Dynamic Graph Learning with Content-guided Spatial-Frequency Relation Reasoning for Deepfake Detection

Yuan Wang[1,2,4] Kun Yu[2] Chen Chen[1*] Xiyuan Hu[3] Silong Peng[1,4,5]
[1]Institute of Automation, Chinese Academy of Sciences [2]Alibaba Group
[3]School of Computer Science and Engineering, Nanjing University of Science and Technology
[4]University of Chinese Academy of Sciences [5]Beijing Visystem Co.Ltd
{wangyuan2020, chen.chen}@ia.ac.cn yukun.yk@alibaba-inc.com xiyuan.hu@foxmail.com

Abstract

With the springing up of face synthesis techniques, it is prominent in need to develop powerful face forgery detection methods due to security concerns. Some existing methods attempt to employ auxiliary frequency-aware information combined with CNN backbones to discover the forged clues. Due to the inadequate information interaction with image content, the extracted frequency features are thus spatially irrelavant, struggling to generalize well on increasingly realistic counterfeit types. To address this issue, we propose a Spatial-Frequency Dynamic Graph method to exploit the relation-aware features in spatial and frequency domains via dynamic graph learning. To this end, we introduce three well-designed components: 1) Content-guided Adaptive Frequency Extraction module to mine the content-adaptive forged frequency clues. 2) Multiple Domains Attention Map Learning module to enrich the spatial-frequency contextual features with multiscale attention maps. 3) Dynamic Graph Spatial-Frequency Feature Fusion Network to explore the high-order relation of spatial and frequency features. Extensive experiments on several benchmark show that our proposed method sustainedly exceeds the state-of-the-arts by a considerable margin.

1. Introduction

Recent years have witnessed the continuous advances in deepfake creation [11, 27, 36]. Utilizing booming open-source tools such as Deepfakes [41], novices can readily manipulate the expression and identity of faces to generate visually untraceable videos. Face forgery technology has stimulated many applications [12, 14, 44, 46] with wide acceptance. These techniques can whereas be abused by malicious intentions to make pornographic movies, fake news and political rumors. In the circumstances, it is desperately in need to develop powerful forgery detection methods.

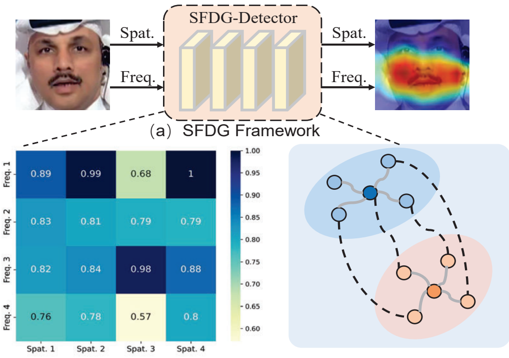

Figure 1. The motivation of our proposed approach. Our SFDG method (a) delves into the high-order relationships (b) of spatial and frequency domain via cross-domain graph reasoning (c).

Early face forgery detection methods [7, 30, 52] treat this challenge as vanilla dichotomy tasks in the prevailing view. They use off-the-shelf backbones to extract the global feature of faces and a binary classifier follow-up to identify the real and counterfeit faces. However, as the counterfeits become increasingly realistic, it is intractable for these methods to spot subtle and local forgery traces. One recent study [50] reformulates deepfake detection as a fine-grained classification task and designs a multi-attentional framework to extract local discriminative features from multiple attention maps. It is susceptible to common disturbances and the generalized features remain therefore poorly understood. Some other works resort to specific forgery patterns to encourage better classification such as DCT [24, 28], SRM [25] and steganalysis features [45]. Although promising advances have been achieved by these previous works,

*Chen Chen is the corresponding author. This work is supported by the National Key R&D Program of China under Grant 2021YFF0602101, Alibaba Group through Alibaba Research Intern Program.

they always extract frequency features with hand-crafted filter banks which are content-irrelevant, thus incapable to adapt the changes of complex scenarios. Moreover, they fuse multi-domain information via adding directly or attentional projection. However, these approaches devote little efforts to discover the high-order relation of spatial and frequency features and integrating them in a reasonable way.

In this paper, we provide seminal insights to exploit adaptive frequency features and delve into the interactions of spatial-frequency domains. To this end, we propose an *adaptive extraction-multiscale enhancement-graph fusion* paradigm for deepfake detection via dynamic graph learning, which prompts to excavate content-aware frequency clues and the high-order relation of multiple domains.

Firstly, for adaptive frequency extraction, we tailor a Content-guided Adaptive Frequency Extraction (CAFÉ) module with coarse-grained DCT and fine-grained DCT to capture the local frequency cues guided by content-aware masks. Different from PEL [9] and F^3Net [28], our customized frequency learning protocol provides potential for more combinations of frequency features, which is indispensable to spot complicated counterfeit patterns.

To further enhance the representation of content-guided frequency features, we introduce a Multiple Domains Attention Map Learning (MDAML) module to generate multiscale spatial and frequency attention maps with high-level semantic features. Specifically, we first propose a Multi-Scale Attention Ensemble (MSAE) module, which produces multi-scale semantic attention maps with large receptive fields and endows rich contextual information to spatial and frequency domains. Moreover, an Attention Map Refinement Block (AMRB) is included in the MSAE module to refine the obtained semantic attention maps conducive to the following feature learning. In comparison with MADD [50] that merely emphasizes the spatial domain, we further introduce the semantic-relevant frequency attention map with rich semantic information retained for the subsequent spatial-frequency relation-discovery paradigm.

