the baseline and variants as follows: (1) Baseline. The MC3-18 [75] is exploited as the baseline, whose classification head is modified to a binary classifier. (2) Variant with only STA

TABLE IV

QUANTITATIVE RESULTS OF ABLATION STUDIES. INTRA-DATASET EVALUATION IS CONDUCTED ON FF++ (LQ). CROSS-DATASET EVALUATIONS ARE CONDUCTED ON CELEB-DF AND DFDC

Variant	FF++ (LQ) ACC	FF++ (LQ) AUC	Celeb-DF AUC	DFDC AUC
Image-based	85.14	87.34	63.75	63.41
Spatial attention	87.36	94.64	77.66	66.54
Mask in inference	90.55	94.14	80.55	67.89
Max pooling	86.63	92.95	76.86	67.38
w/o $\mathcal{L}_{od}$	90.41	94.06	76.76	68.64
w/o $\mathcal{L}_{tc}$	89.39	94.77	81.98	68.88
Full	**91.81**	**96.18**	**83.58**	**71.53**

TABLE V

ABLATION STUDY OF THE PROPOSED COMPONENTS. INTRA-DATASET EVALUATION IS CONDUCTED ON FF++ (LQ). CROSS-DATASET EVALUATIONS ARE CONDUCTED ON CELEB-DF AND DFDC. THE RESULTS OF THE BASELINE ARE SHOWN IN THE FIRST ROW

STA	Graph	Masking	FF++ (LQ) ACC	FF++ (LQ) AUC	Celeb-DF AUC	DFDC AUC
✗	✗	✗	83.87	92.98	70.83	66.19
✓	✗	✗	87.65	94.01	73.80	67.52
✗	✓	✗	86.92	89.44	67.70	66.82
✓	✓	✗	90.52	94.86	78.15	68.95
✓	✓	✓	**91.81**	**96.18**	**83.58**	**71.53**

module. We remove MRL module from the proposed method. (3) Variant with only Graph module. We replace the attention maps with matrices of ones. The masking strategy is also canceled in the training procedure. (4) Variant with STA and Graph modules. Only the masking strategy is canceled.

1) Spatiotemporal Attention: Four questions about STA need to be investigated. (i) Is STA effective for DeepFake detection? (ii) Can average pooling be replaced with max pooling? (iii) How many attention heads are appropriate for STA? (iv) Are the 3D convolution layers necessary?

Firstly, we ablate STA for comparison. We replace the attention maps F_A^t with N matrices of ones $\mathbf{1}^{H\times W}$. Then vertices $\mathcal{V}^t$ are obtained by Eq. (3). Obviously, STA extracts features of different facial regions, which are the elements of relation learning. TABLE V shows that the variant without STA has the worst performance. Compared with the baseline, it decreases the AUC score by 3.54% in the intra-dataset evaluation and decreases the AUC score by 3.13% in the cross-dataset evaluation of Celeb-DF.

Secondly, we construct a variant in which spatiotemporal attention produces attention maps with max pooling [78]. As shown in TABLE IV, the variant has worse performance than the proposed method. Although max pooling is widely used to enhance textural artifacts [26], it loses partial features of facial regions and is susceptible to maximum. In contrast, normalized average pooling preserves attention features of facial regions and allows relation learning to model the global relationship among regions.

Thirdly, we construct variants with different numbers of attention heads. As shown in Fig. 6, STA with 12 attention heads achieves the best performance. We observe that the results on both FF++ (LQ) and DFDC firstly increase with the number of attention heads. However, the results on FF++ (LQ)

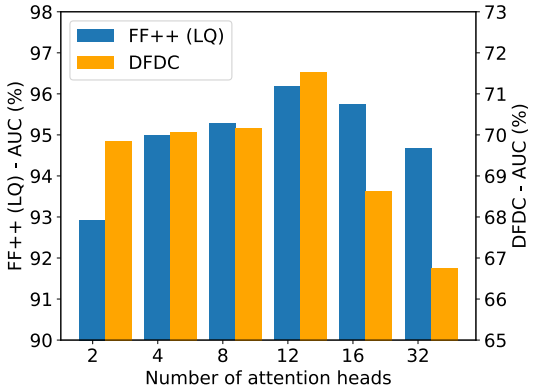

Fig. 6. **Quantitative analysis of the number of attention heads.** The blue bars and orange bars indicate AUC scores of cross-dataset evaluations on FF++ (LQ) and DFDC, respectively. Model is trained on FF++ (LQ).

