International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02339-6

Robust Sequential DeepFake Detection

Rui Shao[1] · **Tianxing Wu**[2] · **Ziwei Liu**[2]

Received: 5 August 2023 / Accepted: 1 December 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

Abstract

Since photorealistic faces can be readily generated by facial manipulation technologies nowadays, potential malicious abuse of these technologies has drawn great concerns. Numerous deepfake detection methods are thus proposed. However, existing methods only focus on detecting *one-step* facial manipulation. As the emergence of easy-accessible facial editing applications, people can easily manipulate facial components using *multi-step* operations in a sequential manner. This new threat requires us to detect a sequence of facial manipulations, which is vital for both detecting deepfake media and recovering original faces afterwards. Motivated by this observation, we emphasize the need and propose a novel research problem called Detecting Sequential DeepFake Manipulation (Seq-DeepFake). Unlike the existing deepfake detection task only demanding a binary label prediction, detecting Seq-DeepFake manipulation requires correctly predicting a sequential vector of facial manipulation operations. To support a large-scale investigation, we construct the first Seq-DeepFake dataset, where face images are manipulated sequentially with corresponding annotations of sequential facial manipulation vectors. Based on this new dataset, we cast detecting Seq-DeepFake manipulation as a specific image-to-sequence (e.g., image captioning) task and propose a concise yet effective Seq-DeepFake Transformer (SeqFakeFormer). To better reflect real-world deepfake data distributions, we further apply various perturbations on the original Seq-DeepFake dataset and construct the more challenging Sequential DeepFake dataset with perturbations (Seq-DeepFake-P). To exploit deeper correlation between images and sequences when facing Seq-DeepFake-P, a dedicated Seq-DeepFake Transformer with Image-Sequence Reasoning (SeqFakeFormer++) is devised, which builds stronger correspondence between image-sequence pairs for more robust Seq-DeepFake detection. Moreover, we build a comprehensive benchmark and set up rigorous evaluation protocols and metrics for this new research problem. Extensive quantitative and qualitative experiments demonstrate the effectiveness of SeqFakeFormer and SeqFakeFormer++. Several valuable observations are also revealed to facilitate future research in broader deepfake detection problems. The code has been released at https://github.com/rshaojimmy/SeqDeepFake/.

Keywords DeepFake detection · Sequential facial manipulation · Multimodal Learning

1 Introduction

In recent years, hyper-realistic face images can be generated by deep generative models (Li et al., 2024; Chen et al., 2024; Shen et al., 2024) which are visually extremely indistinguishable from real images. Meanwhile, the significant improvement for image synthesis brings security concerns on potential malicious abuse of these techniques that produce misinformation and fabrication, which is known as *deepfake* (Shao et al., 2025, 2023, 2024). To address this security issue, various deepfake detection methods have been proposed to detect such forged faces. As illustrated in Fig. 1a, given the manipulated face image generated by face swap algorithm (Rossler et al., 2019) and the original face image,

Communicated by Gang Hua.

✉ Ziwei Liu
ziwei.liu@ntu.edu.sg

Rui Shao
shaorui@hit.edu.cn

Tianxing Wu
twu012@ntu.edu.sg

[1] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China

[2] S-Lab, Nanyang Technological University, Singapore, Singapore

Published online: 04 January 2025

the existing deepfake detection task requires the model to predict the correct binary labels (Real/Fake).

With the increasing popularity of easy-accessible facial editing applications, such as YouCam Makeup,[1] FaceTune2,[2] and YouCam Perfect,[3] it is convenient for people to edit face images in daily life. Compared to existing deepfake techniques mainly carrying out *one-step* facial manipulation (Rossler et al., 2019; He et al., 2021), we can now easily manipulate face images using *multi-step* operations in a *sequential* manner. As shown in Fig. 1b, the original image can be manipulated by adding eyeglasses, making a bigger smile and removing beard sequentially. This expands the scope of existing deepfake problem by adding sequential manipulation information and poses a new challenge for current *one-step* deepfake detection methods. This observation motivates us to introduce a new research problem—Detecting Sequential DeepFake Manipulation (**Seq-DeepFake**). We summarize several key differences between detecting Seq-DeepFake and the existing deepfake detection: (1) rather than only predicting binary labels (Real/Fake), detecting Seq-DeepFake aims to detect sequences of facial manipulations with diverse sequence lengths. For example, the model is required to predict a 3-length sequence as 'Eyeglasses-Smiling-Beard' for the manipulated image as shown in Fig. 1b and (2) As illustrated in Fig. 1b, beyond pure forgery detection, we can further *recover* the original faces (refer to Sect. 7.6 in Experiments) based on the detected sequences of facial manipulation in Seq-DeepFake. This greatly enriches the benefits of detecting Seq-DeepFake manipulation.

To facilitate the study of detecting Seq-DeepFake, this paper contributes the first Seq-DeepFake dataset. Figure 2 shows some samples in Seq-DeepFake dataset. From Fig. 2, it can be seen that one face image can be sequentially manipulated with different number of steps (from minimum 1 step to maximum 5 steps), which leads to facial manipulation sequences with diverse lengths. It is extremely difficult to distinguish the original and manipulated face images, and even harder to figure out the exact manipulation sequences. To make our study more comprehensive, we consider two different facial manipulation techniques, i.e., facial components manipulation (Kim et al., 2021) and facial attributes manipulation (Jiang et al., 2021), which are displayed in the first and second row, respectively in Fig. 2.

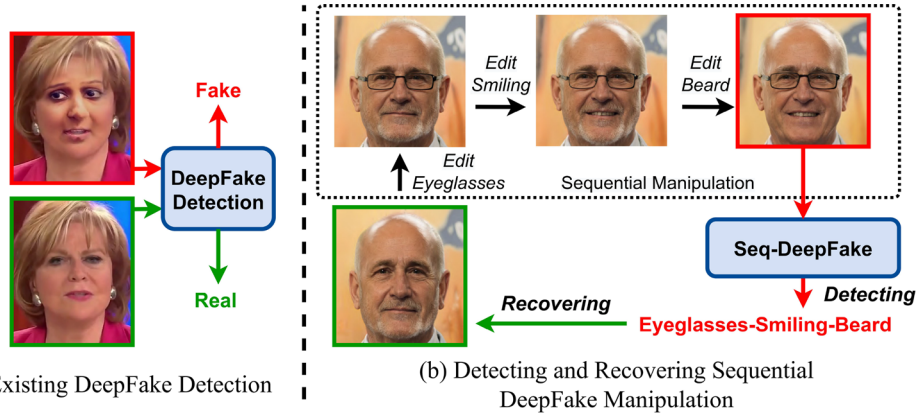
Most of current facial manipulation applications are built based on Generative Adversarial Network (GAN). It is well known that the semantic latent space learned by GAN is difficult to be perfectly disentangled (Shen et al., 2020a; Lee et al., 2020b). We argue that this defect is likely to leave some spatial as well as sequential manipulation traces unveiling sequential facial manipulations. Based on this observation, to detect such two types of manipulations traces, we cast detecting Seq-DeepFake as a specific image-to-sequence (e.g., image captioning) task and thus propose a concise yet effective Seq-DeepFake Transformer (**SeqFakeFormer**). Two key components are devised in SeqFakeFormer: **Spatial Relation Extraction** via Image Encoder and **Sequential Relation Modeling with Spatially Enhanced Cross-Attention** via Sequence Decoder. Given a manipulated image, to adaptively capture subtle spatial manipulation regions, SeqFakeFormer feeds the image into a deep convolutional neural network (CNN) to learn its feature maps. Then we extract the relation of spatial manipulations captured in feature maps using the self-attention modules of Image Encoder, obtaining features of spatial relation, i.e., spatial manipulation traces. After that, Sequence Decoder of SeqFakeFormer models the sequential relation of extracted spatial relation features via cross-attention modules in an autoregressive mechanism. This captures sequential manipulation traces, and thereby detects the facial manipulation sequences. To enable more effective cross-attention given limited annotations of facial manipulation sequences, SeqFakeFormer further integrates a Spatially Enhanced Cross-Attention (SECA) module in the decoder. This module enriches the spatial information of manipulation sequences by learning a spatial weight map. After fusing the spatial weight map with the cross-attention map, a spatially enhanced cross-attention can be achieved.