Finally, to fully discover the spatial and frequency relationships, we propose a Dynamic Graph-based Spatial-Frequency Feature Fusion network (DG-SF^3Net) to formulate the interaction of two domains via a graph-based relation discovery protocol. Specifically, DG-SF^3Net is composed of two ingredients: Dynamic GCN [16] and Graph Information Interaction layers. The former constructs kNN-graph dynamically and performs graph convolution to reason high-order relationships in spatial and frequency domains, while the latter is designed to enhance the mutual relation with several graph-weighted MLP-Mixer [40] layers via channel-wise and node-wise interaction.

The achievements, including contributions are threefold:

- From a new perspective, we propose a novel Spatial-Frequency Dynamic Graph (SFDG) framework, which is qualified to exploit relation-aware spatial-frequency features to promote generalized forgery detection.
- We first harness a CAFÉ module for content-aware frequency feature extraction, and then tailor an MDAML scheme to dig deeper into multiscale spatial-frequency attention maps with rich contextual understanding of forgeries. Finally, a seminal DG-SF^3Net module is proposed to discover the multi-domain relationships with a graph-based relation-reasoning approach.
- Our method achieves state-of-the-art performance on six benchmark datasets. The cross-dataset experiment and the perturbation analysis show the robustness and generalization ability of the proposed SFDG method.

2. Related Works

Over the past several years, with the remarkable progress of forgery creation techniques, increasing efforts have been made to boost the development of face forgery detection in computer vision communities. In this section, we briefly review previous works exploring the authenticity of faces.

Rudimentary deepfake detection approaches base primarily on obvious counterfeit artifacts, which utilize hand-crafted features to detect anomalies in spatial domain, *e.g.*, inconsistent head pose [47] and unnatural eye blinding [21]. However, these traditional methods feel intractable to deal with the improved realistic deepfakes. With the overwhelming success of deep learning, some works [1,4,7] adopt off-the-shelf classification backbones to extract high-level semantic features for deepfake detection. Although considerable performances have been achieved on specific datasets, their vanilla structures would lead to catastrophic overfitting and lag behind the advanced face synthesis technology.

To further exploit the essential forged clues, general deepfake detection [15, 20, 26, 34, 35] has been an area of intense investigation. Nguyen et al. [26] employ multi-task learning strategy to detect forged artifacts and locate manipulated regions simultaneously. Face X-ray [20] observes the specific blending step of face swapping and locates forgery boundary in a self-supervised manner. FReTAL [15] adopts the knowledge distillation to prevent catastrophic forgetting and enhance adaptability in different domains. DCL [35] performs dual-granularity contrastive learning to further improve the generalization. However, these approaches captures category-level differences instead of the intrinsic discrepancies between authentic and forged images.

Recently, other studies focus on mining specific forgery patterns, such as local texture, high-frequency noise, reconstruction residual and frequency clues. Among them, Zhao et al. [50] propose a fine-grained deepfake detection framework that aggregates the local texture and high-level semantic information into multiple attention maps. However, it fails to distinguish highly compressed videos with

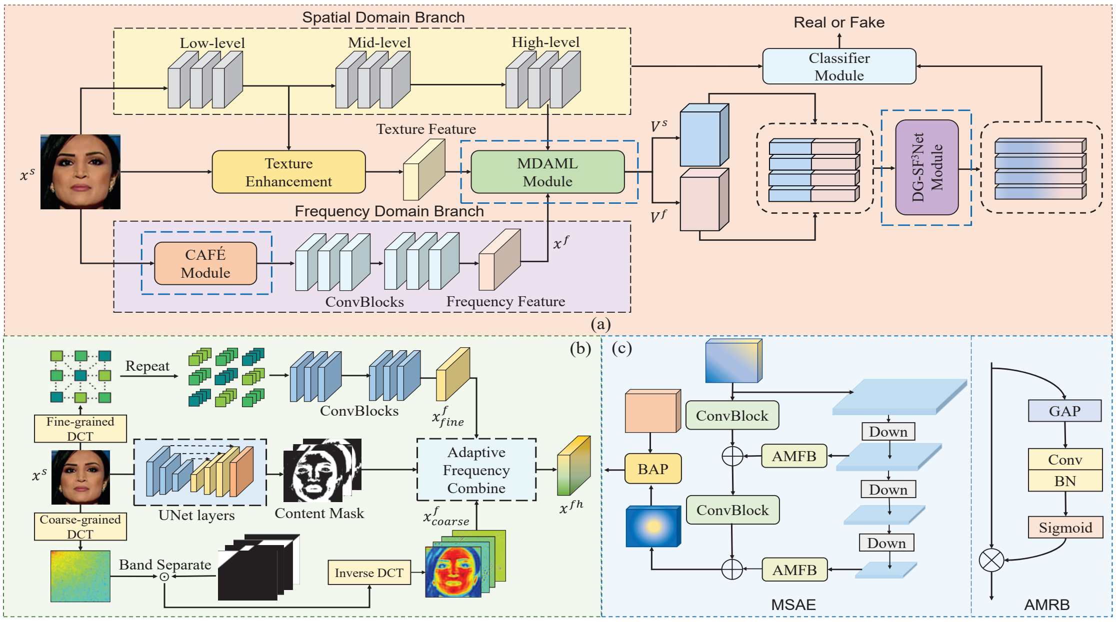

Figure 2. (a) Overview of the proposed Spatial-Frequency Dynamic Graph (SFDG) framework. (b) The Content-guided Adaptive Frequency Extraction (CAFÉ) module. (c) The Multiple Domains Attention map Learning (MDAML) network.

blurry texture. Luo et al. [25] utilize the high-pass filter SRM [8] to extract high-frequency noise guiding face forgery detection. RECCE [2] employs the reconstruction difference of authentic faces as guidance of forgery traces. F^{3}Net [28] collaboratively mines the forged frequency clues with frequency-aware decomposition and local frequency statistics. However, the extracted frequency feature is coarse-grained, incapable of assembling discriminative feature patterns. To ameliorate this issue, PEL [9] learns the fine-grained frequency features using sliding window DCT combined with RGB images in a progressive enhancement learning fashion. Nevertheless, the fine-grained frequency is content-irrelevant and in compliance with the vanilla spatial-frequency fusion ethos, thus failing to achieve the utter information interaction of both domains.

