Detecting Sequential Deepfake Manipulation via Spectral Transformer with Pyramid Attention in Consumer IoT

Guisheng Zhang, Mingliang Gao*, Qilei Li, Siyou Guo, Gwanggil Jeon*

***Abstract*—Recently, the Consumer Internet of Things (CIoT) has brought great convenience to people. In CIoT, face image information is indispensable for payment and checking the identity of the user in the transaction. However, the misuse of deepfake face information in CIoT transactions is a growing problem. It has seriously violated the property and privacy of individuals. Moreover, with the proliferation of easily accessible facial editing applications, individuals can effortlessly manipulate facial components through sequential multi-step manipulations. To solve this issue, we propose a Spectral Transformer with a Pyramid Attention (STPA) model to detect sequence permutations in manipulated facial images. Specifically, we introduce a pyramid attention module that integrates both spatial and channel attention mechanisms to prioritize the face region over the background region. Additionally, a spectral Transformer is employed concurrently to extract global and local features to facilitate the fine-grained extraction of the face forgery region. Comprehensive experiments prove that the proposed method can enhance the detection accuracy of the sequential deepfake manipulation task through the fine-grained extraction of features in the face forgery region.**

***Keywords*—Consumer Security, Privacy Preservation, Sequential Deepfake Detection, Spectral Transformer, Pyramid Attention**

I. Introduction

The evolution of the intelligent Consumer Internet of Things (CIoT) has greatly facilitated people's lives. Nevertheless, it also introduces various avenues for distinct security threats and compromises the privacy of users [1]. For example, in CIoT, criminals engage in the theft of user facial information and use deepfake technologies to fabricate false identities. Subsequently, they use this forged face information to commit fraud in transactions. Moreover, with the advancement of deep learning techniques, particularly the emergence of sophisticated methods such as Generative Adversarial Networks (GANs) [2], hyper-realistic face images can be effortlessly generated. Such misuses of deepfake technology in CIoT [3], [4] raises potential societal concerns, *e.g.,* misinformation and privacy infringements. For example, some face images of celebrities have been maliciously manipulated for explicit content, and bring them substantial harassment and privacy concerns.

To mitigate the adverse effects resulting from the abuse of deepfake technology, numerous detection methods [5]–[7] have been proposed and extensively investigated. Zhao *et al.* [8] introduced a multi-attentional deepfake detection network designed to uncover subtler distinctions between authentic and manipulated images. Guarnera *et al.* [9] introduced a multi-level deepfake detection approach to recognize fake images generated by various GANs and diffusion models. Jeong *et al.* [10] introduced a method termed Frequency Perturbation GAN (FrePGAN), which incorporates a frequency-level perturbation during training. These traditional studies can attain a high level of accuracy in discerning the authenticity of images. In real-life scenarios, a facial image may undergo multiple deepfake manipulations, each following a specific sequence. As illustrated in Fig. 1 (b), an artificial facial image is generated by manipulating the original image through two successive deepfake steps, namely "Lip" and "Eyebrow" manipulation. Consequently, the manipulation sequence for this synthesized image is designated as "Lip-Eyebrow". However, conventional deepfake detection methods struggle to identify the sequence of this facial manipulation.

To address this issue, the Sequential Manipulation Detection method was built to detect the sequence of deepfake manipulation. The distinctions between the sequential manipulation detection method and the traditional deepfake detection method are shown in Fig. 1. The conventional deepfake detection method is limited to discerning the authenticity of an image, whereas the sequential manipulation detection method can detect the manipulation sequence of an image. Recently, a Seq-DeepFake Transformer (SeqFakeFormer) [11] was proposed to detect sequences of facial manipulations with diverse lengths. The framework consists of two key parts: spatial relation extraction and sequential relationship modeling with spatially enhanced Cross-attention. These parts work synergistically, each complementing the other's functionality. The detection of sequential manipulations can be conceptualized as a specialized image-to-sequence task, which can be forecasted utilizing the autoregressive algorithm embedded within the SeqFake-

This work is supported in part by the National Natural Science Foundation of Shandong Province (No.ZR2022MF307). (Corresponding authors: Mingliang Gao and Gwanggil Jeon)

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, and siyouweiyi66@gmail.com.)

Qilei Li is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, United Kingdom (e-mail: qilei.li@outlook.com).

Gwanggil Jeon is with the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, 255000, China. Also, he is with the Department of Embedded Systems Engineering, Incheon National University, Incheon 22012, South Korea (e-mail:ggjeon@gmail.com).

* Mingliang Gao and Gwanggil Jeon are the corresponding authors.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3414319

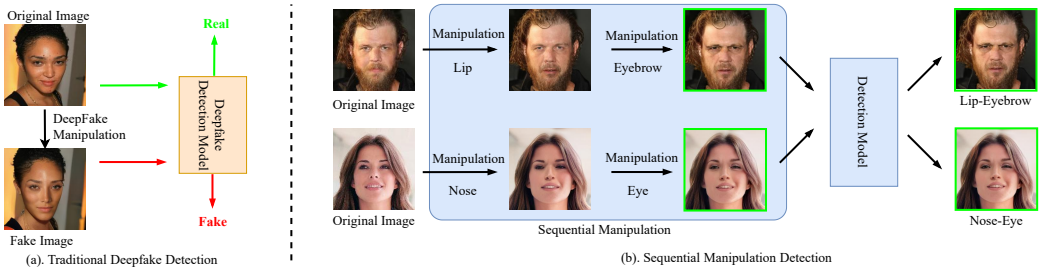

Fig. 1. Comparison between (a) the traditional deepfake detection method and (b) the sequential manipulation detection method.

Former. Given the diverse sizes of facial manipulation regions, the detection method requires meticulous feature extraction from these regions. However, the SeqFakeFormer framework falls short of achieving satisfactory performance for fine-grained feature extraction.