To better imitate deepfake data distributions in real-world scenarios, we further apply 6 types of perturbations, such as color distortion, multiplicative noise, random compression and blur, to the original Seq-DeepFake data at 3 intensity levels. This constructs a more challenging version of Seq-DeepFake dataset with perturbations (**Seq-DeepFake-P**). To deal with such a more difficult case, we further propose a more dedicated model dubbed Seq-DeepFake Transformer with Image-Sequence Reasoning (**SeqFakeFormer++**). Apart from the extraction of image-sequence correlation via cross-attention in SeqFakeFormer, SeqFakeFormer++ further performs deeper correlation reasoning between images and sequences via **Image-Sequence Contrastive Learning (ISC)** and **Image-Sequence Matching (ISM)**. The former reasoning module aligns the features of spatial relation and features of sequential relation by pulling those of matched image-sequence pairs close while pushing those of unmatched pairs apart. Meanwhile, the latter one achieves more fine-grained feature alignment between images and sequences by predicting whether an image-sequence pair is matched or unmatched based on aggregated multi-modal features. Aided by these two reasoning mod-

[1] https://apps.apple.com/us/app/youcam-makeup-selfie-editor/id863844475.

[2] https://apps.apple.com/us/app/facetune2-editor-by-lightricks/id1149994032.

[3] https://apps.apple.com/us/app/youcam-perfect-photo-editor/id768469908.



(a) Existing DeepFake Detection

(b) Detecting and Recovering Sequential DeepFake Manipulation

Fig. 1 Comparison between **a** existing deepfake detection and **b** proposed detecting and recovering sequential deepfake manipulation

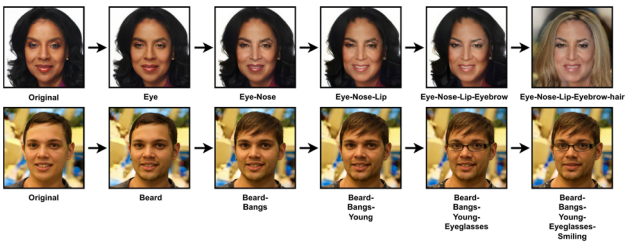


Fig. 2 Illustration of sequential facial manipulation. Two types of facial manipulation approaches are considered, i.e., facial components manipulation (Kim et al., 2021) in the first row and facial attributes manipulation (Jiang et al., 2021) in the second row

ules, more robust cross-modal correlation between images and sequences could be built and thus SeqFakeFormer++ is more able to unveil sequence manipulation traces under severe post-processing perturbations.

Main contributions of our paper can be summarized as:

- We introduce a new research problem named Detecting Sequential DeepFake Manipulation (**Seq-DeepFake**), with the objective of detecting sequences of facial manipulations, which expands the scope and poses a new challenge for deepfake detection.
- We contribute the first large-scale Seq-DeepFake dataset generating sequential manipulated face images using two different facial manipulation techniques. Corresponding annotations of manipulation sequences are provided. To mimic various noisy scenarios in real-world, we further apply a series of perturbations on manipulated face images and construct the more challenging Seq-DeepFake dataset with perturbations (**Seq-DeepFake-P**).
- We propose a powerful Seq-DeepFake Transformer (**SeqFakeFormer**). Aiming at better detecting sequence of deepfake manipulation under perturbations, a Seq-DeepFake Transformer with Image-Sequence Reasoning (**SeqFakeFormer++**) is devised, providing deeper cross-modal reasoning for more robust detection under perturbations. A comprehensive benchmark is built and rigorous evaluation protocols and metrics are designed for this novel research problem. Extensive quantitative and qualitative experiments demonstrate the superiority of SeqFakeFormer and SeqFakeFormer++.

This paper is a substantial extension of our previous ECCV 2022 work (Shao et al., 2022). We have made three major improvements in this journal version: (1) We construct a more realistic and challenging Seq-DeepFake dataset with perturbations (**Seq-DeepFake-P**). It is designed to evaluate the robustness of Seq-DeepFake detection to various perturbations in real-world; (2) To promote such robustness, a more powerful Seq-DeepFake Transformer with Image-Sequence Reasoning (**SeqFakeFormer++**) is developed. Two more reasoning modules, Image-Sequence Contrastive Learning via Sequence Encoder and Image-Sequence Matching via Image-attended Sequence Encoder, are integrated and complemented along with Sequence Decoder; (3) More visualizations corresponding to Seq-DeepFake and Seq-DeepFake-P datasets are provided. More detailed and comprehensive quantitative and qualitative experiments regarding SeqFakeFormer and SeqFakeFormer++ are performed in this journal version.

2 Related Work

2.1 DeepFake Detection

The issue of facial information security (Shao et al., 2018, 2019, 2020, 2022b, 2020, 2017, 2022a) is receiving increasing attention. Current deepfake detection methodscan be roughly categorized into spatial-based and frequency-based deepfake detection. The majority of spatial-based deepfake detection methods focus on capturing visual cues from spatial domain. Face X-ray (Li et al., 2020a) is proposed to detect the blending boundary left in the face forgery process as visual cues for real/fake detection. A multi-attentional deep-

fake detection network is proposed in Zhao et al. (2021a) to integrate low-level textural features and high-level semantic features. Zhu et al. (2021) introduce 3D decomposition into forgery detection and propose a two-stream network to fuse decomposed features for detection. Pair-wise self-consistency learning (PCL) (Zhao et al., 2021) is introduced to detect inconsistency of source features within the manipulated images. Inconsistencies in semantically high-level mouth movements are captured in Haliassos et al. (2021) by fine-tuning a temporal network pretrained on lipreading. On the other hand, some methods pay attention to the frequency domain for detecting spectrum artifacts. There exist distinct spectrum distributions and characteristics between real and fake images in the high-frequency part of Discrete Fourier Transform (DFT) (Durall et al., 2019; Dzanic et al., 2020). Qian et al. (2020) propose a F^3-Net to learn local frequency statistics based on Discrete Cosine Transform (DCT) to mine forgery. Liu et al. (2021) present a Spatial-Phase Shallow Learning method to fuse spatial image and phase spectrum for the up-sampling artifacts detection. A two-stream model is devised in Luo et al. (2021) to model the correlation between extracted high-frequency features and regular RGB features to learn generalizable features. A frequency-aware discriminative feature learning framework (Li et al., 2021b) is introduced to integrate metric learning and adaptive frequency features learning for face forgery detection.