Different from existing studies, our proposed end-to-end SFDG approach excavates content-adaptive frequency clues and facilitates the comprehensive fusion of spatial and frequency features via dynamic graph learning.

3. Method

In this section, we propose a Spatial-Frequency Dynamic Graph (SFDG) framework, which consists of three major modules, *i.e.*, Content-guided Adaptive Frequency Extraction (CAFÉ), Multiple Domains Attention Map Learning (MDAML), and Dynamic Graph Spatial Frequency Feature Fusion Network (DG-SF^{3}Net). The primary idea of our approach is to exploit content-aware frequency features utilizing CAFÉ module. As limited context information do these features embrace, we present the MDAML module to generate high-level frequency attention maps in a multiscale learning manner. Beyond these improvements, we finally design a DG-F^{3}Net to perform relation reasoning of spatial and frequency features via dynamic graph learning.

3.1. Content-aware Frequency Extraction

Towards the frequency-aware face forgery detection, former studies generally use DCT to transform the input image into frequency domain. However, for the scarcity of spatial information in DCT process, the extracted frequency feature is presumably not compatible to the image content. To this end, we propose the CAFÉ module to fully exploit the forgery clues via content-aware frequency learning in order to handle complex or simple manipulated patterns adaptively. Without loss of generality, let $x^s \in \mathbb{R}^{3\times H\times W}$ denotes the RGB image, where H and W are the height and width of the input image. As shown in Fig. 2(b), we generate an attentional content mask M^s using a UNet [29] submodule with the spatial information retained. We then propose a coarse-grained counterpart to partition the compact frequency spectrum into several bands. Specifically, we manually design N_f binary filters $\{f_i|i = 1, ..., N_f\}$ to decompose the frequency domains into low, middle and high frequency bands. The low-frequency component lies on the top-left corner while the high-frequency response locates on the bottom-right corner [28]. As described, the coarse-grained process can be formulated as:

$$x_{coarse}^{f,i} = \mathcal{H}^{-1}\left[\mathcal{H}\left(x^s\right) \odot f^i\right], \quad i = \{1, 2, \ldots, N_f\}, \quad (1)$$

where $\odot$ is the Hadamard product. $\mathcal{H}$ and $\mathcal{H}^{-1}$ represent the DCT and inverse DCT respectively. $x_{coarse}^{f} \in \mathbb{R}^{3N_f \times H \times W}$ is the extracted coarse frequency feature, which indicates global frequency information to some extent.

Further, we design a fine-grained frequency extraction method to emphasize on the localized frequency features. Specifically, we slice x^s via a sliding window to obtain a set of $l \times l$ patches. $p_{m,n} \in \mathbb{R}^{3 \times l \times l}$ denotes the patch sliced by the (m-th, n-th) sliding window. After performing DCT on each patch, we get the localized fine-grained frequency features $d_{m,n}^f \in \mathbb{R}^{3 \times l \times l}$. We repeat the channel of each $d_{m,n}^f$ to match the coarse-grained frequency features abovementioned. All localized patched are then gathered into a whole feature map $d^f \in \mathbb{R}^{3N_f \times H \times W}$ and go through several convolution Blocks composed by Conv2d, BN, and ReLU to form the output fine-grained frequency features x_{fine}^f.

Finally, we obtain the overall frequency head x^{fh} through an adaptive frequency combine module guided by the learned content mask M^s, which can be formulated as:

$$x^{fh} = (1 - M^s) \odot x_{fine}^f + M^s \odot x_{coarse}^f. \quad (2)$$

The extracted frequency head $x^{fh} \in \mathbb{R}^{3N_f \times H \times W}$ will be fed into several stride convolution layers to obtain the final frequency feature C_f channels x^f with size of $H_f \times W_f$.

3.2. Multiple Domains Attention Map Learning

To fully exploit the forged clues in spatial domain, we adopt EfficientNet-b4 as the backbone and further partition the entire network into low-level, mid-level and high-level layers, respectively. As shown in Fig. 2(c), to acquire spatial-frequency features with rich semantics, we further design a Multi-Scale Attention Ensemble (MSAE) module in spatial and frequency domains to generate multiscale attention maps and aggregate them with hierarchical feature pyramid. The MSAE includes light weighted ConvBlocks, which consist of several convolution layer, BN and non-linear activation layers ReLU, and certain global average layers to downsample the attention maps into multiple scales. With the multi-scale feature representation, we can obtain a sufficient receptive field and rich contextual information, which are of great significance for the deepfake detection task. Inspired by [48], we propose an Attention Map Refinement Block (AMRB) to get refined feature maps.