In this work, to address this issue, we introduce a Spectral Transformer with Pyramid Attention (STPA) framework for refined extraction of facial manipulation. Specifically, the proposed model captures a face image. Meanwhile, the image is fed to a backbone network to extract coarse-grained visual features. Subsequently, the extracted features are channeled into the pyramid attention module. The primary objective of employing the pyramid attention module is to emphasize the facial region in the image while minimizing attention to irrelevant background areas. Then, by utilizing the Spectral transformer encoder, the proposed method adeptly captures spatial relationships within the features of manipulated regions. Finally, a decoder incorporating spatially enhanced cross-attention is employed to derive the final features related to sequential manipulation. These features are input into a Fast Forward Network (FFN) to unveil the sequence of deepfake manipulations. In sum, the contributions of the work are three-fold:

- We design a Spectral Transformer with Pyramid Attention (STPA) to improve the accuracy of sequential deepfake manipulation detection.
- We introduce a pyramid attention to emphasize the facial region while disregarding the background. This module initially acquires multi-scale features through grouped convolution, followed by deriving attention weights using spatial and channel attention mechanisms.
- We propose a spectral transformer encoder for extracting features from fine-grained face manipulation regions. This transformer encoder employs a spectral model to address the deficiency of attention modules in capturing localized features.

The rest paper 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 paper is concluded in Section V.

II. Related Work

A. Consumer Internet of Things

The concept of Consumer Internet of Things (CIoT) is to apply IoT architecture and technologies to consumer electronic devices and products. It has a wide range of applications, such as smart home systems, wearable technology, connected household appliances, and intelligent health devices [12], [13]. Although its principle is to improve user experience and overall quality of life, the CIoT-connected devices concurrently pose significant privacy and security challenges due to the extensive collection of user information [14]. For example, criminals engage in the theft of user facial information and employ deepfake technologies to fabricate false identities. Subsequently, they utilize this forged facial information to perpetrate fraud in transactions.

To address such issues, numerous works have been conducted to protect consumer privacy within CIoT. For example, Xu *et al.* [15] introduced an evaluation model to assess privacy risks in federated learning within the CIoT. This model integrates model inversion and white-box membership inference attacks for efficient and high-quality privacy data reconstruction. Rabieinejad *et al.* [16] proposed a two-level privacy-preserving framework combining federated learning with partially homomorphic encryption. This framework aims to address security vulnerabilities and enhance privacy protection in CIoT. In this work, we introduce a spectral transformer with a pyramid attention model for the detection of deepfakes.

B. Deepfake detection

In the contemporary digital landscape, the pervasive use of deepfake technologies has introduced a new set of challenges in verifying the authenticity of digital content. As these techniques advance, the boundary between genuine and manipulated material becomes progressively indistinct. Consequently, distinguishing between reality and fabrication becomes a challenging endeavor. To mitigate the adverse effects of deepfakes, researchers are actively developing various methods for their detection. Recently, FST-Matching [17] was proposed to address the challenge of suboptimal performance exhibited by detection modules when applied to compressed images. This model enhances the detection performance of the network by acquiring representations through image matching. Reiss *et al.* [18] introduced a practical recipe for deepfake fact checking (FACTOR) to detect fake images generated by unknown forgery methods. In contrast to other methods, the FACTOR is a non-training-based approach that exhibits a certain degree of generalization. Shuai *et al.* [19] addressed the lack of universality and overfitting to image content of

current deep forgery detection methods through a dual-stream network. They employed a semi-supervised strategy for plaque similarity learning to estimate annotations at the plaque level for forged locations. Unlike the aforementioned methods that concentrate solely on a single stage of facial manipulation, Shao *et al.* [11] proposed the SeqFakeFormer model to detect the sequential deepfake manipulations. The SeqFakeFormer breaks the deep forgery research paradigm of binary classification of true and false labels. Nevertheless, it lacks of extracting fine-grained features.

In this work, we introduce a Spectral Transformer with a Pyramid Attention (STPA) model to detect sequential deepfake manipulations. The proposed framework incorporates a pyramid attention model and a spectral transformer to extract fine-grained features from facial forgery regions.

C. Attention mechanism

The incorporation of the attention mechanism in deepfake detection plays a pivotal role in elevating the sensitivity of the network to signs of forgery [20]. The attention mechanism enhances the accuracy and robustness of forgery detection by concentrating on crucial feature areas. Many attention mechanisms are proposed in the literature [21]. Among them, the most representative attention mechanisms are spatial attention and channel attention. The channel attention is primarily derived from the originally obtained feature maps by channel as the weights of each channel. The spatial attention enables the network to focus on specific regions of the input data.

The most representative channel attention is Squeeze-and-Excitation (SE) [22]. The SE module enhances attention toward pivotal channels while diminishing dependence on less crucial channels. Wang *et al.* [23] proposed an Efficient Channel Attention (ECA) module by fast one-dimensional convolution, and it is lightweight channel attention. Wang *et al.* [24] introduced a spatial attention mechanism termed non-local attention. The non-local attention captures more complex structural information by considering any two positions in an image. The Spatial-reduction attention (SRA) [25] was introduced to reduce the computational cost by shrinking the spatial scale of key and numerical inputs. Moreover, recent studies have shown that spatial attention and channel attention can synergize to improve the performance of networks. For example, Woo *et al.* [26] proposed the Convolutional Block Attention Module (CBAM) to focus on channel and spatial attention, simultaneously. The channel attention module emphasizes key channels, whereas the spatial attention module enhances the discernment of distinct spatial locations. The Triplet Attention [27] was built to capture cross-dimensional interactions between channels and spatial dimensions through a three-branch structure.

The aforementioned work aims to combine channel and spatial attention, but they both struggle to capture multi-scale spatial information effectively. To address this issue, the Pyramid Split Attention (PSA) [28] was introduced. Utilizing operations such as spatial pyramid convolution (SPC) and complex SE weighting modules, the PSA excels at fusing contextual information at multiple scales. In this work, we introduce the pyramidal attention module to emphasize facial regions while disregarding background areas unrelated to the face.

