Space–Frequency and Global–Local Attentive Networks for Sequential Deepfake Detection

Guisheng Zhang, Qilei Li, Mingliang Gao, Siyou Guo, Gwanggil Jeon, and Ahmed M. Abdelmoniem

***Abstract*—The widespread misinformation generated by deepfake systems has emerged as a significant challenge in the dynamic realm of digital media. It poses threats to credibility, privacy, and security of information in daily life. Moreover, the increasing accessibility to facial editing tools further enables users to alter facial characteristics subtly through a series of intricate steps. To address the issue, we introduce a space–frequency and global–local attentive network (SFGLA-Net) for sequential deepfake detection. This method is designed to identify and analyze the sophisticated manipulated attributes of deepfake images. Specifically, we introduce a space–frequency fusion module to leverage the deep feature extracted in spatial and frequency domains, so as to exploit subtle inconsistencies and artifacts that are not perceptible in the spatial domain alone. Additionally, we design a global–local attention module to pinpoint the manipulated areas more accurately. Extensive experiments demonstrate the superior performance of the proposed method by significantly outperforming existing techniques in sequential deepfake detection. The code is available at https://github.com/guishengzhanga/SFGLA.**

***Index Terms*—Deepfake, global–local consistency, sequential deepfake detection, space–frequency fusion.**

I. Introduction

DEEP generative models have recently been widely adopted to create hyper-realistic facial images virtually indistinguishable from genuine ones. These deep generative models, commonly called deepfake, were initially used for entertainment, image restoration, and other similar applications [1]. However, malicious actors have exploited deepfake to create fabricated news, satirical images, phishing attacks, identity fraud, and other harmful content [2]. The misuse of deepfake has raised numerous societal and legal concerns.

To address the security concerns associated with deepfakes, researchers have proposed various methods to detect forged faces [3], [4], [5], [6], [7], [8]. For example, Khalil et al. [9] employed an Integrated capsule-based Deepfake (iCaps-Dfake) to address the issue of deepfake detection models that do not generalize well to different datasets. Recently, Wang et al. [10] proposed a noise-based deepfake detection model, named NoiseDF. The NoiseDF can focus on scrutinizing the inherent forensic noise traces present in deepfakes. These studies have achieved high accuracy in recognizing the authenticity of images. With the continuous advancement of deepfake, facial images are increasingly susceptible to a series of deepfake manipulations. Each of these facial images follows a specific procedural sequence. For example, in Fig. 1(a), the original image is sequentially manipulated to obtain a fake image. The operation sequence of the fake image is “lip-eyebrow-hair”.

Existing deepfake detection methods cannot identify sequences of deepfake manipulations. Therefore, sequential deepfake detection is proposed to identify and determine the correct sequence of manipulated regions. Fig. 1(b) and 1(c) summarizes the key dissimilarities between deepfake detection and sequential deepfake detection. Deepfake detection and sequential deepfake detection converge on the primary goal of discerning genuine images from altered imagery. Nevertheless, sequential deepfake detection focuses explicitly on pinpointing and mapping out the manipulated regions of faces. It also recognizes the order of these manipulations. Recently, Shao et al. [11] proposed a SeqFakeFormer to detect sequential deepfake. The SeqFakeFormer considers sequential deepfake detection as a specialized image captioning task. An enhanced cross-attention and an autoregressive block were employed to identify these manipulated regions in SeqFakeFormer. Nevertheless, due to the high variability of sequential deepfake manipulations, the sole use of pretrained ResNet and standard transformer architectures proves challenging for learning representations of various manipulation traces in deepfake images.

To address this problem, we propose a space–frequency and global–local attentive network (SFGLA-Net) to detect sequential deepfakes. Specifically, a space–frequency fusion module is proposed to consider the spatial positions and frequency distributions within images. Therefore, this module can provide

Received 28 April 2024; revised 4 November 2024 and 3 February 2025; accepted 10 February 2025. This work was supported in part by the National Natural Science Foundation of Shandong Province under Grant ZR2022MF307. *(Guisheng Zhang and Qilei Li contributed equally to this work.) (Corresponding author: Mingliang Gao.)*

Guisheng Zhang, Mingliang Gao, and Siyou Guo are with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255000, China (e-mail: sdut_guisheng@163.com; mlgao@sdut.edu.cn; siyouweiyi66@gmail.com).

Qilei Li and Ahmed M. Abdelmoniem are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. (e-mail: q.li@qmul.ac.uk; ahmed.sayed@qmul.ac.uk).

Gwanggil Jeon is with the Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea (e-mail: ggjeon@gmail.com).

Digital Object Identifier 10.1109/TCSS.2025.3541346

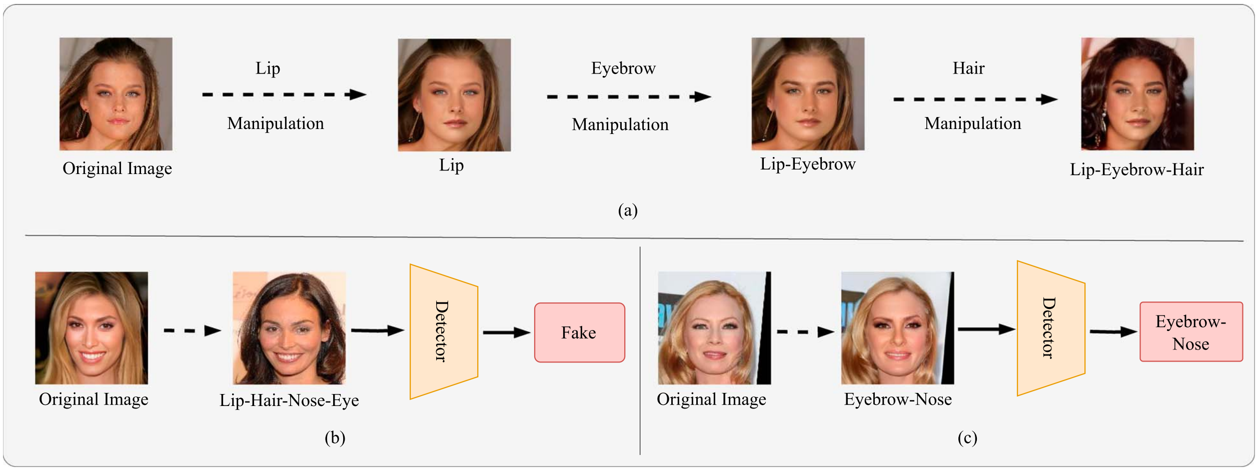

Fig. 1. Processes of sequential deepfake manipulation, deepfake detection and the sequential deepfake detection. (a) Sequential deepfake manipulation. (b) Deepfake detection. (c) Sequential deepfake detection.