So far, several deepfake datasets have been released to public, such as FaceForensics++ (Rossler et al., 2019), Celeb-DF (Li et al., 2020b), DeepFake Detection Challenge (DFDC) (Dolhansky et al., 2019), and DeeperForensics−1.0 (DF1.0) (Jiang et al., 2020). However, only binary labels are provided in most of existing deepfake datasets, and thus most of the above works are trained to carry out binary classification, which results in performance saturation and poor generalization.

2.2 Facial Editing

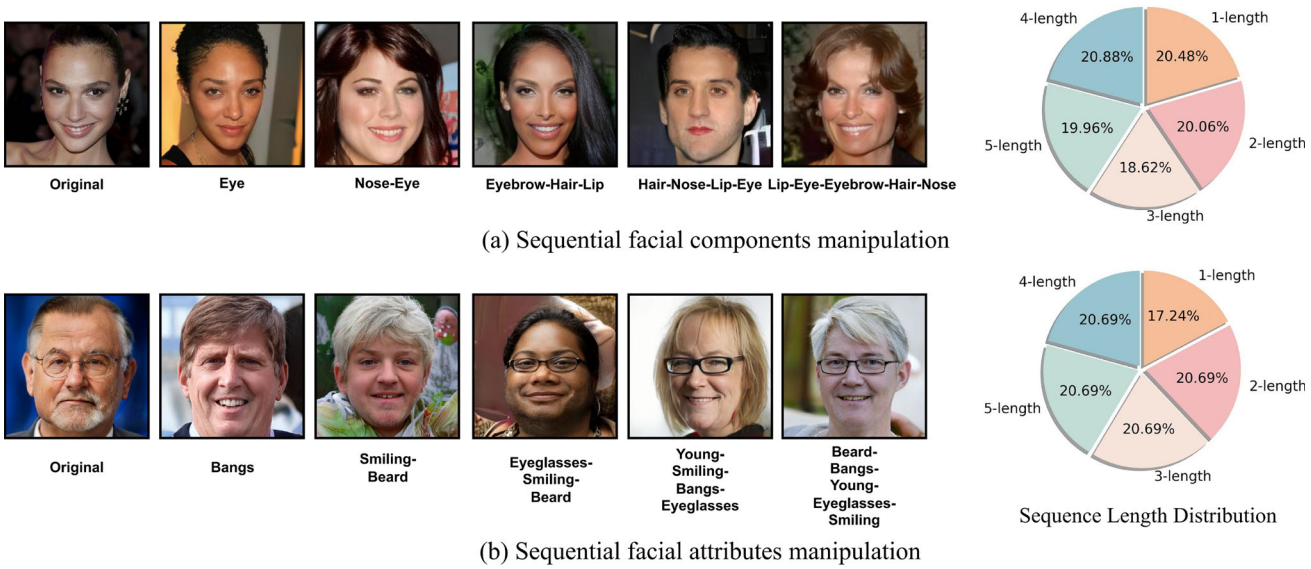
Several methods have been proposed for editing facial components (i.e. eye, nose, month). Lee et al. (2020a) present a geometry-oriented face manipulation network MaskGAN for diverse and interactive face manipulation guided by semantic masks annotations. A semantic region-adaptive normalization (SEAN) (Zhu et al., 2020) is proposed to facilitate manipulating face images by encoding images into the per-region style codes conditioned on segmentation masks. StyleMapGAN (Kim et al., 2021) introduces explicit spatial dimensions to the latent space and manipulates facial components by blending the latent spaces between reference and original faces. Moreover, some works target editing specific facial attributes such as age progression (Yang et al., 2018), and smile generation (Wang et al., 2018). Some recent works discover semantically meaningful directions in the latent space of a pretrained GAN so as to carry out facial attributes editing by moving the latent code along these directions (Shen et al., 2020a, b; Zhuang et al., 2021; Voynov & Babenko, 2020; Shen & Zhou, 2021). InterFaceGAN (Shen et al., 2020a, b) tries to disentangle attribute representations in the latent space of GANs by searching a hyperplane, of which a normal vector is used as the editing direction. Fine-grained facial attributes editing is achieved by Jiang et al. (2021) through searching a curving trajectory with respect to attribute landscapes in the latent space of GANs.

3 Sequential DeepFake Dataset

To support the novel research problem, we generate a large-scale Sequential DeepFake (Seq-DeepFake) dataset consisting of sequential manipulated face images based on two representative facial manipulation techniques, facial components manipulation (Kim et al., 2021) and facial attributes manipulation (Jiang et al., 2021). Unlike most of existing deepfake datasets (Rossler et al., 2019; He et al., 2021) only providing binary labels, the proposed dataset contains annotations of manipulation sequences with diverse sequence lengths. Details of generation pipelines based on the two facial manipulation techniques are as follows.

3.1 Sequential Facial Components Manipulation

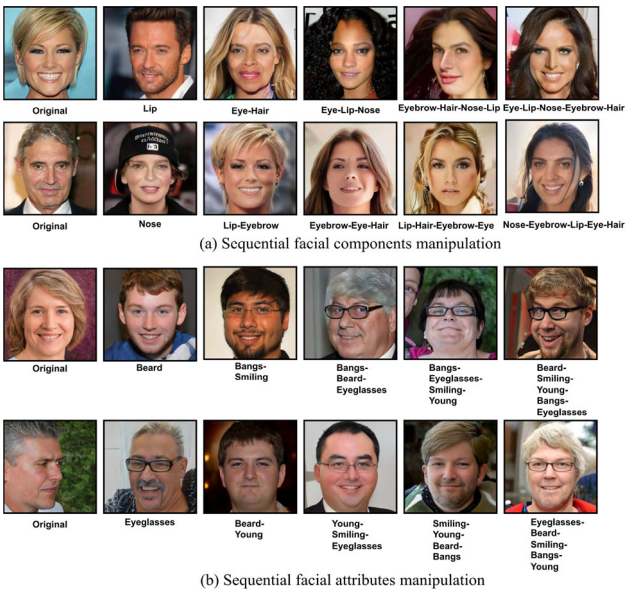
We adopt the StyleMapGAN proposed in Kim et al. (2021) for facial components manipulation. Facial components manipulation is carried out based on original images from CelebA-HQ (Liu et al., 2015; Karras et al., 2018) and corresponding facial component masks from CelebAMask-HQ (Lee et al., 2020a) dataset. Facial components manipulation aims to transplant some facial components of a reference image to an original image with respect to a mask that indicates the components to be manipulated. Specifically, we project the original image and the reference image through the encoder of StyleMapGAN to obtain stylemaps, which are intermediate latent spaces with spatial dimensions. Then, the facial components manipulation is carried out by blending the stylemaps extracted from reference and original faces based on facial component masks. Due to the inevitable appearance of degraded images in the generation process, we adopt the Generated Image Quality Assessment (GIQA) algorithm (Gu et al., 2020) to quantitatively evaluate the quality of each generated image and then filter out some low-quality ones based on the pre-defined threshold. Figure 3a shows some samples with corresponding annotations of sequential facial components manipulation. Through this data generation pipeline, we can finally generate 35,166 manipulated face images annotated with 28 types of manipulation sequences in different lengths (including original). As illustrated in Fig. 3a,