D. Spectral transformer

The Spectral Transformer is a deep learning model based on the Transformer architecture, meticulously designed to analyze time-series signals and spectrogram data. By integrating a frequency-domain attention mechanism, it adeptly captures the inherent frequency-domain information in signals, thereby enhancing the capability to extract local features from images.

Recently, the spectral transformer holds paramount significance in analyzing information within the frequency domain of images [29]. Rao *et al.* [30] put forward a Global Filter Network (GFNet) to improve the computational efficiency and robustness of the network by replacing the attention module with a 2D inverse Fourier transform. Lee *et al.* [31] proposed a Transformer-like model termed Mixing Tokens with Fourier Transforms (FNet). In FNet, a self-attention sublayer is replaced by an unparameterized Fourier Transform. Therefore, the FNet enhances training speed while concurrently reducing the number of parameters. A wavelet vision transformer (Wave-ViT) [32] was reported by Yao *et al.*. In Wave-ViT, the Discrete Wavelet Transform (DWT) is employed for lossless downsampling of self-attentive keys and values. Additionally, the inverse discrete wavelet transform (IDWT) is employed to enhance self-attention outputs by aggregating local contexts with an expanded receptive field. Patro *et al.* [29] introduced a SpectFormer, which incorporates a combination of spectral attention and later multi-headed attention. The spectral layer is employed to capture the different frequency components of the image to capture localized frequencies. This enhancement contributes to the capability of the transformer to capture specific localized features.

In this work, we incorporate a spectral module into a transformer encoder to conduct fine-grained extraction of facial images. This structure addresses the limitations associated with the challenge of the transformer in accurately capturing localized features.

III. Proposed Method

A. Overview

In contrast to traditional deepfake detection methods, this work aims to develop a framework proficient in addressing the intricate array of falsification techniques applied to a single facial image. Moreover, it seeks to predict the sequence of these deepfake manipulations accurately. Detecting forged manipulation sequences necessitates the extraction of spatial relationship features from these sequences, which constitutes a challenging task. Consequently, to precisely forecast sequences of manipulated manipulations, a critical challenge involves capturing nuanced spatial manipulation regions and conducting meticulous extraction of spatial features.

To capture nuanced spatial manipulation regions and conduct detailed spatial feature extraction, we propose a framework termed Spectral Transformer with Pyramid Attention (STPA). This framework aims to augment precision in

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3414319

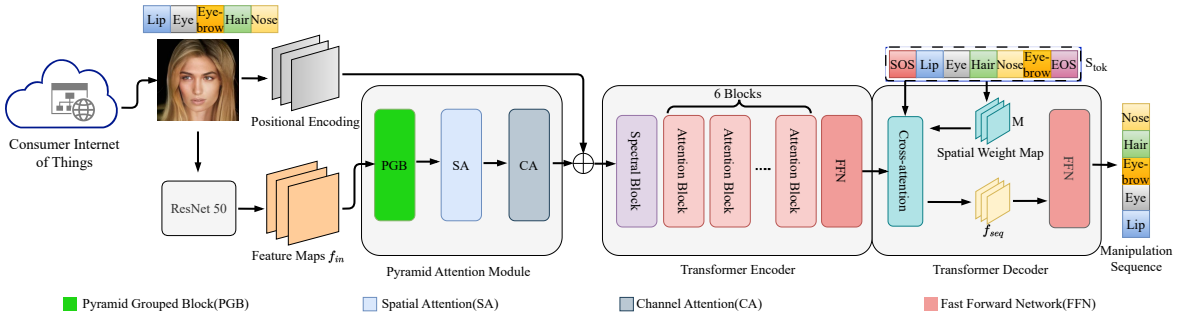

Fig. 2. The overall structure of the proposed STPA for sequential deepfake manipulation detection.

the detection of sequential deepfake manipulation. The overall structure of the STPA framework is illustrated in Fig. 2. Given an input image I captured by an imaging device in CIoT network, the proposed framework takes it as input and processes it with a pre-trained ResNet50 [33] for coarse-grained visual feature extraction. The coarse-grained visual feature f_{in} is represented as:

$$f_{in} = f_{resnet}(I), \tag{1}$$

where the $f_{resnet}(\cdot)$ is the ResNet50. $f_{in} \in R^{C\times H\times W}$, the variables H and W denote the height and width of the feature f_{in} respectively, C represents the number of channels in the feature f_{in}. Subsequently, the feature f_{in} is fed into the pyramid attention module to guide the proposed framework toward emphasizing the facial region in the image while minimizing attention to irrelevant background areas. The proposed method employs a spectral transformer encoder, which integrates a spectral model with self-attention modules, to effectively capture spatial relationships in features of manipulated regions. Subsequently, a decoder featuring spatially-enhanced cross-attention acquires the final sequential relation features. These features are then processed through a Fast Forward Network (FFN), enabling the framework to deduce the sequence of deepfake manipulation. The architecture of this framework is intricately designed to accurately discern the subtleties of sequential manipulation.

B. Pyramid attention module

To enhance the detection of deepfake manipulation sequences, it is imperative to accurately identify features in subtly manipulated areas, typically found within the portrait region. The pre-trained ResNet50 model is utilized for initial feature extraction from input images. However, given its unfamiliarity with the specific dataset, ResNet50 may inadvertently focus on areas not associated with facial features. To address this, the proposed method incorporates a pyramid attention module, designed to steer the network's focus more precisely towards facial regions. This targeted approach aids in the more effective discernment of manipulations relevant to facial characteristics.

This pyramid attention module comprises three components, namely Pyramid Grouped Block (PGB), Spatial Attention (SA), and Channel Attention (CA). The PGB is a group convolution module capable of extracting features across various scales through the utilization of diverse convolutional kernels. The SA is a spatial attention module, and it enables the model to capture information within the facial region. As a channel attention model, the CA allows the system to adjust the weights of various channels dynamically, and it emphasizes the image features crucial for the sequence of deepfake manipulation. The architecture of the pyramid attention module is shown in Fig. 3.