a more comprehensive understanding of the information. The traditional self-attention mechanisms are less effective at processing local information. Furthermore, the scales of different manipulated regions vary. Therefore, we employ a global–local attention module to enhance the network grasp of both global and localized information. In sum, the contributions of this work are three-fold.

1) A space–frequency and global–local attentive network (SFGLA-Net) is proposed to improve the accuracy of sequential deepfake detection.
2) A Space–frequency fusion (SFF) module is proposed to consider the spatial positions and frequency distributions within images, which can reduce information loss. Integrating spatial and frequency information can capture images' textures, structures, and patterns more effectively.
3) A global–local attention module captures features of manipulated regions at various scales. This module includes self-attention and learnable local region attention mechanisms, and it effectively captures global and local information in facial images.

The rest of the article is structured as follows. Section II presents the related work. Section III illustrates the proposed method in detail. Section IV analyses the experimental results. The article is concluded in Section V.

II. Related Work

A. Deepfake Detection

Deepfake detection involves identifying digital media manipulated using deep learning technologies. It has gained urgency with the rise of deepfakes that can spread misinformation or harm reputations. Recently, Tan et al. [12] proposed a learning on gradients (LGrad), which uses a pretrained CNN model to transform images into gradients to visualize general artifacts and classifies images based on these representations. Qian et al. [13] introduced a dual-branch framework to learn the frequency-aware statistical differences between real and forged faces. Liu et al. [14] proposed FedForgery which integrates residual feature learning (RFL) with federated learning (FL) for enhanced forgery detection. Both deepfake detection and sequential deepfake detection aim to distinguish between real and fake images. However, traditional deepfake detection methods primarily focus on binary classification to assess the authenticity of facial images. Sequential deepfake detection aims to identify the manipulated facial region and determine the optimal sequence of operations. Recently, Hong et al. [15] proposed a contrastive learning framework with integrated multilabel ranking to achieve both classification and localization of deepfakes. Xia et al. [16] introduced a multicollaboration and multisupervision network (MMNet) for detecting manipulations and recovering images without needing knowledge of the manipulation techniques. Differing from these methods, the proposed SFGLA-Net considers using spatial-frequency domain information to detect forgery traces. Meanwhile, a learnable local attention mechanism is introduced to capture features of forgery regions at various scales.

B. Attention Mechanism

The attention mechanism boosts the accuracy and robustness of forgery detection by concentrating on essential feature areas. Generally, the attention mechanism enables networks to adjust the weights of input features dynamically. Among the attention mechanisms, spatial and channel attention are considered the most representative and have been extensively explored in the literature [17]. The groundbreaking work of SENet [18], which introduced the concept of decomposing attention into compression and excitation operations, has profoundly influenced the field of deep learning. Lin et al. [19] proposed a local learnable region attention module to concentrate on local information. By applying the local attention mechanism, each feature can discern and prioritize the most relevant local regions for processing. This work introduces a global–local attention module to boost performance in tasks requiring nuanced spatial awareness.

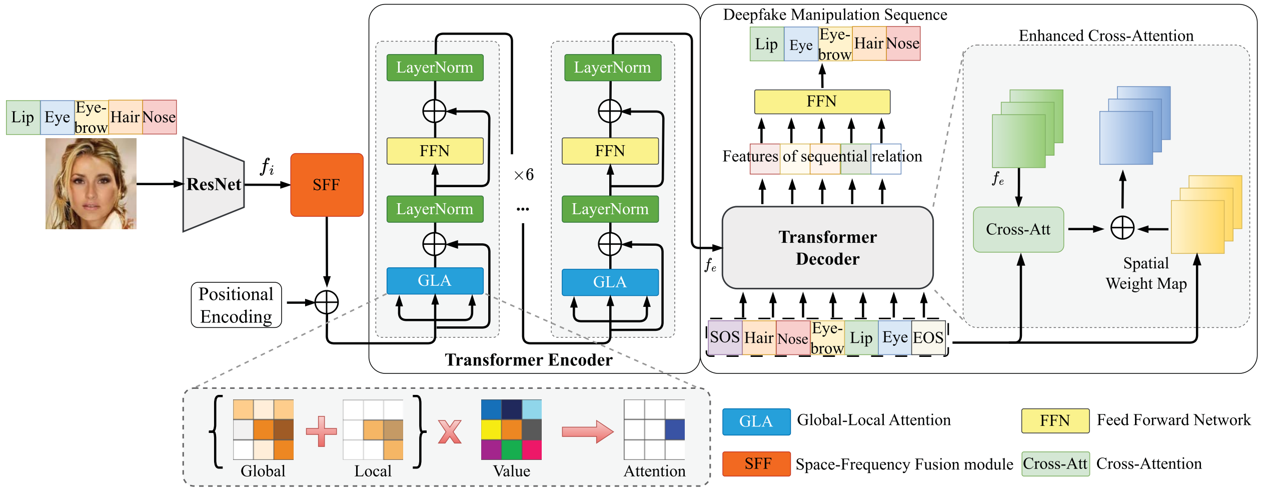

Fig. 2. Pipeline of the proposed SFGLA-Net for sequential deepfake detection.