(a) Sequential facial components manipulation

(b) Sequential facial attributes manipulation

Fig. 3 Illustration of Seq-DeepFake dataset. Samples of Seq-DeepFake are provided with annotations of manipulation sequences. We also show the distribution of sequence length in Seq-DeepFake dataset



(a) Sequential facial components manipulation

(b) Sequential facial attributes manipulation

Fig. 4 More samples of Seq-DeepFake dataset. Various sequential facial manipulations are produced with diverse manipulation steps, expressions, ages, and genders

the proportions of 1–5 different lengths of manipulation sequences are: 20.48%, 20.06%, 18.62%, 20.88%, 19.96%.

3.2 Sequential Facial Attributes Manipulation

Unlike facial components manipulation methods that swap certain local parts from a reference image to an original image, facial attributes manipulation approaches directly change specific attributes on the original face image without any reference images. To take this manipulation type into consideration, we utilize the fine-grained facial editing method proposed by Jiang et al. (2021). This method aims to learn a location-specific semantic field for each editing type on the training set, then edit this attribute of interest on the given face image to a user-defined degree by stepping forward or backward on the learned curve in latent space. Based on this idea, we further generate face images with sequential facial attributes manipulation by performing the editing process in a sequential manner. Specifically, we first sample latent codes from the StyleGAN trained on FFHQ dataset (Karras et al., 2019) to generate original images. Then according to pre-defined attribute sequences, we progressively manipulate each attribute on the original face to another randomly chosen degree using the above method. After generating the final manipulation results, we also perform GIQA algorithm to filter out low-quality samples. Using this pipeline, we generate 49,920 face images with 26 manipulation sequence types, with the length of each sequence ranging from 1 to 5. Since this generation pipeline is more controllable than facial components manipulation, we construct a more balanced dataset, as shown in Fig. 3b. More samples generated by the above two types of manipulation approaches in Seq-DeepFake dataset are displayed in Fig. 4. It can be seen from Figs. 3 and 4 that various sequential facial manipulations are produced with diverse manipulation steps, expressions, ages, and genders.

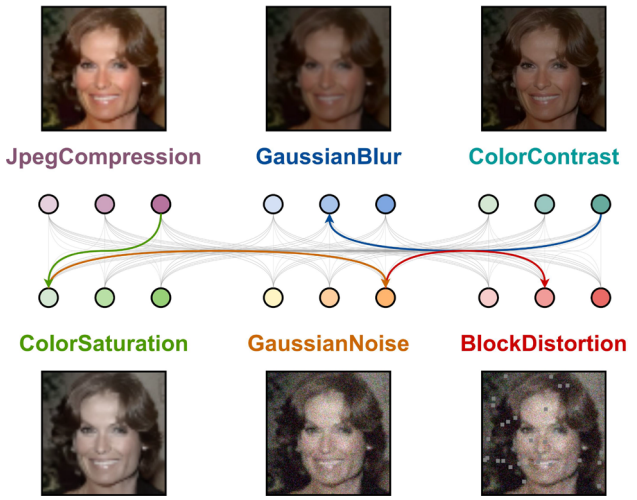
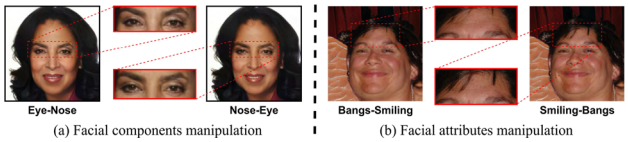


Fig. 5 Illustration of mixing process of perturbations in Seq-DeepFake-P dataset. Different perturbation types and intensity levels are marked in different colors. Arrows represent the mixture order, e.g., the image on the top-middle is first added Color Contrast Change then followed by Gaussian Blur (Color figure online)

4 Sequential DeepFake Dataset with Perturbations

In real-world scenarios, face images are likely to be affected by various post-processing like compression and noise. These post-processing operations degrade the visual quality of face images and tend to cover or alter the manipulation traces. This brings greater challenges for detecting Seq-DeepFake. To better reflect real-world deepfake data distributions under this more realistic scenario, we further impose a series of perturbations on the original Seq-DeepFake dataset and construct a more challenging Sequential DeepFake dataset with perturbations (Seq-DeepFake-P).

We depict some samples of Seq-DeepFake-P dataset in Fig. 5. Total 6 types of manipulations, e.g., JEPG Compression, White Gaussian Noise, Gaussian Blur, Color Saturation Change, Color Contrast Change, and Local Block-Wise Distortion, are applied on all the original Seq-DeepFake images. In particular, JEPG Compression mimics the information loss of images in capture and storage in the real life. White Gaussian Noise, Gaussian Blur and Local Block-Wise Distortion simulate some common visual noise appearing in low-quality images. In addition, Color Saturation Change and Color Contrast Change represent some typical color distortions in images. As illustrated in Fig. 5, 1–4 types of the above perturbations are randomly selected to be added or mixed on samples. Each perturbation is randomly applied by 1–3 intensity levels. Meanwhile, to avoid severe distribution imbalance, we guarantee each perturbation type does not co-occur for each sample. All of these combinations contribute to in total 360 distinct perturbations with different intensity levels in Seq-DeepFake-P dataset. The variety and comprehensiveness of perturbations facilities the diversity and realness of Seq-DeepFake-P dataset.



(a) Facial components manipulation (b) Facial attributes manipulation

Fig. 6 Effect of different sequential order for facial manipulation. Switching the sequential order of manipulations between **a** eye and nose and **b** bangs and smiling results in different facial manipulations

5 SeqFakeFormer

5.1 Motivation

Most of current facial manipulation applications are constructed using algorithms of Generative Adversarial Network (GAN). However, it is a well known fact that due to imperfect semantic disentanglement in the latent space of GAN (Shen et al., 2020a; Lee et al., 2020b), manipulating one facial component or attribute is likely to affect the others. As shown in the first row of Fig. 2, manipulating the nose in the step of 'Eye-Nose' simultaneously results in some little modification on the eye and mouth components compared to the former step 'Eye', which alters the overall **spatial relation** among facial components. We can thus discover some **spatial manipulation traces** from the spatial relation. Furthermore, as illustrated in Fig. 6, switching the sequential order of manipulations (e.g., manipulation order between eye and nose in (a) and bangs and smiling in (b) of Fig. 6) causes different facial manipulation results (e.g., distinct gazes in (a) and distinct amount of bangs in (b) in Fig. 6). This indicates once changing the sequential order of manipulations, the above overall spatial relation of facial components altered by manipulations will also be changed. This means there exists sequential information from spatial relation that reflects the sequential order of manipulations, which corresponds to the facial manipulation sequence. That is, we can extract the spatial relation among facial components to unveil the **spatial manipulation traces** and model their **sequential relation** to detect the facial manipulation sequence. We thus regard the sequential relation as **sequential manipulation traces**.