Pyramid grouped block To capture features across various scales, the input feature f_{in} is evenly partitioned into four segments along the channel dimension. The channel $m^{'}$ of each segmented part can be denoted as:

$$m^{'} = \frac{m}{4}, \tag{2}$$

where m is the number of channels in the feature f_{in}. To enhance the capacity of the proposed method in capturing diverse features across distinct spatial locations, multi-scale grouped convolution is employed to extract multi-scale features. The size of the group g can be denoted as:

$$g_i = 2^{\frac{k_i-1}{2}}, \tag{3}$$

where k is convolution kernel size and $k_i = 2i + 3, i = 0, 1, 2, 3$. We employ convolution kernels of varying sizes to extract multi-scale features in the four segmented portions. The multi-scale features f_i are formulated as:

$$f_i = \mathrm{Conv}(k_i \times k_i, g_i)(f_{in}), i = 0, 1, 2, 3, \tag{4}$$

where f_{in} represents the coarse-grained visual feature output from the ResNet. Subsequently, the multi-scale features fed into the spatial attention module for further process.

Spatial attention The spatial attention module aims to highlight the importance of specific locations within the image, focusing on areas of interest while diminishing background distractions. By applying the Spatial Attention (SA) module across each of the four branches, the network's attention is redirected away from the background, intensifying the focus on facial region features. The configuration of SA is illustrated in Fig. 3, and it is detailed as follows.

First, the average pooling and maximum pooling operations are operated along the channel axis. This process yields a combination of global and local features. Then, the features

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3414319

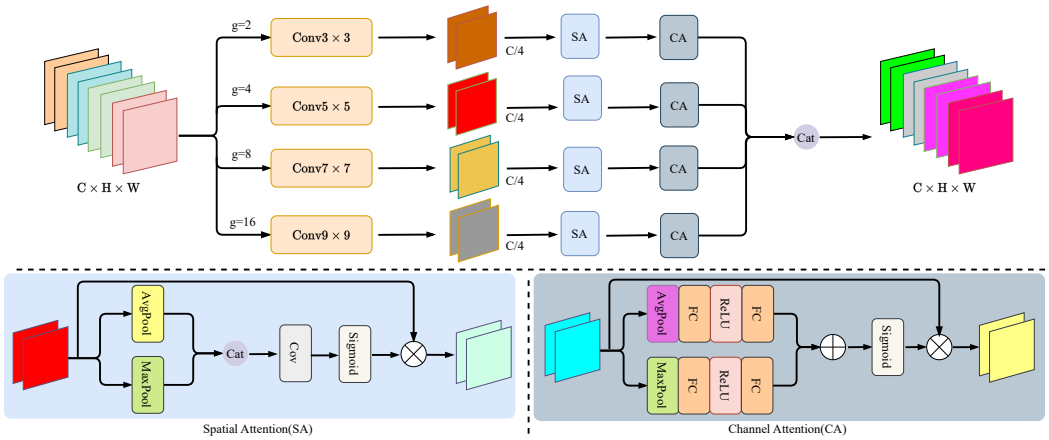

Fig. 3. The structure of the pyramid attention model.

obtained from both maximum pooling and average pooling are concatenated along the channel dimensions. This concatenation results in a comprehensive feature map that encapsulates contextual information across various scales, effectively enriching the feature representation for subsequent analysis. This feature map undergoes processing through a convolutional layer to generate spatial attention weight W_{si}. The spatial attention weight W_{si} is defined as:

$$W_{si} = \mathrm{Cat}(\mathrm{AvgPool}(f_i), \mathrm{MaxPool}(f_i)), i = 0, 1, 2, 3, \quad (5)$$

where Cat represents the concatenation operation. Then, the sigmoid activation function is employed to confine the spatial attention weights within the range of 0 to 1. These processed attention weights are then applied to the input features to derive the output features f_{si}. It is obtained as:

$$f_{si} = f_i \odot \mathrm{Sigmoid}(\mathrm{Conv}(W_{si}))), i = 0, 1, 2, 3, \quad (6)$$

where $\odot$ denotes the channel-wise multiplication [26].

Channel attention Different from spatial attention, the objective of the channel attention module is to enhance the feature representation of each channel. Similar to the spatial attention module, both maximum pooling and average pooling are utilized to compute the maximum eigenvalue and average eigenvalue for each channel. The resultant feature vectors from the maximum pooling and average pooling operations undergo processing through two fully connected layers. These layers serve the purpose of extracting attention weights specific to each channel. The combination of maximum and average features is accomplished through a summation operation, yielding the attention weight vector. This process enhances the model's ability to discern significant channel-wise information within the given context. The attention weight W_{ci} is computed as:

$$W_{ci} = \mathrm{Sigmoid}(\mathrm{FC}(\mathrm{AvgPool}(f_{si})), \mathrm{FC}(\mathrm{MaxPool}(f_{si}))), \quad (7)$$

where FC is the fully connected layer, $i = 0, 1, 2, 3$. The channel feature map is derived by multiplying the attention weights with each channel of the input feature map. The channel feature map is obtained as:

$$f_{ci} = f_{si} \odot W_{ci}, i = 0, 1, 2, 3. \quad (8)$$

Finally, the feature maps obtained after each branch has undergone the SA and CA modules yield the final feature of the Pyramid Attention module. The final feature f_{att} is obtained as:

$$f_{att} = \mathrm{Cat}(f_{c0}, f_{c1}, f_{c2}, f_{c3}). \quad (9)$$

C. Transformer encoder

In the proposed framework, a spectral transformer encoder is employed to capture the spatial relationships within fine-grained manipulation regions. Compared with the traditional transformers, the spectral transformer incorporates a spectral model on top of the original attention module. The attention module can capture global features but falls short of accurately capturing local features. Conversely, the spectral model excels in capturing local features. The proposed spectral transformer can combine the strengths of the attention layer and spectral model to capture both global attributes and local features accurately. The structure of the spectral transformer encoder is illustrated in Fig. 4.