As Fig. 2(c) shows, AMRB employs global average pooling to capture global context and uses a subsequent sigmoid layer to produce an attention vector that determines the feature distribution of attention maps in each scale. Finally, we upsample the refined multi-scale attention maps to the original size and add them spatially. We use AMRB in both spatial and frequency streamline to acquire the multi-scale spatial attention map F^s and frequency feature map F^f, which have T_s and T_f channels on behalf of the corresponding number of attention maps in these two domains.

After obtaining the refined multi-scale attention maps F^s and F^f, we perform Bilinear Attention Pooling (BAP) [23] to get the relation-aware feature matrix P^s and P^f respectively. As illustrated in Fig. 2(c), we use the Textural Feature Enhancement block [50] to capture the manipulated artifacts hidden in low-level layers. After the enhancement of the texture features, we get a textual feature map $F^{tex} \in \mathbb{R}^{C_{tex} \times H_{tex} \times W_{tex}}$. For the spatial domain, we element-wisely multiply textural feature map F^{tex} by each spatial attention map F_k^s to obtain the partial textural feature maps F_k^{tex} and compute the final spatial feature matrix employing BAP, which can be formulated as:

$$p_k^s = \frac{\sum_{m=1}^{H_{tex}} \sum_{n=1}^{W_{tex}} F_{k,m,n}^{tex}}{\left\| \sum_{m=1}^{H_{tex}} \sum_{n=1}^{W_{tex}} F_{k,m,n}^{tex} \right\|_2} \quad (3)$$

The spatial attention vector $p_k^s \in \mathbb{R}^{1 \times C_{tex}}$ is stacked together to get the texture-relevant spatial feature matrix $P^s \in \mathbb{R}^{T_s \times N_s}$. For the frequency domain, we perform the BAP with the frequency attention maps F^f and the spatial feature after high-level backbone layers as shown in Fig. 2(a). Similar to Eq. 3, we can obtain the content-relevant frequency feature matrix $P^f \in \mathbb{R}^{T_f \times N_f}$. Finally, the learnt relation-aware feature matrices will be fed into the follow-up DG-SF^{3}Net to collaboratively learn a comprehensive feature representation in a graph-based relation reasoning way.

3.3. Dynamic Graph Learning

After obtaining the spatial feature matrix $P^s \in \mathbb{R}^{T_s \times N_s}$ and frequency feature matrix $P^f \in \mathbb{R}^{T_f \times N_f}$, we propose a novel DG-SF^{3}Net to discover the comprehensive relation of them as shown in Fig. 3. Inspired by [19, 49, 51], we provide the intuition that there exists high-order relation of spatial and frequency features and Graph Convolution Network (GCN) performs tremendous potential in relation reasoning. As illustrated in Fig. 1(c), we confirm the above intuition by visualizing the corelation heatmap of spatial and frequency features. Therefore, we build a graph-based relation-discovery module to conduct reasoning with improved GCN to integrate spatial and frequency features.

Firstly, we concatenate P^s and P^f in the channel dimension to get $V^{(0)} \in \mathbb{R}^{T^{(0)} \times N^{(0)}}$, where $T^{(0)} = T_s + T_f$ and $N^{(0)} = max(N_s, N_f)$. We treat each column of $V^{(0)}$ as a node. As shown in the right bottom of Fig. 3, a dynamic GCN module is proposed to aggregate features based on their similarity. Different from original GCN, our model learns to dynamically design appropriate graph structures in each layer, rather than treating it as a fixed constant one.

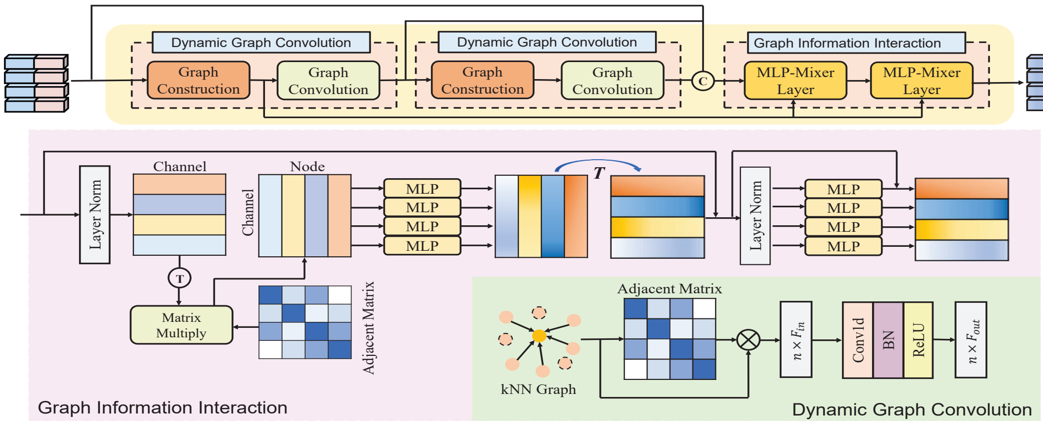

Figure 3. The proposed Dynamic Graph Spatial-Frequency Feature Fusion Network (DG-SF^{3}Net), which includes two modules: Dynamic Graph Convolution Network (DGCN) and Graph Information Interaction (GI^2) layers.

At t-th layer, we have a tailored sparse-connected graph $G^{(t)} = (V^{(t)}, E^{(t)})$ for graph convolution, where the local neighbor of each center node $v_i^{(t)}$ can be defined as:

$$\mathcal{N}^{(t)}(i) = \left\{ v_{jim}^{(t)} \mid v_{jim}^{(t)} \in kNN\left(v_i^{(t)}\right), m = 1, \ldots, k \right\},$$

where $v_{ji1}^{(t)}, \ldots, v_{jik}^{(t)}$ are k nodes closest to $v_i^{(t)}$ according to the Euclidean distance. After getting the localized kNN graph, we formulate the adjacency matrix $A^{(t)}$ based on k neighbors attached to center node with self-loop added.