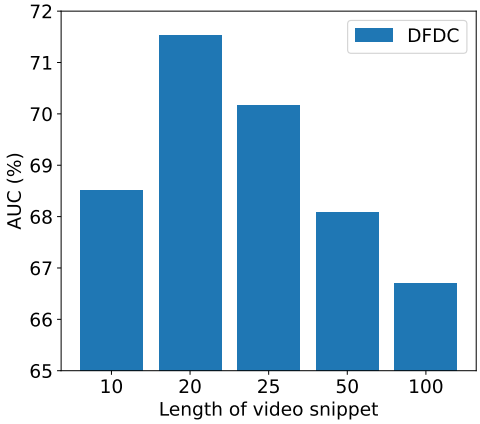

Fig. 7. **Quantitative analysis of length of video snippet.** Bars indicate AUC scores of the cross-dataset evaluation on DFDC.

Fig. 8. **Qualitative analysis of temporal consistency loss.** This video is sampled from FF++ NT. The warmer color indicates higher confidence in localizing the forged region.

gradually decline when the number of attention heads is larger than 12. This is a typical signal of over-fitting. Meanwhile, the decline of accuracy on DFDC is greater than that of accuracy on FF++ (LQ), which implies that excessive attention heads lead to over-fitting and weak generalization ability.

Finally, we investigate the necessity of 3D convolution in STA. A variant is constructed with a spatial attention module (SA) that only uses 2D convolution. It produces attention maps for each frame of the snippets. We use average pooling to gather attention maps of frames into attention maps $F_A^t \in \mathbb{R}^{N\times H\times W}$. TABLE IV shows the variant with SA is weaker than the proposed method. Because 3D convolution allows STA to capture temporal incoherence, our model can learn the consistency of relational information for better generalization.

2) Masked Relation Learning: We investigate the effectiveness of masked relation learning and the influence of snippet length. As shown in TABLE V, relation learning