Spectral block The spectral layer analyzes image frequencies using a spectral gating network, which includes a Fast Fourier Transform (FFT) layer, weighted gating with a trainable parameter, and an inverse FFT layer. The spectral layer converts physical space into spectral space using FFT. By employing a trainable weight parameter W_g, the assignment of weights to each frequency component ensures the precise capture of lines and edges within an image. This parameter is acquired through the implementation of back-propagation techniques. The IFFT is employed to convert the spectral space back to the physical space. Meanwhile, to acquire the localized feature f_l, the spectral block employs a residual connection to add the feature generated by IFFT to the input feature f_{pos}. The localized feature f_l is denoted as:

$$f_l = \mathrm{IFFT}(W_g \times \mathrm{FFT}(f_{pos})) + f_{pos}, \quad (10)$$

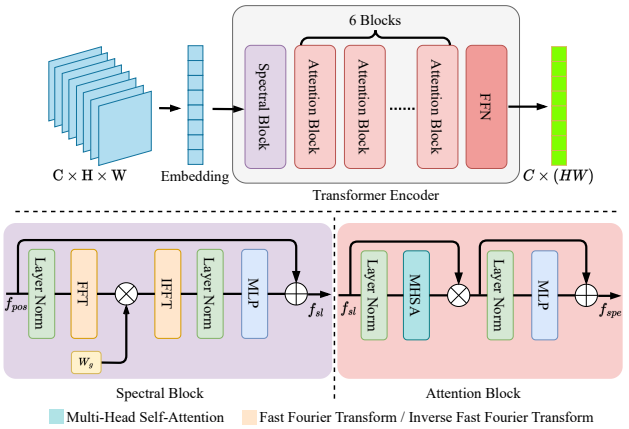

Fig. 4. The structure of the spectral transformer encoder.

where W_g is a trainable weight parameter, the feature f_{pos} is the result of complementing the feature f_{att} with a fixed positional encoding. Subsequently, a layer normalization block is introduced to facilitate stable training and improve convergence. Following the layer normalization block, the Multi-Layer Perceptron (MLP) is used for channel mixing. The output f_{sl} of the spectral block is obtained as:

$$f_{sl} = \text{MLP}(\text{LayerNorm}(f_l)) + f_l. \tag{11}$$

Attention block The attention block is a multi-head self-attention model. The objective of employing multiple attention mechanisms is to empower the model to capture long-range dependencies within the input features more effectively. Specifically, feeding the feature f_{sl} into the attention blocks, the input feature f_{sl} is segmented into multiple groups along the channel dimension. The query (Q), key (K), and value (V) are generated by the multi-head self-attention. Subsequently, the attention model employs the Q, K, and V to extract relations among all spatial positions. The multi-head self-attention can be defined as:

$$\begin{aligned} f_{spe}^{i} &= (\left(\frac{K_i^T Q_i}{\sqrt{d}}\right) V_i), \\ f_{spe} &= \text{Cat}(f_{spe}^{1}, f_{spe}^{2}, ..., f_{spe}^{N}), \end{aligned} \tag{12}$$

where N represents the number of groups into which the feature f_{sl} is partitioned, and $N = 4$. We subsequently concatenate all the groups to compose the spatial relation features f_{spe} as the output of the encoder.

D. Transformer decoder

To detect sequences of manipulations, it is essential to model the sequential relationships between features. Therefore, given the extracted features f_{spe}, the proposed framework employs an autoregressive multi-head cross-attention architecture to process these features and their corresponding manipulation sequences. The proposed method employs a spatially augmented cross-attention module to model sequence relationships with limited annotations of operational sequences.

Specifically, the STPA framework converts every manipulation in the sequence into a single token. It includes Start of Sentence (SOS) and End of Sentence (EOS) tokens at the commencement and conclusion of the sequence. By this, the tokenized sequence of manipulations S_{tok} is obtained. It is argued that each manipulation within S_{tok} corresponds to a distinct facial attribute or component, distinguished by a robust spatial region precedent. Therefore, the STPA model employs this knowledge to steer the detection of operational sequences. To this end, a Gaussian spatial weight map can be dynamically generated for each manipulation component or attribute by predicting the spatial center and scale. The Gaussian-shape spatial weight map $M(h, w)$ is obtained as:

$$\begin{aligned} (m_h, m_w) &= \text{Sigmoid}\left(\text{MLP}\left(S_{tok}\right)\right), d_h, d_w = \text{FC}\left(S_{tok}\right), \\ M(h, w) &= \exp\left(-\frac{(h-m_h)^2}{\alpha d_h^2} - \frac{(w-m_w)^2}{\alpha d_w^2}\right), \end{aligned} \tag{13}$$

where the coordinates (m_h, m_w) denote the two-dimensional spatial center points of the specific manipulation. d_h and d_w are two-dimensional parameters defining the scale of the specific manipulation. The coordinates (h, w) represent the two-dimensional parameters of the map M, and α is a hyper-parameter. In a spatial weight map, areas in proximity to the centroid receive high weights, while areas more distant from the centroid are assigned low weights. Furthermore, the spatial weight map possesses the capability to dynamically adjust its height-to-width ratio to accommodate variations in manipulation areas. This advantage facilitates the generation of spatial weight maps that are more adaptive. Finally, the produced spatial weight map M can be employed to enhance the multi-head cross-attention model. The S_{tok} serves as queries (Q), whereas the feature f_{spe} supplies both keys (K) and values (V). Subsequently, they are input into the cross-attention layer for semantic interaction. The multi-head cross-attention can be obtained as:

$$\begin{aligned} Q &= \text{FC}\left(S_{tok}\right), K, V = \text{FC}\left(f_{spe}\right), \\ f_{seqi} &= \text{Softmax}\left(K_i^T Q_i / \sqrt{d} + \log M\right) V_i, \\ f_{seq} &= \text{Cat}\left(\sum_{i=1}^{D} f_{seqi}\right), \end{aligned} \tag{14}$$

where D is the number of heads of multi-head cross-attention.