With this sparse-connected GCN, the node features can be updated with the message passing as follows:

$$V^{(t+1)} = \mathrm{ReLU}\left(\tilde{D}^{(t)-\frac{1}{2}} \tilde{A}^{(t)} \tilde{D}^{(t)-\frac{1}{2}} V^{(t)} W^{(t)}\right), \quad (4)$$

where $\tilde{D}^{(t)} \in \mathbb{R}^{T^{(t)} \times T^{(t)}}$ is the degree matrix. $W^{(t)} \in \mathbb{R}^{N^{(t)} \times N^{(t+1)}}$ is the learnable graph weight. In addition, the output channel is conforming with the input one.

Finally, we propose a Graph Information Interaction (GI^2) component for the spatial-frequency feature interaction. The GI^2 component is composed of several MLP-Mixers [40] layers weighted by the graph adjacent matrix. Each of them leverages a channel-mixing MLP layer and a token-mixing MLP layer, which perform information interaction across all channels and node positions embedded in high-dimensional feature space. These two types of layers are interleaved to enable interaction of both input dimensions. Particularly, in order to highlight the localized features defined on the kNN graph abovementioned, we utilize the adjacent matrix derived from the first GCN layer to weight the feature of the token-mixing MLP.

3.4. Loss Function

After dynamic graph learning, we obtain the semantic feature matrix $\tilde{V} \in \mathbb{R}^{M \times N}$. Following MADD, we regard Region Independent Loss (RIL) with Cross Entropy (CE) loss as the loss function. The auxiliary RIL is defined as:

$$\begin{aligned} \mathcal{L}_{RIL} = &\sum_{i=1}^{B} \sum_{j=1}^{M} \mathrm{ReLU}\left(\left\|\tilde{V}_j^i - c_j^t\right\|_2^2 - m_{in}(y_i)\right) \\ &+ \sum_{i \neq j}^{M} \mathrm{ReLU}\left(m_{out} - \left\|c_i^t - c_j^t\right\|_2^2\right) \end{aligned} \quad (5)$$

$c \in \mathbb{R}^{M \times N}$ are feature centers of $\tilde{V}$ that can be updated iteratively during training. B is the batch size, M is the number of attention maps and y_i is the label. m_{in} and m_{out} represent the margin of intra-class and inter-class.

In conclusion, the total loss function $\mathcal{L}$ of our proposed SFDG framework can be described as:

$$\mathcal{L} = \lambda_1 * \mathcal{L}_{CE} + \lambda_2 * \mathcal{L}_{RIL} \quad (6)$$

where λ_1, λ_2 are hyper-parameters for balancing different terms. We set $\lambda_1 = \lambda_2 = 1$ in the following experiments.

4. Experiment

4.1. Experimental Setup

Datasets. We appraise our proposed SFDG method and current state-of-the-art approaches on FaceForensics++ (FF++) [30], WildDeepfake [52], DFDC [5], Celeb-DF [22], DF-v1.0 [13] and DFD [6]. FF++ (low quality (LQ) and high quality (HQ) counterparts) is a widely-used benchmark dataset composed of 1000 real videos from YouTube and corresponding fake videos generated by four types of manipulated techniques: Deepfakes (DF) [41], Face2Face (F2F) [39], FaceSwap (FS) [17] and Neural-Textures (NT) [38]. WildDeepfake is a small real-scenario dataset including 7314 face sequences. Celeb-DF contains

Method	FF++ (LQ)		FF++ (HQ)		WildDeepfake		Celeb-DF	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
Xception [4]	86.86	89.30	95.73	96.30	79.99	88.86	97.90	99.73
EfficientNet-b4 [37]	86.67	88.20	96.63	99.18	82.33	90.12	98.19	99.83
Add-Net [52]	87.50	91.01	96.78	97.74	76.25	86.17	96.93	99.55
SCL [18]	89.00	92.40	96.69	99.30	——	——	——	——
MADD [50]	88.69	90.40	97.60	99.29	82.62	90.71	97.92	99.94
F3Net [28]	90.43	93.30	97.52	98.10	80.66	87.53	95.95	98.93
PEL [9]	90.52	94.28	97.63	99.32	**84.14**	91.62	——	——
RECCE [2]	91.03	95.02	97.06	99.32	83.25	92.02	98.59	99.94
Local Relation [3]	91.47	**95.21**	97.59	99.46	——	——	——	——
M2TR [43]	**92.35**	94.22	**98.23**	**99.48**	——	——	——	——
SFDG (Xception)	91.08	94.49	97.61	99.45	83.36	**92.15**	**98.95**	**99.94**
SFDG (Ours)	**92.28**	**95.98**	**98.19**	**99.53**	**84.41**	**92.57**	**99.22**	**99.96**

Table 1. Quantitative comparison in terms of Acc(%) and AUC(%) on FF++, WildDeepfake and Celeb-DF dataset. The red represents the best performance while the blue indicates the second-best results.

590 real videos and 5,639 high quality fake videos tampered with the improved deepfake methods. DFDC is a remarkable large-scale dataset with over 100,000 video clips. DFD and DF-v1.0 are two another large-scale publicly available dateset to evaluate the model's generalization ability.