C. Space–Frequency Fusion

Space–frequency fusion is a technique that processes signal or image data. It integrates information from the spatial and frequency domains to enhance data attributes or to extract additional useful information. The space–frequency fusion module enriches the understanding of the underlying data and facilitates a more comprehensive analysis by leveraging the strengths of spatial and frequency domains [20]. It can enhance the performance of deepfake detection. Liu et al. [21] presented spatial-phase shallow learning (SPSL) for face forgery detection. SPSL uses the phase spectrum in the frequency domain to detect artifacts from up-sampling, a common step in forgeries. Wang et al. [22] proposed a method called spatial-frequency dynamic graph (SFDG) for detecting deepfakes. The SFDG combines spatial and frequency features through a dynamic graph model to enhance the detection of subtle manipulation traces. Guo et al. [23] proposed a space–frequency interactive convolution (SFIConv) to capture the subtle manipulation cues of Deepfake creations. Liu et al. [24] employed a cross-domain local forensics (XDLF) for more general deepfake detection. This model is designed to concurrently leverage local tampering patterns across space, frequency, and time domains, thereby acquiring cross-domain features for forgery detection. Byeon et al. [25] proposed spatial-frequency and computer graphics decomposition to furnish extra cues for forgery identification. Although spatial-frequency fusion methods are widely applied in deepfake detection, their use in sequential deepfake detection is relatively rare. This work employed a space–frequency fusion module to capture more comprehensive information.

III. Proposed Method

A. Overview

To enhance the accuracy of sequential deepfake detection, we propose a space–frequency and global–local attentive network (SFGLA-Net). The SFGLA-Net aims to provide more comprehensive facial information from spatial and frequency domains and detect manipulation traces at various scales through global–local attention.

The pipeline of the SFGLA-Net model is illustrated in Fig. 2. Specifically, each facial image I is fed to a pretrained RestNet50 [26] to extract the feature f_i. The ResNet50 is to perform a preliminary extraction of features. Then, the space–frequency fusion (SFF) module is utilized to assist the model in capturing spatial and frequency information simultaneously. The SFF module enables the model to focus on the most informative parts of the image. Afterward, the output features from the SFF module are enriched with fixed positional encodings to yield a tensor. The fixed positional encodings are generated using sine and cosine functions. This tensor is fed into a transformer encoder to capture spatial manipulation traces. After that, the transformer decoder is employed to model sequential relationships based on spatial relational features. The transformer decoder enables the network to capture traces of sequence manipulation. The input of the transformer decoder is a sequence with seven tokens. It starts with “SOS” and ends with “EOS”. The other five tokens are randomly sorted. For example, it can be [“SOS, “Hair”, “Nose”, “Eyebrow”, “Lip”, “Eye”, “EOS”] or [“SOS”, “Lip”, “Eye”, “Eyebrow”, “Hair”, “Nose”, “EOS”]. The final output is the predicted forgery sequence generated by the proposed SFGLA-Net. In the transformer decoder, we adhere to the design principles of SeqFakeFormer [11] and implement a spatially enhanced cross-attention module with an autoregressive mechanism. This module initially generates varied spatial weight maps corresponding to specific manipulations. Then, it employs these maps to enhance the cross-attention mechanism.

B. Space–Frequency Fusion Module

In the computer vision domain, spatial and frequency information provide complementary benefits. The spatial information is paramount in discerning images’ overall structure and semantic content. Conversely, frequency information excels at capturing local details and textures within images. Sequential

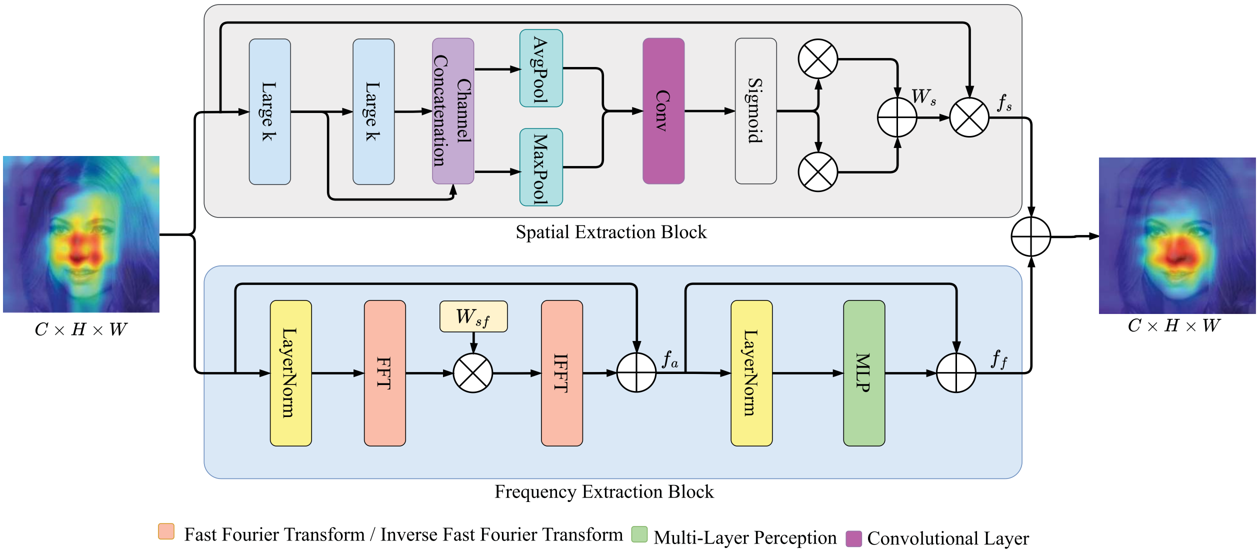

Fig. 3. Structure of the SFF Module. The module consists of two distinct branches: the first is dedicated to extracting spatial information, while the second focuses on capturing frequency information.

deepfake detection necessitates the identification of a diverse array of facial forgeries. This requires the network to conduct more detailed and comprehensive feature extraction on the facial regions. Therefore, a space–frequency fusion (SFF) module is proposed to capture local details and global information. The architecture of the SFF module is shown in Fig. 3. It consists of two branches, i.e., spatial extraction block and frequency extraction block. After extracting spatial and frequency-domain information, the SFF module merges them to obtain detailed facial features.