Subsequently, an autoregressive mechanism is integrated into the multi-head cross-attention module to tackle the predictive aspects of sequential manipulation. The autoregressive mechanism decodes the facial manipulation sequence by predicting the next manipulation element based on the preceding elements. The process concludes when the end marker is predicted. This approach enables the prediction of facial manipulation sequences with variable lengths as needed. Then, the final feature f_{seq} is fed into the FFN, and the sequence of facial manipulation will be generated as output. Finally, the proposed network is trained by minimizing the cross-entropy loss between each class score in the sequence and the corresponding operational annotation.

IV. Experimental results and analysis

A. Implementation details

In this work, we employed a ResNet50 backbone network to perform coarse-grained feature extraction from images. The batch size was set to 64, and the epoch was set to 150. We implemented a learning rate warm-up phase, and the epoch was set to 20. The initial learning rates were set as $1e-3$ for the transformer part and $1e-4$ for the ResNet50 part. The parameter N in Eq. (12) was set to 4. The hyperparameter α in Eq. (13) was experimentally set as $\alpha = 4$. The cross-attention heads D in Eq. (14) was set as $D = 4$. The trainable weight parameter W_g was defined as a matrix with all initial values 1. The framework was implemented in PyTorch [34] framework with 2 NVIDIA 3090 GPUs.

B. Dataset

The Sequential Deepfake (Seq-Deepfake) dataset was collected and produced by Shao *et al.* [11]. The Seq-Deepfake dataset has been meticulously curated to facilitate the detection of sequences of facial manipulations. It is divided into two distinct subsets, namely the facial components dataset and the facial attributes dataset. The two sub-datasets vary in the forgery techniques employed to create synthetic facial images. One involves facial components manipulation [35], while the other utilizes facial attributes manipulation [36]. The details of the two sub-datasets are as follows.

Facial components dataset The objective of the facial components manipulation technique is to relocate specific facial components from one individual face to another face. The original images are sourced from the CelebA-HQ [37] dataset. The fake images are produced by applying facial component masks and StyleMapGAN [35] within CelebAMask-HQ [38] to execute forgery operations on the original images. Ultimately, this sub-dataset comprises 35,166 manipulated face images labeled with 28 manipulation sequences with various lengths. Each image has the corresponding annotations for the sequential face component operations. The distribution of manipulation sequence lengths, ranging from 1 to 5, is as follows: 20.48%, 20.06%, 18.62%, 20.88%, and 19.96%.

Facial attributes dataset The method of facial attributes manipulation endeavors to alter the style of the original human face, such as changing hair color, wearing glasses, and adjusting age. The fake images are produced by inputting images from the FFHQ dataset [39] into StyleMapGAN. The facial attributes dataset includes 49,920 face images. There are 26 types of operation sequences, each with a length ranging from 1 to 5.

C. Evaluation metrics

To assess the effectiveness of the STPA model, we conducted a comprehensive analysis objectively and subjectively. For objective evaluation, Fixed Accuracy (Fixed-Acc) and Adaptive Accuracy (Adaptive-Acc) are introduced to evaluate the proposed framework.

Fixed-Acc In assessing the proposed network, we standardized the length of the predicted facial manipulation sequences to 5. In the course of training, if the predicted sequence falls short of 5, the "no manipulation" (NM) category is integrated into the annotated manipulation sequences. Then, the model compares each manipulation class in the predicted sequences with the real sequence to calculate the evaluation accuracy.

Adaptive-Acc In the proposed method, the prediction is automatically stopped when the EOS token is detected. Consequently, the proposed approach enables the discernment of sequences of facial operations with adaptable lengths. This evaluation metric is introduced to assess the performance of models in scenarios involving adaptive sequence lengths.

D. Comparison with state-of-the-art methods

To evaluate the efficacy of the proposed framework, we performed a comparative analysis against four state-of-the-art (SOTA) methods. The DRN [40], Multi-Cls [11], and DETR [41] approach sequential deepfake detection by considering each manipulation sequence as a distinct class. These can be regarded as a multi-categorization task. Nevertheless, these methods overlook the pivotal analysis of manipulated sequential data. In contrast, the SeqFakeFormer [11] and the proposed method STPA are designed for the concurrent processing of both spatial and sequential forged information. Compared with the SeqFakeFormer, the proposed method introduces pyramid attention to accentuate the facial region while neglecting the background. Additionally, to augment the capability of the transformer encoder in extracting global and local features, the proposed model employs a spectral transformer encoder to extract features from fine-grained face manipulation regions. The comparison results between the proposed method and SOAT methods are shown in Table I.

In Table I, the DRN, Multi-Cls, and DETR methods exhibit a lower level of detection performance in comparison to SeqFakeFormer and STPA methods. The proposed method STPA surpasses other SOAT methods in the detection of sequential deepfake manipulation. Specifically, in the facial components dataset, the STPA achieves scores of 72.34 and 54.20 on the Fixed-Acc and Adaptive-Acc metrics, respectively. In contrast to the SeqFakeFormer, the STPA enhances the Fixed-Acc and Adaptive-Acc metrics by 1.06% and 1.08%, respectively. In the facial attributes dataset, the proposed STPA exhibits strong performance. Compared to the SeqFakeFormer, the STPA improves the Fixed-Acc and Adaptive-Acc metrics by 0.87% and 2.02%, respectively. Nevertheless, Table I shows that the Fixed-Accuracy metric of the proposed method consistently surpasses the Adaptive-Accuracy metric. This indicates that the proposed method faces challenges in detecting sequences with adaptive length. In future work, we will study feature extraction algorithms for detecting facial manipulation sequences with adaptive length to improve detection performance.