Evaluation Metrics. Following previous works [9, 28], we employ the commonly used Accuracy (Acc), Area Under the Receiver Operating Characteristic Curve (AUC) and Equal Error Rate (EER) as our evaluation metrics.

Implementation Details. Following the official split of FF++ dataset, we use 740 videos for training, 140 for validation and 140 remained for testing. We sample 50 frames per video at equal interval during training and validation phases, 20 frames in testing period. We use the remarkable DLIB [32] for face detection and alignment. We employ EfficientNet-b4 [37] pretrained on ImageNet [31] as the backbone of our network. We set m_{out} to 0.2 and m_{in} to [0.05,0.1] in Eq. 5. The whole network is trained with Adam optimizer with an initial learning rate of 1×10^{-4}, the weight decay of 1×10^{-5}. A step learning scheduler is used to reduce the learning rate half every 5 epochs.

4.2. Experimental Results

Intra-testing. We compare our proposed method versus current state-of-the art approaches on the FF++ dataset under different quality settings (HQ and LQ), WildDeepfake and Celeb-DF datasets. As shown in Table 1, our method consistently achieves admirably performance on all quality settings and trumps all reference methods by a considerable margin. Specifically, in terms of Acc, our method achieves 92.28% and 98.19% on LQ and HQ settings respectively, with a remarkable improvement in comparison to PEL [9], *i.e.*, 2.1% performance gain on LQ and 0.6% on HQ. Different from PEL which only utilizes fine-grained frequency features, our model extracts adaptive frequency features with a content mask providing rounded frequency representations. More importantly, we boost the information interaction between spatial and frequency domains via dynamic graph learning instead of vanilla convolution paradigm in PEL. Further, our method achieves a noteworthy performance compared with Local Relation [3] and M2TR [43], which introduce powerful priors of forgery masks. Instead, our SFDG only employs real/fake images as the input, but the AUC on both LQ and HQ settings surpass these two works strikingly. As discussed, these fruitful results give explanation of the effectiveness of our SFDG method.

Cross-testing. To further appraise the generalization ability of our method on unseen manipulated types, we herein conduct cross-dataset experiments by training and testing on different datasets. Following PEL [9], we reimplement several state-of-the-art models for a fair comparison on FF++ (LQ) dataset and testing them on Celeb-DF, DFDC, DFD, DF-v1.0 and WildDeepfake dataset. Comparisons under AUC and EER metrics are detailedly shown in Table 2. It implicates that our SFDG method generally outperforms all competitors conspicuously on all testing datasets. Instead of overfitting with specific forged patterns as in most existing methods, SFDG explores the essential forgery with content-aware semantic attention maps and reasons about generalized forged cues via graph-based high-order relation discovery in spatial and frequency domains, which guarantees the superior generalization ability of our proposed method.

Robustness. Considering the image quality will be deteriorated seriously by inevitable noise in video acquisition process, we further investigate the robustness of our SFDG model under several common perturbations. Specifically,

Training DataSet	Method	Testing Dataset									
		Celeb-DF		DFD		WildDeepfake		DFDC		DF-v1.0	
		AUC	EER↓	AUC	EER↓	AUC	EER↓	AUC	EER↓	AUC	EER↓
FF++	Xception [4]	60.05	0.432	65.43	0.393	60.59	0.619	55.65	0.461	80.27	0.265
	Ef-b4 [37]	64.29	0.419	83.17	0.235	64.27	0.376	60.12	0.428	85.31	0.228
	Add-Net [52]	57.83	0.444	57.16	0.453	54.21	0.462	51.60	0.548	——	——
	MADD [50]	68.64	0.371	74.18	0.327	65.65	0.397	63.02	0.410	89.34	0.173
	F3Net [28]	67.95	0.368	69.50	0.354	60.49	0.434	57.87	0.442	82.27	0.246
	PEL [9]	69.18	0.357	75.86	0.308	67.39	0.383	63.31	0.404	——	——
	SFDG (Ours)	**75.83**	**0.303**	**88.00**	**0.197**	**69.27**	**0.377**	**73.64**	**0.337**	**92.10**	**0.151**

Table 2. Cross-testing results in terms of AUC(%) and EER training on FF++ dataset. The bold indicates the best performance.

Method	+GaussianNoise		+SaltPepperNoise		+GaussianBlur	
	ΔAcc(FF)	ΔAcc(Wild)	ΔAcc(FF)	ΔAcc(Wild)	ΔAcc(FF)	ΔAcc(Wild)
Xception [4]	-2.65%	-0.98%	-32.44%	-27.80%	-6.22%	-12.71%
Add-Net [52]	-41.51%	-11.66%	-11.28%	-18.21%	-11.28%	-12.91%
F3Net [28]	-9.86%	-1.17%	-31.08%	-43.57%	-11.08%	-12.43%
MADD [50]	-1.79%	-0.99%	-49.30%	-29.47%	-12.23%	-14.86%
PEL [9]	-0.10%	-0.86%	**-9.39%**	-4.25%	-7.41%	-10.88%
SFDG (Ours)	**-0.10%**	**-0.75%**	-10.10%	**-3.74%**	**-3.76%**	**-5.12%**

Table 3. Robustness evaluation under three types of perturbations. Our SFDG performs admirably under several common perturbations.

following PEL [9], we apply GaussianNoise, SaltPepperNoise and GaussianBlur to the FF++ (LQ) and WildDeepfake testing images, and adopt the decay of Acc to indicate the robustness of face forgery detection model. As shown in Table 3, a series of experiments throw light on that our method with the least performance decline is more robust to the perturbations abovementioned. We attribute it to the elaborate CAFÉ module which extracts the adaptive frequency features among multiple bands and acts as an image denoiser. We provide sufficient experiments results in terms of AUC in the supplementary materials.