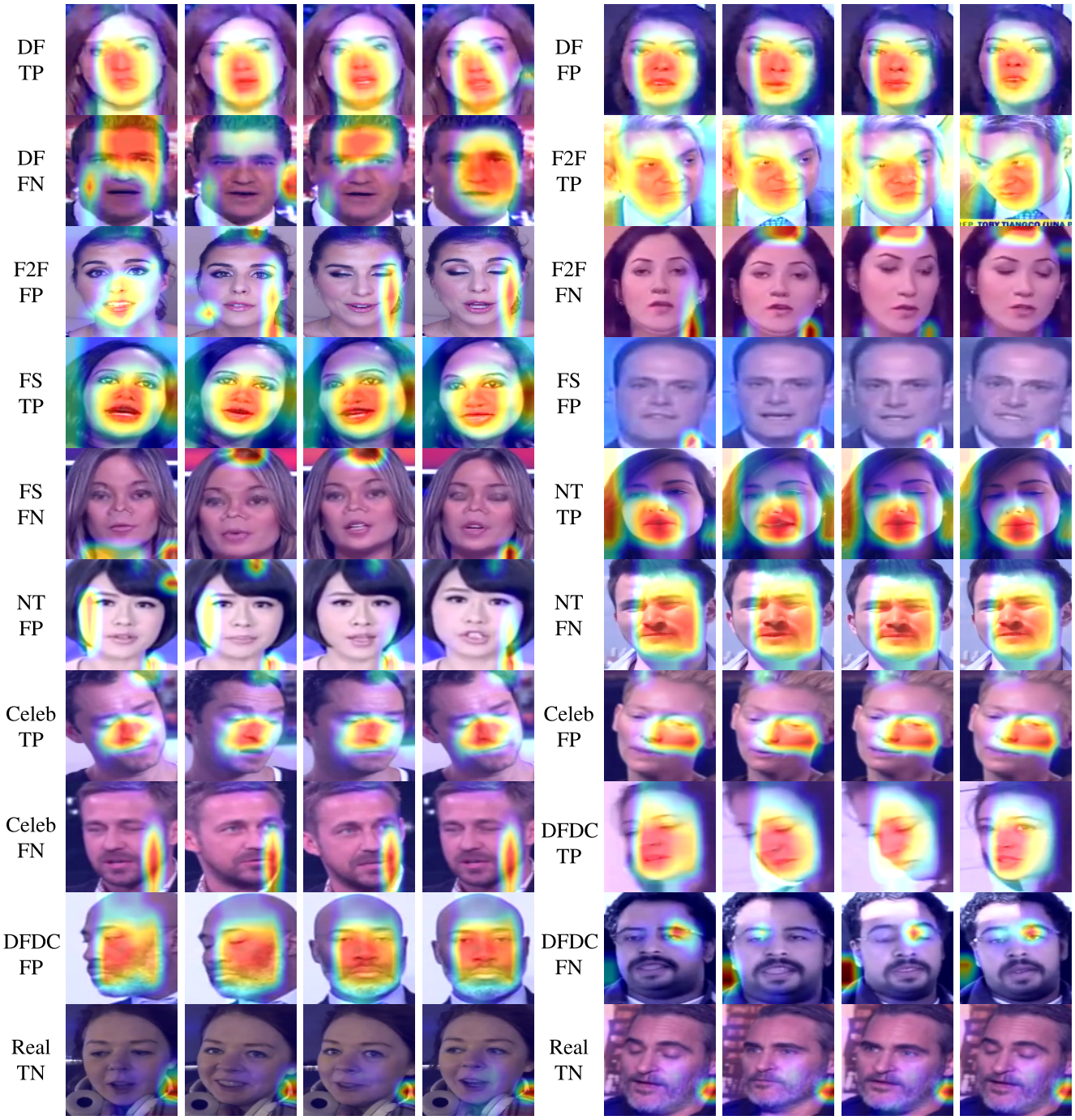

Fig. 9. **Results of DeepFake localization.** Celeb-DF is abbreviated to Celeb. TP means True Positive. FP means False Positive. FN means False Negative. TN means True Negative.

effectively promotes the generalization of DeepFake detection. Compared with the baseline, it improves the AUC scores from 70.83% to 78.15% and from 66.19% to 68.95% on the cross-dataset evaluations of Celeb-DF and DFDC, respectively. Masked graph modeling further improves the AUC scores by 5.43% and 2.58% on the cross-dataset evaluation of Celeb-DF and DFDC. It verifies the significant contribution of masked relation learning.

As for the influence of snippet length, we train four variants with different lengths of snippets. The lengths of each snippet include 10, 20, 25, 50, and 100. Fig. 7 illustrates that 20 and 25 are the appropriate lengths of snippets. Since the number of sampled frames is fixed, the snippets become fewer when they are longer. Too short or too long snippets hinder the relational interaction across different snippets from capturing long-term or short-term inconsistency.

3) Loss Functions: We test the effectiveness of the orthogonal diversity loss $\mathcal{L}_{od}$ and the temporal consistency loss $\mathcal{L}_{tc}$. We construct two variants namely w/o $\mathcal{L}_{od}$ and w/o $\mathcal{L}_{tc}$. These two variants are trained without the orthogonal diversity loss or the temporal consistency loss. Quantitative results are shown in TABLE IV. In the intra-dataset evaluation on FF++ (LQ), our method achieves 2.12% and 1.41% higher AUC scores than variants w/o $\mathcal{L}_{od}$ and w/o $\mathcal{L}_{tc}$, respectively. In the cross-dataset

evaluation on DFDC, our method achieves 2.89% and 2.65% higher AUC scores than these two variants. It indicates that both orthogonal diversity loss and temporal consistency loss are essential for extracting attention features of facial regions, which are the basics of relation learning. We further conduct qualitative analysis with Gradient-weighted Class Activation Map (Grad-CAM) [94]. Grad-CAM highlights the regions corresponding to the decisions of neural networks for a visual explanation. As shown in Fig. 8, the heatmaps of w/o $\mathcal{L}_{tc}$ gradually drift away from the face. The lack of temporal consistency regularization results in unstable attention features.