The visualized results of detecting sequential deepfake manipulations are depicted in Fig. 5 and Fig. 6. The subjective results indicate that the proposed network proficiently discerns the sequence of deepfake manipulations with varying lengths.

TABLE I. COMPARISON WITH THE SOTA METHODS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Methods	facial components		facial attributes	
	Fixed-Acc	Adaptive-Acc	Fixed-Acc	Adaptive-Acc
DRN [40]	66.06	45.79	64.42	43.20
Multi-Cls [11]	69.65	50.57	66.66	46.00
DETR [41]	69.75	49.84	67.62	47.99
SeqFakeFormer [1] [11]	71.58	53.62	68.12	48.58
STPA (our)	**72.34**	**54.20**	**68.71**	**49.56**

TABLE II. ABLATION STUDY ON THE KEY COMPONENTS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	facial components		facial attributes	
	Fixed-Acc	Adaptive-Acc	Fixed-Acc	Adaptive-Acc
Baesline	71.58	53.62	68.12	48.58
Baesline+STM	71.83	54.09	68.65	48.72
Baesline+STM+CA	72.11	54.01	67.92	48.11
Baesline+STM+SA	71.27	53.13	68.15	48.62
Baesline+STM+SA || CA	71.68	53.49	68.15	48.62
Baesline+STM+CA+SA	71.45	53.50	67.93	48.36
Baesline+STM+SA+CA	**72.34**	**54.20**	**68.71**	**49.56**

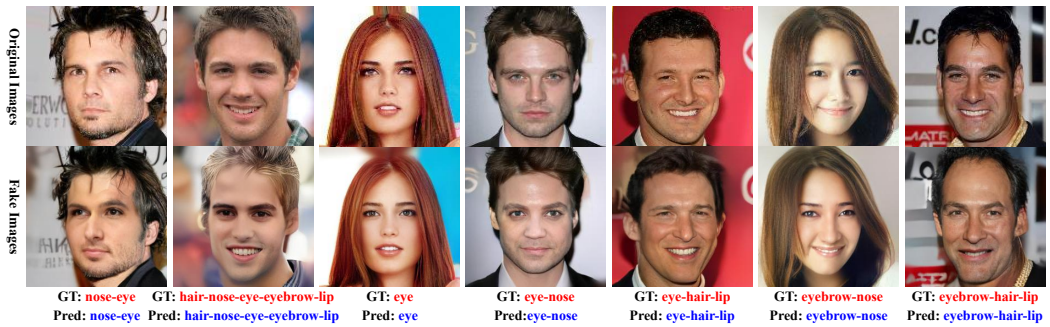

Fig. 5. The qualitative results sampled from the facial components dataset. The texts in red and blue signify the ground truth (GT) and the prediction (Pred).

E. Ablation study

The proposed model introduces a spectral transformer module and a pyramid attention module. Moreover, the channel attention module and the spatial attention module can be combined either in parallel or sequentially. To assess the effectiveness of the essential elements, the ablation experiments are conducted in the Seq-Deepfake dataset, as shown in Table II.

In the ablation studies, the Baseline is the SeqFakeFormer [11]. The STM denotes the spectral transformer module, which mixes spectral modeling and multi-attention modules. The CA denotes the channel attention module, and the SA is an attention module. The "SA || CA" represents the parallel arrangement of the spatial attention and the channel attention modules. Table II proves that the proposed framework achieves optimal performance when the SA is connected before the CA module in a serial configuration. It shows that the proposed method outperforms the baseline in detecting sequential deepfake manipulations. The visualization of the proposed pyramid attention is depicted in Fig. 7. It depicts that the pyramid attention can focus on the facial region while ignoring the background region. This facilitates the subsequent extraction of spatial relationships within the manipulated facial region.

[1]Performance evaluation was conducted with the officially released code and performed on the same platform as ours.

V. CONCLUSION

In this paper, we designed a Spectral Transformer with Pyramid Attention (STPA) network to detect the sequential deepfake manipulations in the CIoT. Compared to existing methods for sequential deepfake manipulation detection, the proposed STPA method excels in extracting the spatial features of facial manipulation regions at a finer granularity to enhance the detection accuracy of the network. The pyramid attention module is employed to prioritize attention on the facial region over the background area. Furthermore, a spectral block is incorporated into the Transformer to address its deficiency in extracting localized features. The experimental results demonstrate that STPA outperforms SOTA methods in detecting sequential deepfake manipulations.

This article has been accepted for publication in IEEE Transactions on Consumer Electronics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCE.2024.3414319

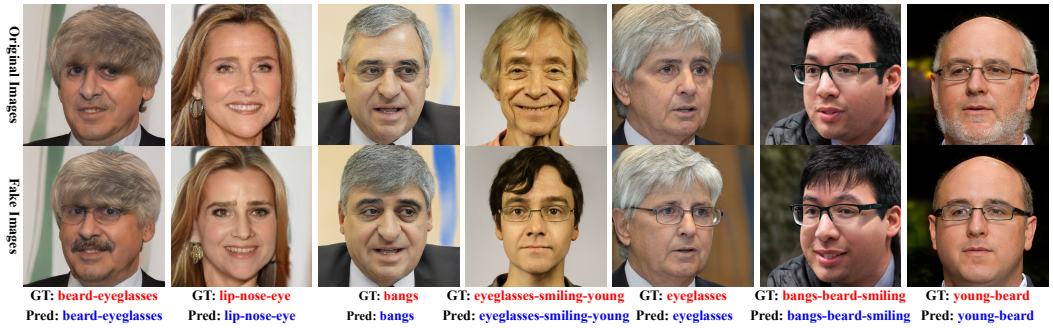

Fig. 6. The qualitative results sampled from the facial attributes dataset.

Fig. 7. Visualization of the pyramid attention map.

REFERENCES

[1] K. Sharma, A. Malik, I. Batra, A. Sanwar Hosen, M. A. Latif Sarker, and D. S. Han, "Technologies behind the smart grid and internet of things: A system survey." *Computers, Materials & Continua*, vol. 75, no. 3, 2023.