4.3. Ablation Study

Effectiveness of Proposed Components. As shown in Table. 4, a series of ablation experiments on FF++ (LQ) benchmark have been conducted to verify the effectiveness of different components in our framework. Specifically, we develop the following variants: (a) the baseline model equipped with the MADD [50] pipeline, (b) the baseline model with the proposed CAFÉ module, (c) the proposed method w/o. DG-SF^{3}Net, (d) the proposed method w/o. MDAML module. Comparing (a) and (b), we observe that the proposed CAFÉ module brings a considerable improvement of Acc and AUC metrics on FF++ (LQ) dataset. From variant (b) and (c), we empirically demonstrate that adding the MDAML module will bring 2.19% Acc and 1.52% AUC gains, which are attributed to the multi-scale contextual information and large receptive field of refined attention maps. When adding the DG-SF^{3}Net module, as

ID	CAFÉ	MDAML	DG -SF^{3}Net	Acc (LQ)	AUC (LQ)
(a)				88.69	90.4
(b)	√			89.52	93.37
(c)	√	√		91.48	94.79
(d)	√		√	92.05	95.45
Ours	√	√	√	**92.28**	**95.98**

Table 4. Ablation study on the FF++ (LQ) dataset.

shown in variant (c) and Ours, we observe a remarkable advance on both Acc and AUC metrics, which benefits from the dynamic graph learning protocol to reason about the relation-aware forged clues. Finally, the best performance is achieved when combining all components, with the Acc and AUC of 92.28% and 95.98% respectively.

Setting of Hyper-parameters. During graph construction, the number of neighbor nodes k is a hyperparameter which determines the structure and aggregated scope of GCNs [10]. As shown in Table 5, to further evaluate the effectiveness using different k, we conduct a series of experiments on FF++ (LQ) and WildDeepfake datasets. From the experimental results, we conclude that too few neighbors can not guarantee the representative ability of the dynamic GCN module, thus degrades the information exchange of forged clues in spatial and frequency domains. Further, too many neighbors will tamper the regional separability of attention maps and tend to catastrophic overfitting. In our

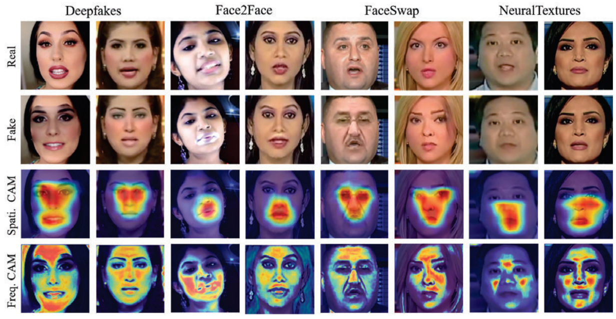

Figure 4. The Grad-CAM visualization for forged faces.

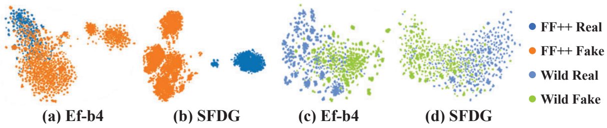

Figure 5. The t-SNE feature visualization of the Ef-b4 and SFDG.

#Neighbors	FF++ (LQ)		WildDeepfake	
	Acc	AUC	Acc	AUC
5	92.10	95.60	82.03	90.57
10	**92.28**	**95.98**	**84.41**	**92.57**
15	92.28	95.49	83.57	91.15
20	92.02	91.59	81.46	90.38

Table 5. Results in terms of Acc(%) and AUC(%) for different number of neighboring nodes in the graph construction.

work, we observe that our model achieves the best performance in terms of Acc and AUC when setting k to 10.

4.4. Visualization

Grad-CAM visualization. To better understand the internal mechanism of our method and explore interested regions for specific forged types, we supply the Grad-CAM [33] visualization on FF++ dataset. As shown in Fig. 4, the forgery artifacts in different domains locate on independent regions. Detailedly, the spatial branch focuses on the facial region with pronounced forgery traces, which are effortlessly discovered by CNN backbones in spatial domain. Conversely, the frequency streamline searches manipulated clues essentially concerned with a wider area, *e.g.*, the background, hair or entire face. These regions are subtle and dynamically change according to the image content thus provides a generalized feature representation. Just as we have speculated, the frequency domain acts as the complementary to the spatial domain and these two domains contribute to each other via graph-based high-order relation discovery for exploiting comprehensive forged cues.

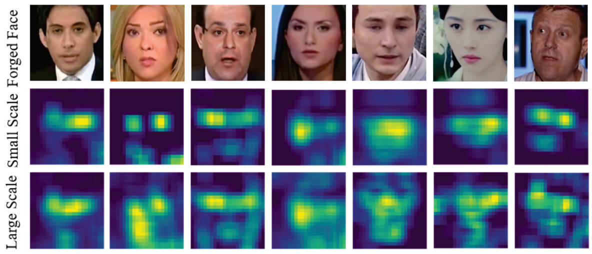

Figure 6. The visualization results of feature maps from MDAML module at different scales on FF++ and WildDeepfake datasets.