Spatial Extraction Block: The challenge of deepfake detection lies in identifying subtle distortions within images generated by deep learning models. This necessitates that the model can capture high-level features and subtle nuances. Consequently, within the spatial feature extraction block, a large selective kernel module is employed to dynamically adjust the receptive field to accommodate features of varying scales. This module mainly includes a large kernel convolution and spatial kernel selection. First, a large kernel is decomposed into a series of depth-wise convolutions with varying sizes and dilation rates. This step aims to expand the block's receptive field more efficiently, allowing it to capture local and global multiscale features. For the ith depth-wise convolution, the incrementally larger kernel size k, receptive field RF, and dilation rate dr are formulated as

$$\begin{aligned} &k_{i-1} \leq k_i \\ &dr_1 = 1, dr_{i-1} < dr_i < RF_{i-1} \\ &RF_1 = k_1, RF_i = dr_i(k_i - 1) + RF_{i-1}. \end{aligned} \tag{1}$$

Next, the spatial feature extraction block dynamically selects convolution kernels appropriate for the present input features via a spatial kernel selection mechanism. This mechanism initially merges the feature maps produced by various kernels. Subsequently, it extracts spatial feature descriptors utilizing both max pooling and average pooling techniques. Following a transformation layer, these descriptors yield multiple spatial attention maps. Each map is associated with a specific convolution kernel. Thus, the block adaptively selects each spatial position's optimal receptive field size. It ensures the efficient capture of contextual information pertinent to that position.

Finally, the feature maps weighted by the spatial selection mechanism are aggregated and multiplied by the original input features, forming the final output features. This process is formulated as

$$f_s = f_i \cdot W_s \tag{2}$$

where f_i is the input feature. The W_s is the weighted feature map computed by the spatial feature extraction block. The f_s is the output of the spatial feature extraction block. This block enables the network to adjust its receptive field dynamically according to objects' varying scales and contexts in remote sensing images.

Frequency Extraction Block: Certain patterns and structures within images can be more readily identified and captured in the frequency domain. An identical spectral layer is employed to extract frequency information in the frequency extraction block. Specifically, the frequency extraction block comprises a fast Fourier transform (FFT) layer, a weighted gate, and an inverse fast Fourier transform (IFFT) layer. The FFT layer is employed to convert physical space into spectral space. A learnable weight parameter W_{sf} is included to assess the importance of each frequency component adaptively. The W_{sf} can effectively capture the lines and edges of the facial image. This parameter is refined through gradient backpropagation. The block restores spectral space to physical space by utilizing the IFFT layer. Subsequently, a layer normalization and an MLP block facilitate the generation of the final output f_f. The f_f is formulated as

$$\begin{aligned} f_a &= \text{IFFT}(W_{sf} \times \text{FFT}(f_i)) + f_i \\ f_f &= \text{MLP}(f_a) + f_a. \end{aligned} \tag{3}$$

TABLE I
HYPER-PARAMETERS OF SFGLA-NET

Epoch	lr (Transformer)	lr (ResNet50)	Batch size	W_{sf}
150	$1e-3$	$1e-4$	64	1

C. Global–Local Attention Module

In the traditional transformer encoder, a self-attention layer concentrates on global information. It treats the input as an unordered sequence and indiscriminately considers all correlations between features. The self-attention module lacks perception of the spatial positional relationships of manipulated region features. Therefore, a global–local attention model focuses on features of manipulated regions at different scales.

Global Attention Block: The global attention block is a self-attention that can capture dependencies among different elements in the input sequence. It is mathematically denoted as

$$\text{Attention}_g = \text{Softmax}\left(QK^T/\sqrt{d_k}\right)V \tag{4}$$

where Q, K, and V denote the query, key, and value. The d_k represents the dimension of the key. To enhance attention to local information by the transformer encoder, a local learnable region attention [19] is employed in the proposed SFGLA-Net.

Local Attention Block: The local attention block dynamically allocates attention to different input image regions. It enables the model to focus on the most relevant local areas and adjust the attention scope based on the size of objects in the image. Since a rectangular region can be defined by two vertices, the regional filtering mechanism is introduced to obtain a dedicated area for each position. Given a point $m = (x_m, y_m)$ in the image, two filter functions at this position are defined as

$$f^{bl}(m \mid bl) = \begin{cases} 1, & \text{if } x_b \leq x_m < W, y_b \leq y_m < H \\ 0, & \text{others} \end{cases}$$

$$f^{ur}(m \mid tr) = \begin{cases} 1, & \text{if } 0 \leq x_m \leq x_t, 0 \leq y_m \leq y_t \\ 0, & \text{others} \end{cases} \tag{5}$$

where $bl = (x_b, y_b)$ represents the bottom-left point and $tr = (x_t, y_t)$ represents the top-right point. Therefore, for a given feature, the filtering region can be represented as

$$\begin{aligned} \hat{R}_i^{bl} &= \left[f^{bl}(m \mid bl)\right]_m^{W\times H} \\ \hat{R}_i^{tr} &= \left[f^{tr}(m \mid tr)\right]_m^{W\times H}. \end{aligned} \tag{6}$$

After that, the final region map R is expressed as

$$\begin{aligned} R &= R^{bl} \circ R^{tr} + R^{tr} \circ R^{bl} \\ R(m) &= \begin{cases} 1, & \text{if } x_b \leq x_m \leq x_t, y_b \leq y_m \leq y_t \\ 0, & \text{others.} \end{cases} \end{aligned} \tag{7}$$