5.2 Overview

Based on the above observation, we cast detecting Seq-DeepFake manipulation as a specific image-to-sequence task, where inputs are manipulated/original images and outputs are facial manipulation sequences. Three challenges will be encountered when addressing this task: (1) From Figs. 2, 3 and 4, it can be seen that distinguishing original faces and

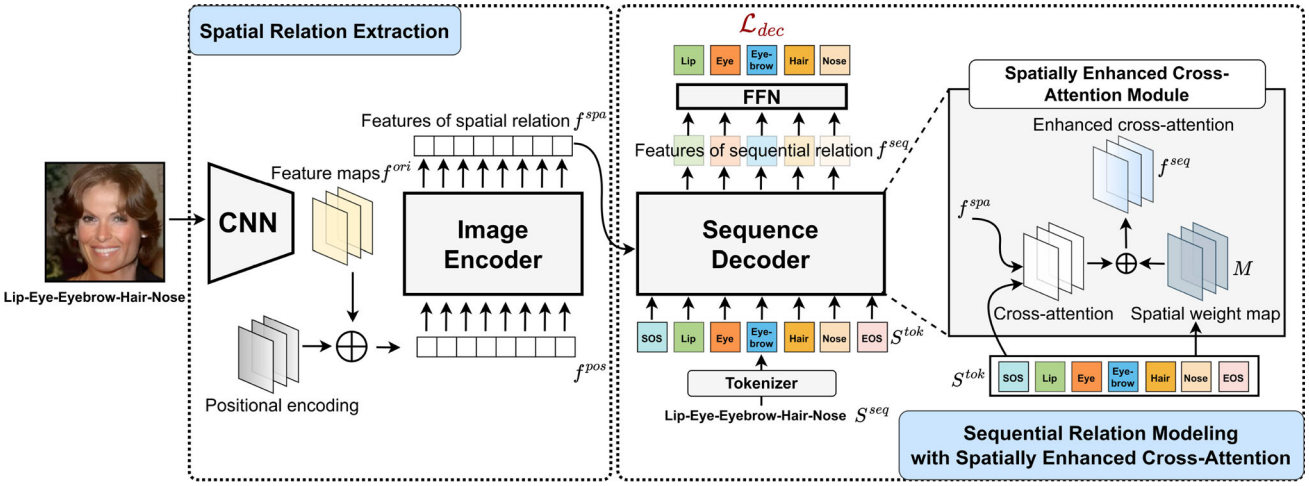


Fig. 7 Overview of proposed Seq-DeepFake Transformer (SeqFakeFormer). We first feed the face image into a CNN to learn features of spatial manipulation regions, and extract their spatial relation via self-attention modules in Image Encoder. Then we model the sequential relation of spatial relation features to detect the manipulation sequences via Sequence Decoder. A spatial enhanced cross-attention module is integrated into the decoder, contributing to a more effective cross-attention

sequential manipulated faces is extremely hard. Besides, with respect to different people, differences in face contour cause diverse manipulation regions for the same type of facial components/attributes manipulation. Thus, given indistinguishable and diverse facial manipulations, how to adaptively capture subtle manipulation regions and model their spatial relation accurately is quite challenging; (2) Based on the spatial relation of manipulated components/attributes, how to precisely model their sequential relation so as to detect the sequential facial manipulation is another challenge; (3) Compared to normal image-to-sequence task (e.g. image captions), the annotations of manipulation sequences are much shorter and thus less informative in our task. Therefore, how to effectively learn the sequential information of facial manipulations given limited manipulation sequences should also be considered.

To cope with the above three challenges, as shown in Fig. 7, we propose a Seq-DeepFake Transformer (**SeqFakeFormer**), which is composed of two key parts: **Spatial Relation Extraction** via Image Encoder, **Sequential Relation Modeling with Spatially Enhanced Cross-Attention** via Sequence Decoder. To capture spatial manipulation traces, features of subtle manipulation regions are first adaptively captured by a CNN and their spatial relation are extracted via self-attention modules in Image Encoder. After that, to capture sequential manipulation traces, we model the sequential relation of spatial relation features through cross-attention layers in Sequence Decoder with an auto-regressive mechanism. To achieve more effective cross-attention given limited annotations of manipulation sequences, a spatially enhanced cross-attention module is devised to generate the adaptive spatial weight map for each corresponding manipulation in the sequence. In following subsections, we describe all components in detail.

5.3 Spatial Relation Extraction

To adaptively capture subtle and various facial manipulation regions, we exploit a CNN to learn feature maps of the input image. Given an input image $x \in \mathrm{R}^{3\times H'\times W'}$ from Seq-DeepFake dataset U, we first feed it into a CNN (He et al., 2016) to extract its visual feature maps $f^{ori} = \mathrm{CNN}(x)$, $f^{ori} \in \mathrm{R}^{C\times H\times W}$, where H', W', and H, W are the height and width of the input image and its corresponding feature maps, respectively. C is the number of channels of feature maps.

Since the transformer architecture is permutation-invariant, we supplement original visual features maps f^{ori} with fixed positional encodings (Parmar et al., 2018; Bello et al., 2019), resulting in feature maps denoted as f^{pos}. Given that a sequence is accepted as input for transformer-based Image Encoder, we reshape the spatial dimensions of f^{pos} to one dimension, generating reshaped features $f^{pos} \in \mathrm{R}^{C\times HW}$. After fed into Image Encoder, f^{pos} conducts self-attention by generating the key, query, and value features K, Q, V so as to extract the relation among all spatial positions. Through this self-attention operation on CNN features, spatial relation of manipulation regions are exploited and thus spatial manipulation traces can be extracted. To further facilitate spatial relation extraction, this paper adopts multi-head self-attention which splits features f^{pos} into multiple groups along the channel dimension. The multi-head normalized

attention based on dot-product is as follows:

$$\begin{aligned} f_i^{spa} &= \text{Softmax}(K_i^T Q_i/\sqrt{d})V_i,\ f^{spa} \\ &= \text{Concat}(f_1^{spa}, ..., f_D^{spa}) \end{aligned} \tag{1}$$

where K_i, Q_i, V_i denote the i-th group of the key, query, and value features, d is dimension of queries and keys, and total D groups are generated. We then concatenate all the groups to form the features of spatial relation f^{spa} as the output of encoder.