[2] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Bing, xu, david warde-farley, sherjil ozair, aaron courville, and, yoshua bengio," *Generative adversarial nets. In, NeurIPS*, vol. 1, no. 2, 2014.

[3] M. Albahar and J. Almalki, "Deepfakes: Threats and countermeasures systematic review," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 22, pp. 3242–3250, 2019.

[4] A. de Rancourt-Raymond and N. Smaili, "The unethical use of deepfakes," *Journal of Financial Crime*, vol. 30, no. 4, pp. 1066–1077, 2023.

[5] G. Zhang, M. Gao, Q. Li, W. Zhai, G. Zou, and G. Jeon, "Disrupting deepfakes via union-saliency adversarial attack," *IEEE Transactions on Consumer Electronics*, vol. 70, pp. 2018–2026, 2023.

[6] Q. Li, M. Gao, G. Zhang, W. Zhai, J. Chen, and G. Jeon, "Towards multimodal disinformation detection by vision-language knowledge interaction," *Information Fusion*, vol. 102, p. 102037, 2024.

[7] Q. Li, M. Gao, G. Zhang, and W. Zhai, "Defending deepfakes by saliency-aware attack," *IEEE Transactions on Computational Social Systems*, 2023.

[8] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2185–2194.

[9] L. Guarnera, O. Giudice, and S. Battiato, "Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models," *arXiv preprint arXiv:2303.00608*, 2023.

[10] Y. Jeong, D. Kim, Y. Ro, and J. Choi, "Frepgan: robust deepfake detection using frequency-level perturbations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 1060–1068.

[11] R. Shao, T. Wu, and Z. Liu, "Detecting and recovering sequential deepfake manipulation," in *European Conference on Computer Vision*. Springer, 2022, pp. 712–728.

[12] D. Pal, V. Vanijja, X. Zhang, and H. Thapliyal, "Exploring the antecedents of consumer electronics iot devices purchase decision: a mixed methods study," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 4, pp. 305–318, 2021.

[13] C. Swain, M. N. Sahoo, A. Satpathy, K. Muhammad, S. Bakshi, and J. J. Rodrigues, "A-dafto: Artificial cap deferred acceptance based fair task offloading in complex iot-fog networks," *IEEE Transactions on Consumer Electronics*, vol. 69, pp. 914–926, 2023.

[14] R. Montasari, "Internet of things and artificial intelligence in national security: Applications and issues," in *Countering Cyberterrorism: The Confluence of Artificial Intelligence, Cyber Forensics and Digital Policing in US and UK National Cybersecurity*. Springer, 2023, pp. 27–56.

[15] S. Xu, H. Xia, L. Xu, R. Zhang, and C. Hu, "Migan: A privacy leakage evaluation scheme for ciot-based federated learning users," *IEEE Transactions on Consumer Electronics*, vol. 70, pp. 3098–3110, 2024.

[16] E. Rabieinejad, A. Yazdinejad, A. Dehghantanha, and G. Srivastava, "Two-level privacy-preserving framework: Federated learning for attack detection in the consumer internet of things," *IEEE Transactions on Consumer Electronics*, vol. 70, pp. 4258–4265, 2024.

[17] S. Dong, J. Wang, J. Liang, H. Fan, and R. Ji, "Explaining deepfake detection by analysing image matching," in *European Conference on Computer Vision*. Springer, 2022, pp. 18–35.

[18] T. Reiss, B. Cavia, and Y. Hoshen, "Detecting deepfakes without seeing any," *arXiv preprint arXiv:2311.01458*, 2023.

[19] C. Shuai, J. Zhong, S. Wu, F. Lin, Z. Wang, Z. Ba, Z. Liu, L. Cavallaro, and K. Ren, "Locate and verify: A two-stream network for improved deepfake detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7131–7142.

[20] S. Waseem, S. A. R. S. Abu-Bakar, Z. Omar, B. A. Ahmed, S. Baloch, and A. Hafeezallah, "Multi-attention-based approach for deepfake face and expression swap detection and localization," *EURASIP Journal on Image and Video Processing*, vol. 2023, no. 1, p. 14, 2023.

[21] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational visual media*, vol. 8, no. 3, pp. 331–368, 2022.

[22] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[23] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.

[24] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[25] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[26] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[27] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3139–3148.

[28] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "Epsanet: An efficient pyramid squeeze attention block on convolutional neural network," in *Proceedings of the Asian Conference on Computer Vision*, 2022, pp. 1161–1177.

[29] B. N. Patro, V. P. Namboodiri, and V. S. Agneeswaran, "Spectformer: Frequency and attention is what you need in a vision transformer," *arXiv preprint arXiv:2304.06446*, 2023.

[30] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," *Advances in neural information processing systems*, vol. 34, pp. 980–993, 2021.

[31] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "Fnet: Mixing tokens with fourier transforms," *arXiv preprint arXiv:2105.03824*, 2021.

[32] T. Yao, Y. Pan, Y. Li, C.-W. Ngo, and T. Mei, "Wave-vit: Unifying wavelet and transformers for visual representation learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 328–345.

[33] H. M. Nguyen and R. Derakhshani, "Eyebrow recognition for identifying deepfake videos," in *2020 international conference of the biometrics special interest group (BIOSIG)*. IEEE, 2020, pp. 1–5.

[34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[35] H. Kim, Y. Choi, J. Kim, S. Yoo, and Y. Uh, "Exploiting spatial dimensions of latent in gan for real-time image editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 852–861.

[36] Y. Jiang, Z. Huang, X. Pan, C. C. Loy, and Z. Liu, "Talk-to-edit: Fine-grained facial editing via dialog," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 799–13 808.

[37] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[38] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558.

[39] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[40] S.-Y. Wang, O. Wang, A. Owens, R. Zhang, and A. A. Efros, "Detecting photoshopped faces by scripting photoshop," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 072–10 081.

[41] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.