According to the aforementioned filtering mechanism, the precision of each dedicated region relies solely on two nonlearnable discrete points. Therefore, a learnable regional filtering mechanism was developed to enable each feature to capture its optimal local region. Specifically, given the query vector $Q \in R^{WH\times d}$ and key vector $K \in R^{WH\times d}$, two predicted coverage probability maps (M^1,$M^2 \in R^{WH\times WH}$) are introduced to obtain two-dimensional learnable attention maps. These predicted coverage probability maps are obtained by

$$\begin{aligned} M^1 &= \text{Softmax}((QW_1^q)(KW_1^k)) \\ M^2 &= \text{Softmax}((QW_2^q)(KW_2^k)) \end{aligned} \tag{8}$$

where $W_1^q, W_1^k, W_2^q, W_2^k$ are the learnable parameter matrices. After that, The tensors M^1 and M^2 are reshaped into a three-dimensional tensor with dimensions $R^{WH\times W\times H}$. For each of $i \in WH$ along the first axis of M^1 and M^2, there are two maps $M_i^1, M_i^2 \in W \times H$. Then, by computing the cumulative distribution function (CDF) for the maps M_i^1, M_i^2, filter region maps are generated. The CDF describes the probability that a random point takes a value less than or equal to a specific point. We redesign the filter region maps by using CDF in two directions: from bottom-left (bl) to the top-right (tr) direction and in the reverse direction. For the specific point m, this process is formulated as follows:

$$\begin{aligned} f_{\text{CDF}}^{bl}(m \mid M_i) &= \sum_{x_j \leq x_m} \sum_{y_j \leq y_m} M_i(x_j, y_j) \\ f_{\text{CDF}}^{tr}(m \mid M_i) &= \sum_{x_j \geq x_m} \sum_{y_j \geq y_m} M_i(x_j, y_j). \end{aligned} \tag{9}$$

Finally, the learnable region mapping $\hat{R}_i$ is calculated for each feature position by the CDFs computed in the two directions. The final learnable region map R is obtained by combining $\hat{R}_i$

$$\begin{aligned} \hat{R}_i^{bl}(M_i) &= \left[f_{\text{CDF}}^{bl}(m \mid M_i)\right]_m^{W\times H} \\ R_i &= \hat{R}_i^{bl}(M_i^1) \circ \hat{R}_i^{tr}(M_i^2) + \hat{R}_i^{bl}(M_i^2) \circ \hat{R}_i^{tr}(M_i^1) \end{aligned} \tag{10}$$

where $\circ$ is the Hadamard product. The map R is employed to guide the attention mechanism towardthe regions of interest. The local attention is obtained by

$$\text{Attention}_l = \text{Softmax}\left(QK^T \circ R/\sqrt{d_k}\right)V. \tag{11}$$

To ensure that the module benefits from a broad overview and detailed local focus, the final attention of the GLA module integrates global and local attention mechanisms. The final attention $\text{Attrntion}_{\text{GLA}}$ is formulated as

$$\text{Attention}_{\text{GLA}} = \text{Attention}_g + \text{Attention}_l. \tag{12}$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Implementation Details

Table I shows the parameters used in this work. We implemented a learning rate warm-up phase, and the epoch was set to 20. A pretrained RestNet50 was employed for preliminary feature extraction. Batch size and epoch count were established at 64 and 150, respectively. Initial learning rates were designated as $1e-3$ for the transformer segment and $1e-4$ for the ResNet50 component. The trainable weight parameter W_{sf} was initialized as a matrix with all values set to 1. The framework was implemented in PyTorch [27] with two NVIDIA 3090Ti GPUs.

TABLE II
COMPARATIVE RESULTS WITH OTHER SOTA METHODS

Methods	Facial Components		Facial Attributes	
	Fix-Acc↑	Ada-Acc↑	Fix-Acc↑	Ada-Acc↑
DRN [28]	66.06	45.79	64.42	43.20
Multi-Cls [11]	69.65	50.57	66.66	46.00
DETR [29]	69.75	49.84	67.62	47.99
MA [30]	71.31	52.94	67.58	47.48
SeqFakeFormer[1] [11]	71.58	53.62	68.12	48.58
Two-Stream [31]	71.92	53.85	68.17	48.81
STPA [32]	72.34	54.20	68.71	49.56
MASFA-Net [33]	72.61	**55.71**	68.86	49.57
SFGLA-Net (our)	**72.85**	55.70	**68.94**	**49.74**

Note: The best results are highlighted in **bold**. [1]Performance evaluations were conducted with the officially released codes and performed on the same platform as ours.

B. Dataset

The sequential Deepfake (Seq-Deepfake) dataset [11] serves the purpose of aiding in the identification of facial manipulation sequences. It segregates into two specialized subsets, designated as the facial components dataset and the facial attributes dataset, each structured to support targeted detection endeavors. The facial components and attributes datasets comprise 35 166 and 49 920 images, respectively. The facial components dataset categorizes manipulations into regions labeled lip, eye, eyebrow, hair, nose. The subdataset comprises 28 distinct facial manipulation sequences. Similarly, the facial attributes dataset categorizes bangs, eyeglasses, beard, smiling, young as manipulated regions. This subdataset presents 26 distinct facial manipulation sequences.

C. Evaluation Metrics

In this work, two indicators, fixed accuracy (Fix-Acc) and adaptive accuracy (Ada-Acc) [11], are utilized to assess the efficacy of the proposed SFGLA-Net. For Fix-Acc, a predetermined length of five is set for manipulation sequence predictions. If the predicted sequence is shorter than five, the “no manipulation” (NM) category is added to the annotated sequence. Finally, the operational categories of the predicted sequences are matched with the corresponding annotations to determine the accuracy of the assessment. Unlike the Fix-Acc, the Ada-Acc assesses the accuracy by predicting manipulation sequences under adaptive lengths. In practice, the proposed method automatically halts prediction once the “EOS” token is detected. This allows it to detect facial manipulation sequences with adaptive lengths effectively. In contrast to the Acc metric, the Fix-Acc and Ada-Acc provide an accuracy assessment of sequential deepfake detection.

D. Comparison With State-of-the-Art Methods

To evaluate the effectiveness and advantages of the proposed SFGLA-Net, we performed comparative analyses against five state-of-the-art (SOTA) methods acknowledged as forefront solutions in sequential deepfake detection. The DRN [28], Multi-Cls [11], MA [30], Two-Stream [31], and DETR [29] merely categorize various deepfake manipulations into distinct classes and identify these manipulations. These models regard sequential deepfake detection as a multiclassification task. Moreover, they solely focus on the spatial information of the image and overlook the sequential manipulation traces. The SeqFakeFormer [11], STPA [32], MASFA-Net [33], and the proposed SFGLA-Net are designed to process sequential forged information concurrently.