5.4 Sequential Relation Modeling with Spatially Enhanced Cross-Attention

Given features of spatial relation f^{spa} extracted from the encoder, we propose to model the sequential relation among them to detect the facial manipulation sequences. To this end, we carry out cross-attention between features of spatial relation f^{spa} and corresponding annotations of manipulation sequences in an auto-regressive manner. To achieve this, we send manipulation sequences $S^{ori} \in R^{C\times N}$ (e.g., $N = 5$ in Fig. 7 before a Tokenizer) into a Tokenizer, where we transform each manipulation in the sequence into one token and insert Start of Sentence `[SOS]` and End of Sentence `[EOS]` tokens into the beginning and end of sequence. After that, we obtain tokenized manipulation sequences $S^{tok} \in R^{C\times(N+2)}$ to be cross-attended with features of spatial relation f^{spa}. Base on the auto-regressive mechanism, the decoding process of manipulation sequence in Sequence Decoder (aided by cross-attention) is triggered by `[SOS]` token and will be automatically stopped once `[EOS]` token is predicted. In this way, we can predict facial manipulation sequences with adaptive lengths.

Normally, cross-attention between tokenized sequences S^{tok} and features of spatial relation f^{spa} should be performed directly. However, as mentioned above, compared to the normal image-to-sequence task, annotations of manipulation sequences are much shorter and thus less informative (S^{tok} only has $(N + 2)$-length and maximum of N is 5). To effectively cross-attend features of spatial relation with limited manipulation sequences, inspired by Gao et al. (2021), we propose a sequential relation modeling with Spatially Enhanced Cross-Attention (SECA). We argue that each manipulation in S^{tok} corresponds to one specific facial component/attribute which has a strong prior of spatial regions, thus we can enrich the information of manipulation sequences guided by this prior. To this end, we generate the spatial weight map for each manipulation by dynamically predicting the spatial center and scale of each manipulation component/attribute in sequences by:

$$t_h, t_w = \text{sigmoid}(\text{MLP}(S^{tok})), r_h, r_w = \text{FC}(S^{tok}) \tag{2}$$

where t_h, t_w and r_h, r_w are estimated 2-dimensional coordinates corresponding to spatial centers and scales of specific manipulations in the sequences, respectively. Then the Gaussian-shape spatial weight map can be generated as:

$$M(h, w) = \exp\left(-\frac{(h - t_h)^2}{\lambda r_h^2} - \frac{(w - t_w)^2}{\lambda r_w^2}\right) \tag{3}$$

where $(h, w) \in [0, H] \times [0, W]$ are 2-dimensional coordinates of the spatial weight map M, and λ is a hyper-parameter to modulate the bandwidth of the Gaussian-shape distribution. From Eq. 3, it can be seen that spatial weight map M can assign higher importance to spatial regions closer to the centers and lower weights to locations farther from the centers. Moreover, as analyzed before, since diverse manipulation regions are presented for different people, the above dynamically learned scales can further tune the height/width ratios of spatial weight map based on each manipulation, contributing to a more adaptive spatial weight map. Based on this idea, we can enhance cross-attention between features of spatial relation and manipulation sequences with generated spatial weight map M as follows:

$$\begin{aligned} &S = \text{FC}(S^{tok}),\ K, V = \text{FC}(f^{spa}), \\ &f_i^{seq} = \text{Softmax}(K_i^T Q_i/\sqrt{d} + log M)V_i, \\ &f^{seq} = \text{Concat}(f_1^{seq}, ..., f_D^{seq}) \end{aligned} \tag{4}$$

where FC denotes a single fully-connected layer, and f_i^{seq} denotes features of sequential relation. The cross-attention of the i-th head is further element-wise added with logarithm of spatial weight map M, which contributes to spatially enhanced cross-attention. Furthermore, to model the sequential relation of facial manipulation, the auto-regressive mechanism is integrated into the above cross-attention process. This is implemented by masking out (setting to $-\infty$) all values in the input of the Softmax function in Eq. 4 which correspond to illegal connections. Through concatenation of features of sequential relation from all cross-attention heads, we can obtain the final features of sequential relation f^{seq} as the output of decoder.

The features of sequential relation are then fed into a Fast Forward Network (FFN) to be predicted as a class score for each manipulation. Finally, we jointly train the CNN, Image Encoder and Sequence Decoder by minimizing cross-entropy loss between each class score and corresponding annotation in the manipulation sequence as follows,

$$\mathcal{L}_{dec} = \mathbb{E}_{(x,S^{seq})\sim U}\mathbf{H}(\text{FFN}(f^{seq}), S^{seq}) \tag{5}$$

where $\mathbf{H}(\cdot)$ is the cross-entropy function.

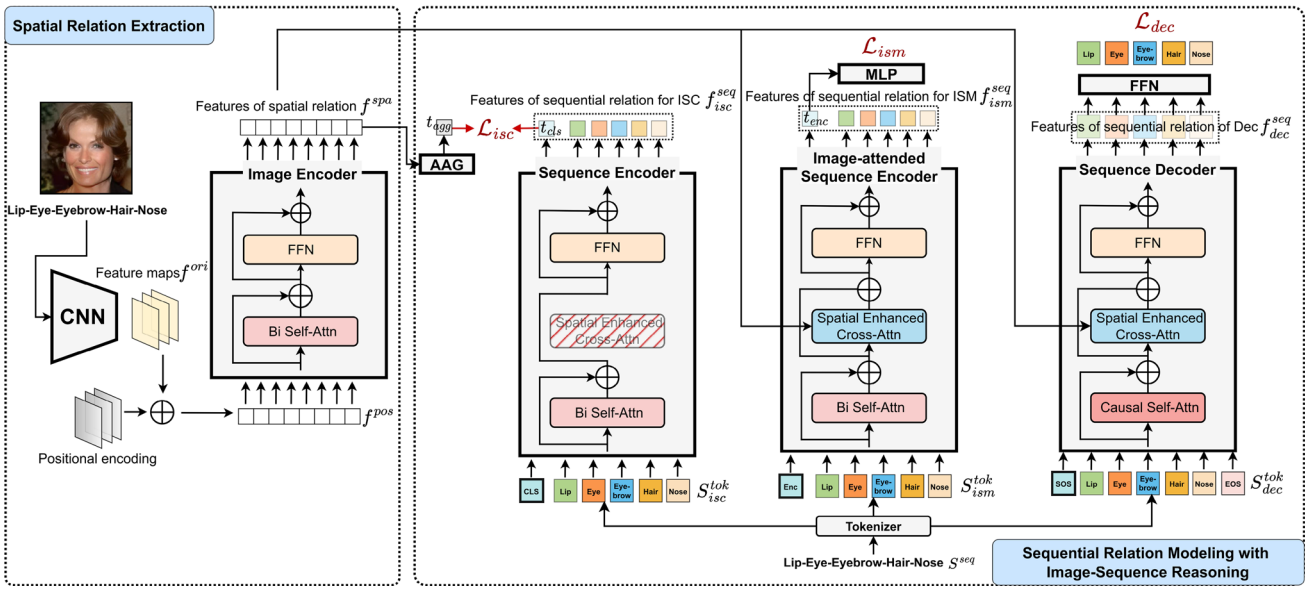


Fig. 8 Overview of proposed Seq-DeepFake Transformer with Image-Sequence Reasoning (SeqFakeFormer++). We further integrate two more Image-Sequence Reasoning modules, i.e., Sequence Encoder with Image-Sequence Contrastive Learning, and Image-attended Sequence Encoder with Image-Sequence Matching into our model. This contributes to deeper reasoning of image-sequence relation and thus more robust detection of Seq-DeepFake under various perturbations