F. DeepFake Localization

Apart from accurate DeepFake detection, the human-centered explanation of DeepFake detectors is expected. Hereby we localize the fake regions by Grad-CAM in Fig. 9. We observe that our method can highlight the fake regions of manipulated faces. For instance, NeuralTextures only modifies the mouth region [15]. High responses locate at the mouth regions and correctly point out the face forgery. On the contrary, the detector has small responses outside the real faces. These areas usually keep unchanged after deepfake manipulation. Qualitative results illustrate the effective explainability of our approach for deepfake localization. Nevertheless, we also demonstrate failure cases of false positives and false negatives in all datasets. Though several false positives (e.g., FS) have small responses outside the face, they are misclassified into fake faces by a deepfake detector. False negatives have a similar problem. In essence, a credible deepfake localization requires an accurate detector. The accuracy and generalization ability of deepfake detection remain to be improved in future work.

V. Conclusion

This paper explores masked relation learning for DeepFake detection. We formulate DeepFake detection as a graph classification problem to leverage relational information between facial regions, where vertex and edge respectively correspond to each region and the edge between two regions. As the relational information has large redundancy, we propose a masked relation learning framework to discard redundant relational information by masking partial edges. The framework uses a spatiotemporal attention module to extract features from multiple facial regions and exploits a masked relation learner to mask partial correlations between regions in the training procedure. Furthermore, the masked relation learner aggregates and propagates relational information across vertices to capture irregularity as clues to detecting fake faces. We observe that masking 50% minimal edges is favorable to learning global relationships among facial regions. Extensive experiments demonstrate that our method achieves superior performance and generalizability for DeepFake detection. We hope our work can boost the relation-aware DeepFake detection.

Acknowledgment

This work is based on MindSpore Framework. The authors sincerely thank the associate editor and the reviewers for their professional comments and suggestions.

References

[1] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–14.

[3] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering malicious DeepFakes: Survey, battleground, and horizon," *Int. J. Comput. Vis.*, vol. 130, no. 7, pp. 1678–1734, Jul. 2022.

[4] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5074–5083.

[5] (2016). *Fakeapp*. Accessed: Jul. 16, 2022. [Online]. Available: https://www.fakeapp.com/

[6] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7184–7193.

[7] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jan. 2019, pp. 38–45.

[8] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020.

[9] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, "OpenForensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10117–10127.

[10] H. Liu et al., "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 772–781.

[11] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5039–5049.

[12] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15044–15054.

[13] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other-audio-visual dissonance-based deepfake detection and localization," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 439–447.

[14] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.

[15] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[16] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting residual-based local descriptors as convolutional neural networks: An application to image forgery detection," in *Proc. 5th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2017, pp. 159–164.

[17] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proc. 4th ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2016, pp. 5–10.

[18] L. Li et al., "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5000–5009.

[19] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15003–15013.

[20] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2020, pp. 86–103.

[21] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6450–6458.

[22] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–21.

[23] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2185–2194.

[24] P. Wang et al., "ADT: Anti-deepfake transformer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1903–2899.

[25] X. Dong et al., "Protecting celebrities from DeepFake with identity consistency transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9468–9478.

[26] Q. Gu, S. Chen, T. Yao, Y. Chen, S. Ding, and R. Yi, "Exploiting fine-grained face forgery clues via progressive enhancement learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 735–743.

[27] W. Zhuang et al., "UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 391–407.

[28] K. Sun et al., "An information theoretic approach for attention-driven face forgery detection," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 111–127.

[29] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, May 2019, pp. 80–87.

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

[31] X. Li et al., "Sharp multiple instance learning for DeepFake video detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1864–1872.