The comparison results are shown in Table II. The comparison results indicate that the proposed SFGLA-Net outperforms other SOAT methods in sequential deepfake detection. On the facial components dataset, the SFGLA-Net method scores 72.85 and 55.70 in Fix-Acc and Ada-Acc, respectively. Compared with the SOAT methods, the proposed approach ranks first in Fix-Acc and second in Ada-Acc. Specifically, compared with MASFA-Net, the proposed method improves Fix-Acc by 0.33%. In comparison to the baseline method, SeqFakeFormer, the proposed method achieves improvements of 1.77% in Fix-Acc and 3.88% in Ada-Acc. Additionally, when contrasted with the third-best method, STPA, the proposed SFGLA-Net shows enhancements of 0.71% in Fix-Acc and 2.77% in Ada-Acc. Moreover, the proposed network remains competitive on facial attribute datasets. The SFGLA-Net outperforms SeqFakeFormer in terms of Fix-Acc by 1.20% and Ada-Acc by 2.36% on facial attribute datasets. Compared with the second-best method, MASFA-Net, the proposed SFGLA-Net improves Fix-Acc and Ada-Acc by 0.12% and 0.34%, respectively. These results underscore the superior efficacy of the proposed SFGLA-Net. It employs a pace-frequency attention mechanism that yields detailed and comprehensive facial information. Furthermore, it incorporates a global–local attention mechanism within the transformer encoder, which can effectively capture manipulated clues. This facilitates the achievement of outstanding results in sequential deepfake detection.

The visualized results of detecting sequential deepfake manipulations are depicted in Fig. 4. The subjective results indicate that the proposed network proficiently discerns the sequence of deepfake manipulations with varying lengths. Moreover, to provide further visual evidence of the effectiveness of the SFGLA-Net, the attention visualization maps of SeqFakeFormer, STPA, MASFA-Net, and the proposed SFGLA-Net are

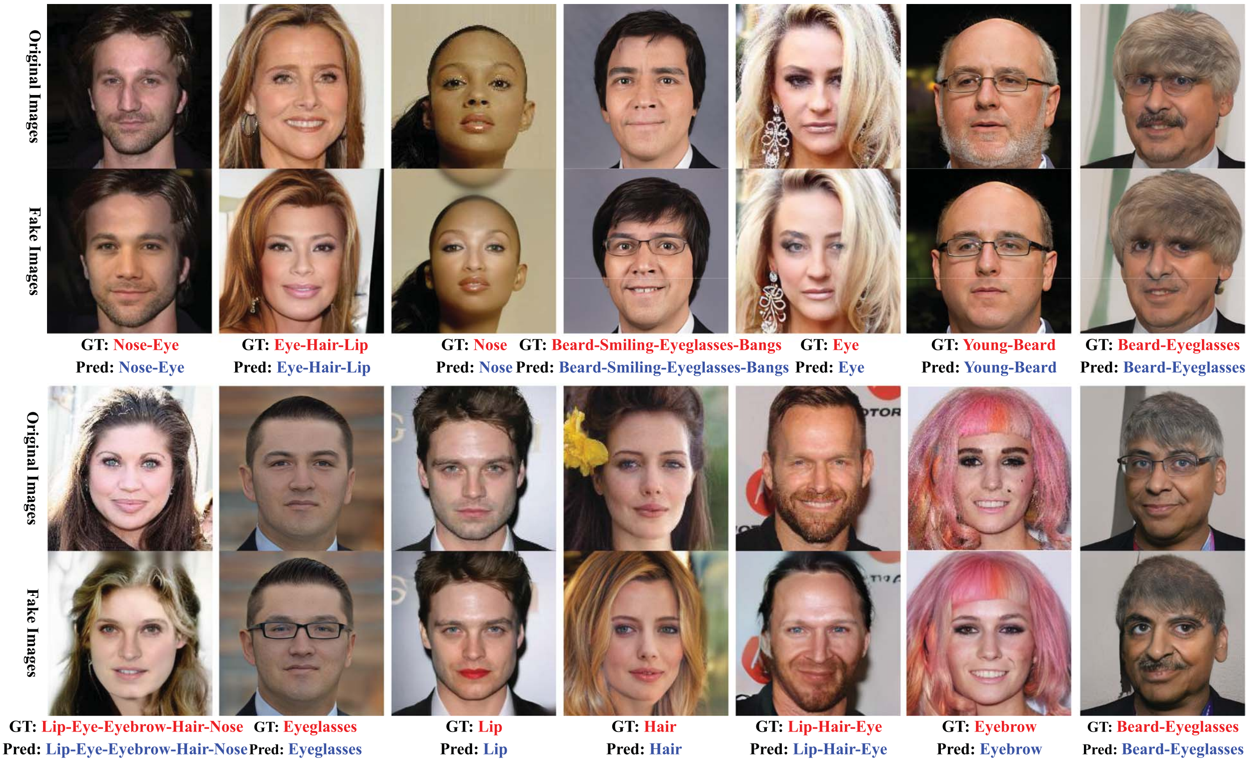

Fig. 4. Qualitative results sampled from the Seq-Deepfake dataset. The ground truth (GT) and the prediction (Pred) are highlighted in red and blue, respectively.