5.5 Multi-head Version of SECA

Similar to Gao et al. (2021), we extend the basic version of Spatially Enhanced Cross-Attention (SECA) into multi-heads version, enhancing cross-attention features adaptively for different cross-attention heads. As formulated in Eq. 2, the basic version of SECA estimates 2-dimensional coordinates corresponding to spatial centers $[t_h, t_w]$. Similarly, the multi-head version of SECA estimates a head-shard spatial center $[t_h, t_w]$ and then predicts a head-specific center offset $[\triangle t_{h,i}, \triangle t_{w,i}]$ and corresponding head-specific scales $[r_{h,i}, r_{w,i}]$ for i-th cross-attention head. In this way, we generate i-th head-specific Gaussian-shape spatial weight map M_i based on the i-th head-specific center $[t_h + \triangle t_{h,i}, t_w + \triangle t_{w,i}]$ and scales $[r_{h,i}, r_{w,i}]$ as:

$$M_i(h, w) = \exp\Bigg(- \frac{(h - (t_h + \triangle t_{h,i}))^2}{\lambda r_{h,i}^2} - \frac{(w - (t_w + \triangle t_{w,i}))^2}{\lambda r_{w,i}^2} \Bigg) \tag{6}$$

Based on this, we can calculate the features of sequential relation f_i^{seq} from i-th cross-attention head enhanced by the i-th SECA as follows:

$$f_i^{seq} = \text{Softmax}(K_i^T Q_i / \sqrt{d} + log M_i) V_i, \tag{7}$$

Different from basic version of SECA, Eq. 7 shows that in the multi-head version of SECA, the cross-attention of the i-th head is element-wise added with logarithm of i-th head-specific spatial weight map M_i, which contributes to a more adaptive and specific enhanced cross-attention.

6 SeqFakeFormer++

Although SeqFakeFormer can perform well in Seq-DeepFake dataset, its performance is prone to degenerate when facing Seq-DeepFake data with perturbations in Seq-DeepFake-P. This is because SeqFakeFormer is only able to exploit the shallow relation between images and sequences merely via cross-attention between Image Encoder and Sequence Decoder. This shallow relation becomes less robust when dealing with data in Seq-DeepFake-P dataset as perturbations could cover or alter some manipulation traces. To more robustly detect sequential manipulation sequences under various post-processing perturbations, we further devise a Seq-DeepFake Transformer with Image-Sequence Reasoning (**SeqFakeFormer++**) as illustrated in Fig. 8. It complements two more Image-Sequence Reasoning modules, i.e., Sequence Encoder and Image-attended Sequence Encoder, to explore deeper image-sequence relation. This strengthens the correspondence between image-sequence pairs and thus improves the robustness of Seq-DeepFake detection to various noises.

6.1 Image-Sequence Contrastive Learning

In essence, feature distances between images and corresponding manipulation sequences should closer to those of unmatched pairs. To make our model more robust to perturbations, this image-sequence correlation should be further strengthened. To this end, we propose Image-Sequence Contrastive Learning to align the uni-modal feature representations of images and sequences. As depicted in Fig. 8, we exploit the features of spatial relation f^{spa} for the uni-modal feature representations of images. Specifically, to adaptively aggregate all the feature tokens in f^{spa}, we propose Attentional Aggregation (**AAG**) to learn a global feature representation for the whole image via an attention mechanism. To be specific, **AAG** is performed by cross-attending a t_{agg} token with all the tokens in f^{spa}:

$$t_{agg} = \mathbf{AAG}(f^{spa}) = \text{Attention}(t_{agg}, f^{spa}, f^{spa}) \tag{8}$$

where $\text{Attention}(Q, K, V) = \text{Softmax}(K^T Q/\sqrt{D})V$, and t_{agg} serves as the feature representation of images.

As for learning uni-modal feature representations of sequences, we add one more Sequence Encoder as shown in Fig. 8, where Bi-directional Self-Attention layers are built without Spatial Enhanced Cross-Attention layers (SECA). We send original annotations of manipulation sequences S^{ori} into the Tokenizer which appends a `[CLS]` token into the beginning of the sequence input to summarize its information. This generates a tokenized manipulation sequence for Image-Sequence Contrastive Learning denoted as S^{tok}_{isc}. We feed it into Sequence Encoder and obtain the corresponding features of sequential relation for ISC f^{seq}_{isc}, where the first feature token is t_{cls} serving as the feature representation of whole sequence. Based on the feature representation of images t_{agg} and sequences t_{cls}, we perform ISC by pulling features of matched image-sequence pairs close while pushing those of unmatched pairs apart. Following InfoNCE loss (Oord et al., 2018), we formulate image-to-sequence contrastive loss by:

$$\mathcal{L}_{i2s}(I, S^+, S^-) = -\mathbb{E}_{p(I,S)}\left[\log\frac{\exp(F(I, S^+)/\tau)}{\sum_{k=1}^{K}\exp(F(I, S_k^-)/\tau)}\right] \tag{9}$$

where τ is a temperature hyper-parameter, $S^- = \{S_1^-, ..., S_K^-\}$ is a set of negative sequence samples that are not matched to I in the current batch (K is batch size). We use two projection heads h_i and h_s to map feature representations of two modalities to a lower-dimensional (256) embedding space for similarity calculation: $F(I, S) = h_i(t_{agg})^{\mathrm{T}} h_s(t_{cls})$. Similarly, sequence-to-image contrastive loss is as follows:

$$\mathcal{L}_{s2i}(S, I^+, I^-) = -\mathbb{E}_{p(I,S)}\left[\log\frac{\exp(F(S, I^+)/\tau)}{\sum_{k=1}^{K}\exp(F(S, I_k^-)/\tau)}\right] \tag{10}$$

where $I^- = \{I_1^-, ..., I_K^-\}$ is a queue of K negative image samples that are not matched to S. We incorporate the above two losses and form Image-Sequence Contrastive Loss as:

$$\mathcal{L}_{isc} = \frac{1}{2}[\mathcal{L}_{i2s}(I, S^+, S^-) + \mathcal{L}_{s2i}(S, I^+, I^-)] \tag{11}$$