Attention Maps Visualization. We verify the effectiveness of our tailored MDAML module and visualize the results in Fig 6. We observe that feature maps with different scales highlight distinctive activated intensities. Detailedly, the large scale features with high resolution representation embrace richer and global manipulated traces, while the small scales concentrate on more localized salient feature around facial landmarks. Further, we aggregate the attention maps via hierarchical pyramid paradigm to exploit the essential discrepancy between authentic and counterfeit faces, which can resist the noise disturbance and visual compression.

Feature Distribution Visualization. In this part, we investigate the discriminative ability of the proposed SFDG model. Leveraging the t-SNE [42] technique, we visualize the semantic feature distribution of the Ef-b4 [37] model and our SFDG on FF++ (LQ) and WildDeepfake dataset. As shown in Fig. 5, our approach encourages the samples of same class into a relatively compact feature space. To explain, our SFDG adequately captures the intrinsic discrepancy between genuine and forged faces in multiple domains through graph-based relationships discovery, thus improving the generalization ability of our method.

5. Conclusion

In this paper, we propose a novel Spatial-Frequency Dynamic Graph network that develops graph model to exploit relationships of spatial and frequency domains for spotting subtle forgery clues. Firstly, the Content-guided Adaptive Frequency Extraction module is proposed to mine the adaptive frequency clues via content-aware frequency learning. Further, a Multiple Domains Attention Map Learning scheme captures rich contextual information of spatial-frequency feature through multi-scale feature ensemble. Finally, Dynamic Graph based Spatial-Frequency Feature Fusion module performs relation reasoning of spatial and frequency domains via improved graph convolution. Extensive experiments and detailed visualizations on widely-used benchmarks confirm the effectiveness and generalizability of our SFDG method compared with other contenders.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security*, pages 1–7, 2018. 2

[2] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022. 3, 6

[3] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1081–1088, 2021. 6

[4] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017. 2, 6, 7

[5] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019. 5

[6] Nick Dufour and Andrew Gully. Contributing data to deepfake detection research. *Google AI Blog*, 1(3), 2019. 5

[7] Nguyen et al. Capsule-forensics: Using capsule networks to detect forged images and videos. In *2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311, 2019. 1, 2

[8] Jessica J. Fridrich and Jan Kodovský. Rich models for steganalysis of digital images. *IEEE Trans. Inf. Forensics Secur.*, 7(3):868–882, 2012. 3

[9] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 735–743, 2022. 2, 3, 6, 7

[10] Kai Han, Yunhe Wang, Jianyuan Guo, Yehui Tang, and Enhua Wu. Vision gnn: An image is worth graph of nodes. *arXiv preprint arXiv:2206.00272*, 2022. 7

[11] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4360–4369, 2021. 1

[12] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3397–3406, 2022. 1

[13] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2889–2898, 2020. 5

[14] Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, and Youngjung Uh. Exploiting spatial dimensions of latent in GAN for real-time image editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 852–861, 2021. 1

[15] Minha Kim, Shahroz Tariq, and Simon S Woo. Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1001–1012, 2021. 2

[16] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 2

[17] M Kowalski. Faceswap. https://github.com/marekkowalski/faceswap. Accessed: 2020-08-01, 2018. 5

[18] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6458–6467, 2021. 6

[19] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4654–4662, 2019. 4

[20] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020. 2

[21] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security*, pages 1–7, 2018. 2

[22] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020. 5

[23] Lin and Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015. 4

[24] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 772–781, 2021. 1

[25] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16317–16326, 2021. 1, 3

[26] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems*, pages 1–8, 2019. 2

[27] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, et al. Deepfacelab: A simple, flexible and extensible face swapping framework. 2020. 1

[28] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of the IEEE Conference on European Conference on Computer Vision*, pages 86–103. Springer, 2020. 1, 2, 3, 6, 7

[29] Ronneberger. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015. 3

[30] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11, 2019. 1, 5

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 6

[32] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016. 6

[33] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *arXiv e-prints*, 2016. 8

[34] Ke Sun, Hong Liu, Qixiang Ye, Yue Gao, Jianzhuang Liu, Ling Shao, and Rongrong Ji. Domain general face forgery detection by learning to weight. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2638–2646, 2021. 2

[35] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2316–2324, 2022. 2

[36] Supasorn Suwajanakorn and Seitz. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017. 1

[37] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 6, 7, 8

[38] Thies. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 5

[39] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 5

[40] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021. 2, 5

[41] M Tora. Deepfakes, 2018. https://github.com/deepfakes/faceswap/tree/v2.0.0. Accessed: 2021-03-29. 1, 5

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008. 8

[43] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 615–623, 2022. 6

[44] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European Conference on Computer Vision*, pages 603–619, 2018. 1

[45] Xi Wu, Zhen Xie, YuTao Gao, and Yu Xiao. Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features. In *2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2952–2956. IEEE, 2020. 1

[46] Yanbo Xu, Yueqin Yin, Liming Jiang, Qianyi Wu, Chengyao Zheng, Chen Change Loy, Bo Dai, and Wayne Wu. Transeditor: Transformer-based dual-space gan for highly controllable facial editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7683–7692, 2022. 1

[47] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265, 2019. 2

[48] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 325–341, 2018. 4

[49] Yulun Zhang, Chen Fang, Yilin Wang, Zhaowen Wang, Zhe Lin, Yun Fu, and Jimei Yang. Multimodal style transfer via graph cuts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5943–5951, 2019. 4

[50] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2185–2194, 2021. 1, 2, 4, 6, 7

[51] Yifan Zhao, Ke Yan, Feiyue Huang, and Jia Li. Graph-based high-order relation discovery for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15079–15088, 2021. 4

[52] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2382–2390, 2020. 1, 5, 6, 7