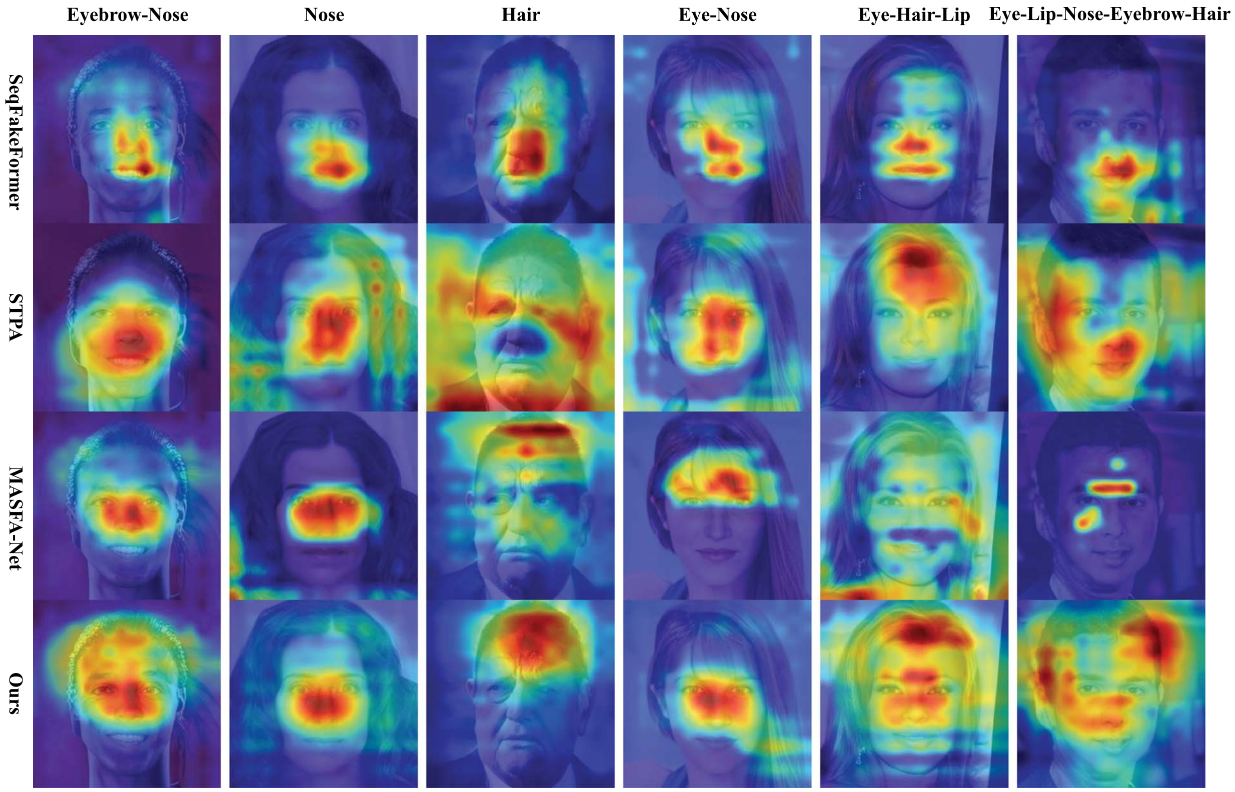

Fig. 5. Visualization of attention map between SeqFakeFormer, STPA, MASFA-Net, and the proposed SFGLA-Net.

TABLE III
ABLATION STUDY ON THE KEY COMPONENTS

Methods			Facial Components		Facial Attributes	
Baseline	**SFF**	**GLA**	Fix-Acc↑	Ada-Acc ↑	Fix-Acc↑	Ada-Acc↑
✓			71.58	53.62	68.12	48.58
✓	✓		71.69	53.69	68.46	49.20
✓		✓	72.12	54.28	68.18	48.71
✓	✓	✓	**72.85**	**55.70**	**68.94**	**49.74**

Note: The best results are highlighted in **bold**.

TABLE IV
ABLATION STUDY ON THE KEY COMPONENTS OF THE SPATIAL-FREQUENCY FUSION MODULE

Methods			Facial Components		Facial Attributes	
Space	**Frequency**	**Space–Frequency**	Fix-Acc↑	Ada-Acc ↑	Fix-Acc↑	Ada-Acc↑
✓			71.85	52.98	68.31	49.47
	✓		72.25	54.42	68.08	49.12
		✓	**72.85**	**55.70**	**68.94**	**49.74**

Note: The best results are highlighted in **bold**.

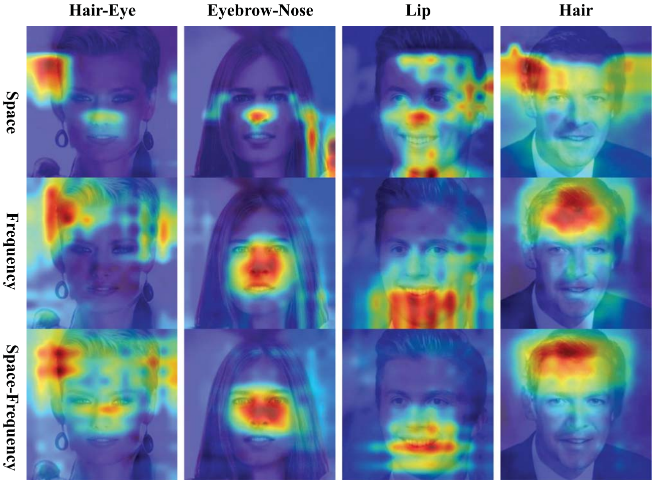

Fig. 6. Visualization of attention map generated with space information, frequency information, and space–frequency information.

shown in Fig. 5. It demonstrates that when detecting the same type of forgery sequence, the proposed SFGLA-Net focuses more on various forged regions than SeqFakeFormer.

E. Ablation Study

To assess the effectiveness of the essential elements, the ablation experiments are conducted on the Seq-Deepfake dataset. The results are shown in Table III. In these ablation studies, the Baseline is the SeqFakeFormer. The SFF denotes the spatial-sequential fusion module. The GLA represents the global–local attention module. It shows that when the Baseline is equipped with the SFF and GLA modules individually, the performance has been substantially improved. These results indicate that these introduced modules contribute significantly to enhancing the detection performance of the network.

To evaluate the effectiveness of the proposed spatial-frequency fusion module, we conducted ablation studies on the Facial components and Facial attributes datasets. The results are shown in Table IV. It indicates that employing the spatial-frequency fusion module outperforms the single spatial and frequency modules. The attention map of the key components in the spatial-frequency fusion module is shown in Fig. 6. It demonstrates that using spatial or frequency domain methods alone cannot effectively focus on the forged regions. However, employing the spatial-frequency fusion module enables the network to focus on these forged regions.