[32] Z. Gu et al., "Spatiotemporal inconsistency learning for DeepFake video detection," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3473–3481.

[33] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, and L. Ma, "Delving into the local: Dynamic inconsistency learning for deepfake video detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1–9.

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

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

[36] Y. Rao, J. Ni, and H. Xie, "Multi-semantic CRF-based attention model for image forgery detection and localization," *Signal Process.*, vol. 183, Jun. 2021, Art. no. 108051.

[37] G. Wang, Q. Jiang, X. Jin, W. Li, and X. Cui, "MC-LCR: Multimodal contrastive classification by locally correlated representations for effective face forgery detection," *Knowl.-Based Syst.*, vol. 250, Aug. 2022, Art. no. 109114.

[38] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16000–16009.

[39] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.

[40] Z. Hou et al., "GraphMAE: Self-supervised masked graph autoencoders," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2022, pp. 594–604.

[41] L. Yang, F. Wu, Y. Wang, J. Gu, and Y. Guo, "Masked graph convolutional network," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4070–4077.

[42] F. Manessi and A. Rozza, "Graph-based neural network models with multiple self-supervised auxiliary tasks," *Pattern Recognit. Lett.*, vol. 148, pp. 15–21, Aug. 2021.

[43] W. Jin et al., "Self-supervised learning on graphs: Deep insights and new direction," 2020, *arXiv:2006.10141*.

[44] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[45] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3204–3213.

[46] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.

[47] F. Lugstein, S. Baier, G. Bachinger, and A. Uhl, "PRNU-based deepfake detection," in *Proc. ACM Workshop Inf. Hiding Multimedia Secur.*, Jun. 2021, pp. 7–12.

[48] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.

[49] S. Fernandes et al., "Predicting heart rate variations of deepfake videos using neural ODE," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1721–1729.

[50] J. Thies, M. Zollhöfer, C. Theobalt, M. Stamminger, and M. Nießner, "HeadOn: Real-time reenactment of human portrait videos," *ACM Trans. Graph.*, vol. 37, no. 4, p. 164, 2018.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[52] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[53] L. M. Binh and S. S. Woo, "ADD: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 122–130.

[54] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6458–6467.

[55] Z. Sun, Y. Han, Z. Hua, N. Ruan, and W. Jia, "Improving the efficiency and robustness of deepfakes detection through precise geometric features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3609–3618.

[56] Z. Gu, T. Yao, Y. Chen, S. Ding, and L. Ma, "Hierarchical contrastive inconsistency learning for deepfake video detection," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 596–613.

[57] M. Cho, T. Kim, I.-J. Kim, K. Lee, and S. Lee, "Relational deep feature learning for heterogeneous face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 376–388, 2021.

[58] H. Park and B. Ham, "Relation network for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11839–11847.

[59] J. Li et al., "Compositional temporal grounding with structured variational cross-graph correspondence learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3032–3041.

[60] A. Goel, B. Fernando, F. Keller, and H. Bilen, "Not all relations are equal: Mining informative labels for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15596–15606.

[61] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

[62] O. Mahmood, E. Mansimov, R. Bonneau, and K. Cho, "Masked graph modeling for molecule generation," *Nature Commun.*, vol. 12, no. 1, pp. 1–12, May 2021.

[63] J. Yu, T. Xu, Y. Rong, J. Huang, and R. He, "Structure-aware conditional variational auto-encoder for constrained molecule optimization," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108581.

[64] Y. Zhang and M. Rabbat, "A graph-CNN for 3D point cloud classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6279–6283.

[65] F. R. Bach, "Graph kernels between point clouds," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, 2008, pp. 25–32.

[66] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3464–3473.

[67] L. Zhao, Y. Song, C. Zhang, and Y. Liu, "T-GCN: A temporal graph convolutional network for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3848–3858, Sep. 2020.

[68] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[69] B.-N. Kang, Y. Kim, B. Jun, and D. Kim, "Attentional feature-pair relation networks for accurate face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5472–5481.

[70] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[71] A. Wettig, T. Gao, Z. Zhong, and D. Chen, "Should you mask 15% in masked language modeling?" 2022, *arXiv:2202.08005*.