V. CONCLUSION

In this article, we introduce the SFGLA-Net for detecting sequential deepfake manipulations. A spatial-frequential fusion module and a global–local attention module are employed to provide comprehensive facial features and precise detection of forgery operations. The proposed SFGLA-Net demonstrates superiority over existing methods for sequential deepfake manipulation detection on several datasets. In the future, our work will focus on enhancing the ability of networks to capture forged region features and improving the accuracy of sequential deepfake detection.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest related to this publication.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

[1] Y. Li, S. Bian, C. Wang, K. Polat, A. Alhudhaif, and F. Alenezi, “Exposing low-quality deepfake videos of social network service using spatial restored detection framework,” *Expert Syst. Appl.*, vol. 231, 2023, Art. no. 120646.

[2] 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, 2020.

[3] Q. Li, M. Gao, G. Zhang, W. Zhai, J. Chen, and G. Jeon, “Towards multimodal disinformation detection by vision-language knowledge interaction,” *Inf. Fusion*, vol. 102, 2024, Art. no. 102037.

[4] Q. Li, M. Gao, G. Zhang, and W. Zhai, “Defending deepfakes by saliency-aware attack,” *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 5060–5067, Aug. 2024.

[5] S. Guo, Q. Li, M. Gao, G. Zhang, J. Pan, and G. Jeon, “Deep learning-based face forgery detection for facial payment systems,” *IEEE Consum. Electron. Mag.*, early access, 2024.

[6] X. Zhang, S. Dadkhah, A. G. Weismann, M. A. Kanaani, and A. A. Ghorbani, “Multimodal fake news analysis based on image–text similarity,” *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 1, pp. 959–972, Feb. 2023.

[7] G. Zhang, M. Gao, Q. Li, W. Zhai, and G. Jeon, “Multi-modal generative deepfake detection via visual-language pretraining with gate fusion for cognitive computation,” *Cogn. Comput.*, vol. 16, no. 6, pp. 1–14, 2024.

[8] G. Zhang, M. Gao, Q. Li, W. Zhai, G. Zou, and G. Jeon, “Disrupting deepfakes via union-saliency adversarial attack,” *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 2018–2026, Feb. 2024.

[9] S. S. Khalil, S. M. Youssef, and S. N. Saleh, “icaps-dfake: An integrated capsule-based model for deepfake image and video detection,” *Future Internet*, vol. 13, no. 4, p. 93, 2021.

[10] T. Wang and K. P. Chow, “Noise based deepfake detection via multi-head relative-interaction,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 12, 2023, pp. 14548–14556.

[11] R. Shao, T. Wu, and Z. Liu, “Detecting and recovering sequential deepfake manipulation,” in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer Nature, 2022, pp. 712–728.

[12] C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei, “Learning on gradients: Generalized artifacts representation for gan-generated images detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12105–12114.

[13] 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.*, Cham, Switzerland: Springer International Publishing, 2020, pp. 86–103.

[14] D. Liu et al., “Fedforgery: Generalized face forgery detection with residual federated learning,” *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 4272–4284, 2023.

[15] C.-Y. Hong, Y.-C. Hsu, and T.-L. Liu, “Contrastive learning for deepfake classification and localization via multi-label ranking,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 17627–17637.

[16] R. Xia, D. Liu, J. Li, L. Yuan, N. Wang, and X. Gao, “MMNet: multi-collaboration and multi-supervision network for sequential deepfake detection,” *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 3409–3422, 2024.

[17] M.-H. Guo et al., “Attention mechanisms in computer vision: A survey,” *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.

[18] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[19] H. Lin, Z. Ma, R. Ji, Y. Wang, and X. Hong, “Boosting crowd counting via multifaceted attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19628–19637.

[20] P. Liu, Q. Tao, and J. T. Zhou, “Evolving from single-modal to multi-modal facial deepfake detection: A survey,” 2024, *arXiv:2406.06965*.

[21] H. Liu et al., “Spatial-phase shallow learning: rethinking face forgery detection in frequency domain,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 772–781.

[22] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng, “Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7278–7287.

[23] Z. Guo, Z. Jia, L. Wang, D. Wang, G. Yang, and N. Kasabov, “Constructing new backbone networks via space-frequency interactive convolution for deepfake detection,” *IEEE Trans. Inf. Forensics Secur.*, vol. 19, pp. 401–413, 2024.

[24] Z. Liu, H. Wang, and S. Wang, “Cross-domain local characteristic enhanced deepfake video detection,” in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 3412–3429.

[25] H. Byeon et al., “Deep learning model to detect deceptive generative adversarial network generated images using multimedia forensic,” *Comput. Elect. Eng.*, vol. 113, 2024, Art. no. 109024.

[26] H. M. Nguyen and R. Derakhshani, “Eyebrow recognition for identifying deepfake videos,” in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Piscataway, NJ, USA: IEEE Press, 2020, pp. 1–5.

[27] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 8024–8035, 2019.

[28] S.-Y. Wang, O. Wang, A. Owens, R. Zhang, and A. A. Efros, “Detecting photoshopped faces by scripting photoshop,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10072–10081.

[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *Proc. Eur. Conf. Comput. Vis*, Springer, 2020, pp. 213–229.

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

[31] Y. Luo, Y. Zhang, J. Yan, and W. Liu, “Generalizing face forgery detection with high-frequency features,” in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2021, pp. 16317–16326.

[32] G. Zhang, M. Gao, Q. Li, S. Guo, and G. Jeon, “Detecting sequential deepfake manipulation via spectral transformer with pyramid attention in consumer IoT,” *IEEE Trans. Consum. Electron.*, early access, 2024.

[33] G. Zhang, Q. Li, M. Gao, and G. Jeon, “Towards sequential deepfake detection using deep learning for privacy protection,” *IEEE Consum. Electron. Mag.*, early access, 2024.