[72] C. Feichtenhofer, H. Fan, Y. Li, and K. He, "Masked autoencoders as spatiotemporal learners," 2022, *arXiv:2205.09113*.

[73] A. Salehi and H. Davulcu, "Graph attention auto-encoders," in *Proc. IEEE 32nd Int. Conf. Tools With Artif. Intell. (ICTAI)*, Nov. 2020, pp. 989–996.

[74] G. Cui, J. Zhou, C. Yang, and Z. Liu, "Adaptive graph encoder for attributed graph embedding," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 976–985.

[75] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.

[76] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.

[77] R. Du et al., "Fine-grained visual classification via progressive multi-granularity training of jigsaw patches," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2020, pp. 153–168.

[78] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[79] J. Guan et al., "Delving into sequential patches for deepfake detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 15–29.

[80] T. Chen, Z. Zhang, Y. Cheng, A. Awadallah, and Z. Wang, "The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12020–12030.

[81] M. Lee, J. Lee, H. J. Jang, B. Kim, W. Chang, and K. Hwang, "Orthogonality constrained multi-head attention for keyword spotting," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 86–92.

[82] K. Ranasinghe, M. Naseer, M. Hayat, S. Khan, and F. S. Khan, "Orthogonal projection loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12333–12343.

[83] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Jun. 2014.

[84] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2387–2395.

[85] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, 2019.

[86] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874.

[87] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[88] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[89] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[90] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[91] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2382–2390.

[92] P. W. Battaglia et al., "Relational inductive biases, deep learning, and graph networks," 2018, *arXiv:1806.01261*.

[93] X. He, Q. Liu, and Y. Yang, "MV-GNN: Multi-view graph neural network for compression artifacts reduction," *IEEE Trans. Image Process.*, vol. 29, pp. 6829–6840, 2020.

[94] 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," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2019.

Ziming Yang received the B.E. degree in computer science and technology from Jinan University, Guangzhou, China, in 2020. He is currently pursuing the M.S. degree in computer application technology with the Institute of Information Engineering, Chinese Academy of Sciences, and the School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China. His research interests include biometrics, pattern recognition, and computer vision.

Jian Liang (Member, IEEE) received the B.E. degree in electronic information and technology from Xi'an Jiaotong University, in July 2013, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in January 2019. He was a Research Fellow with the National University of Singapore, from June 2019 to April 2021. He is currently an Associate Professor with the Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. His research interests focus on transfer learning, pattern recognition, and computer vision.

Yuting Xu received the B.E. degree in software engineering from the Taiyuan University of Technology, Shanxi, China, in 2018. She is currently pursuing the M.S. degree in electronic and information with the University of Chinese Academy of Sciences, Beijing, China, and the National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing. Her research interests include media forensics and machine learning.

Xiao-Yu Zhang (Senior Member, IEEE) received the B.S. degree in computer science from the Nanjing University of Science and Technology, Nanjing, China, in 2005, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010. He is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing. He has authored or coauthored more than 90 refereed publications in international journals and conferences. His research interests include artificial intelligence, data mining, and computer vision. He is a Distinguished Member of the CCF and a Senior Member of the ACM, CSIG, and CAAI. His awards and honors, include the Silver Prize of Microsoft Cup of the IEEE China Student Paper Contest in 2009, the Second Prize of Wu Wen-Jun AI Science and Technology Innovation Award in 2016, the CCCV Best Paper Nominate Award in 2017, the Third Prize of BAST Beijing Excellent S&T Paper Award in 2018, and the Second Prize of CSIG Science and Technology Award in 2019.

Ran He (Senior Member, IEEE) received the B.E. and M.S. degrees in computer science from the Dalian University of Technology, Dalian, China, in 2001 and 2004, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2009. He is currently a Full Professor with the Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences. His research interests include information theoretic learning, pattern recognition, and computer vision. He is an IAPR fellow. He serves on the program committee of several conferences. He also serves as an Associate Editor for *Pattern Recognition* and *Neurocomputing* (Elsevier).