6.2 Image-Sequence Matching

To capture more fine-grained alignment between images and sequences, we further interact image features with sequence features to learn image-sequence multi-modal representation. In this regard, we integrate one more Image-attended Sequence Encoder as displayed in Fig. 8, where interaction of cross-modal information is performed by inserting SECA layers the same as those of Sequence Decoder. To explore the correspondence regarding global feature representation between images and sequences, unlike auto-regressive mechanism performed in Causal Self-attention layers of Sequence Decoder, Bi-directional Self-Attention layers are adopted in Image-attended Sequence Encoder. Notably, a task-specific `[Enc]` token is appended to the beginning of original sequence, which plays the role of the information summary of whole input sequence S^{tok}_{ism}. After cross-modal interaction through Image-attended Sequence Encoder, a token t_{enc} can be generated in the output feature f^{feq}_{ism} which corresponds to the input `[Enc]` token. In this way, t_{enc} summarizes the information of cross-modal interaction and thus can be used as the multi-modal representation of each image-sequence pair. Based on it, we perform ISM by predicting binary classes to determine whether the input image-sequence pair is matched or unmatched. To achieve this, we feed t_{enc} into a Multilayer Perceptron (MLP) and calculate Image-Sequence Matching Loss as follows,

$$\mathcal{L}_{ism} = \mathbb{E}_{(I,S)\sim P}\mathbf{H}(\text{MLP}(t_{enc}), y_{mat}) \tag{12}$$

where $y_{mat} \in \{0, 1\}$ describes whether the input image-sequence pair is matched or not. In order to find more informative unmatched samples, inspired by Li et al. (2021a), we deploy the hard negative mining strategy based on contrastive similarity calculated in Image-Sequence Contrastive Learning. The hardest unmatched samples are selected based on the highest contrastive similarity in a batch.

Based on the aforementioned analysis, we formulate the above objectives into a unified joint optimization model for SeqFakeFormer++ as follows:

$$\mathcal{L} = \mathcal{L}_{isc} + \mathcal{L}_{ism} + \mathcal{L}_{dec} \tag{13}$$

Since SECA and FFN have similar functions in sequence encoder and decoder, we thus share the parameters of these layers when we perform the above joint optimization. In this way, the benefit of deeper Image-Sequence Reasoning brought by ISC and ISM directly contributes to better SECA and FFN layers in Sequence Decoder. This consequently facilitates the auto-regressive sequence decoding carried out in Causal Self-attention layers in Sequence Decoder. All these modules together contribute to a more robust and accurate detection of Seq-DeepFake manipulation sequences even facing perturbations.

7 Experiments

7.1 Experimental Setup

Implementation Details We choose two different CNNs, ResNet-34 (He et al., 2016) and ResNet-50 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) dataset in our paper. To be comparable in the number of parameters, we adopt a transformer model with 2 encoder and 2 decoder layers with 4 attention heads. For the training schedule, we employ 20 epochs warm-up strategy and train for 170 epochs with a learning rate drop to 10% in every 50 epochs. The initial learning rates are set as $1e-3$ for transformer part and $1e-4$ for CNN part. We set $\lambda = 4$.

Implementation is in PyTorch. For the training schedule, we employ a 20-epochs warm-up strategy. The initial learning rate is set as $1e-3$ for transformer part and $1e-4$ for CNN part, with a decay factor of 10 at 70 and 120 epochs, totally 170 epochs. We use the SGD momentum optimizer with weight decay set as $1e-4$. We use a mini-batch size of 32 per GPU and 4 GPUs in total. Model selection for evaluation is conducted by considering the trained model that has produced the best accuracy on the validation set.

Baseline Methods The most straightforward solution for detecting Seq-DeepFake manipulation is to regard it as a multi-label classification problem (Wang et al., 2021). It treats all manipulations in the sequences as independent classes and classifies the manipulated images into multiple manipulation classes. Specifically, we design a simple multi-label classification network (denoted as **Multi-Cls**) as one of the baselines. We use ResNet-34 (He et al., 2016) and ResNet-50 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) dataset as backbones for the multi-label classification network, which is concatenated with N single linear-layer branches as N classification heads ($N = 5$ as maximum manipulation steps are 5 in Seq-DeepFake dataset). Moreover, we study a more complex transformer structure modified for our problem. **DETR** (Carion et al., 2020) is a popular transformer devised for end-to-end object detection. This model detects input images' bounding boxes and corresponding object classes conditioned on object queries. We revise this model by replacing the object queries with annotations of manipulation sequences and only preserve the output of object classes.

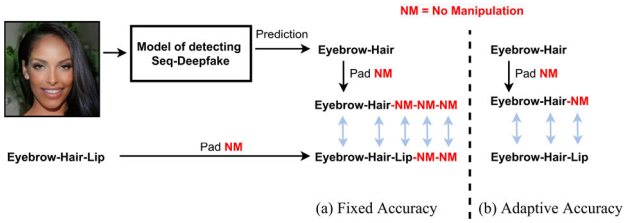


Fig. 9 Comparison between two evaluation metrics **a** fixed accuracy and **b** adaptive accuracy

Furthermore, to examine the performance of existing deepfake detection methods for our research problem, we adapt three state-of-the-art deepfake detection methods, a Dilated Residual Network variant (**DRN**) (Wang et al., 2019), a two-stream network modeling the correlation between high-frequency features and regular RGB features (**TS**) (Luo et al., 2021), and a multi-attentional deepfake detection (**MA**) (Zhao et al., 2021a), into multi-label classification setting. To be specific, we replace their binary label classifier with multiple classification heads to classify sequential manipulations. Please note since all of the above baselines are only able to predict the facial manipulation with fixed length ($N = 5$), 'no manipulation' class will be padded into the annotation sequence if its length is shorter than N so that we can keep the same length between predictions and annotation sequences for training.

Evaluation Metrics We propose two evaluation metrics for this new task.

– **Fixed Accuracy (Fixed-Acc)** Since all the baseline methods are only able to predict the facial manipulation with fixed length ($N = 5$), 'no manipulation' class will be padded into the annotation sequence if its length is shorter than N so that we can keep the same length between predictions and annotation sequences for training. Following this strategy, as shown in Fig. 9, under the evaluation metric of Fixed Accuracy, given the model prediction, such as 'Eyebrow-Hair', we first pad 'no manipulation' class into it to form the padded prediction sequence as 'Eyebrow-Hair-NM-NM-NM' (NM means 'no manipulation' class), generating the predic-

Table 1 Accuracy of detecting Seq-DeepFake based on sequential facial components manipulation

Methods	ResNet-34		ResNet-50	
	Fixed-Acc	Adap-Acc	Fixed-Acc	Adap-Acc
Multi-Cls	69.66	50.54	69.65	50.57
DETR (Carion et al., 2020)	69.87	50.63	69.75	49.84
Ours	72.13	54.80	72.65	55.30
Ours++	**72.66**	**55.16**	**72.81**	**56.12**

The bold value denotes the best results

Table 2 Accuracy of detecting Seq-DeepFake based on sequential facial attributes manipulation

Methods	ResNet-34		ResNet-50	
	Fixed-Acc	Adap-Acc	Fixed-Acc	Adap-Acc
Multi-Cls	66.99	46.68	66.66	46.00
DETR (Carion et al., 2020)	67.93	48.15	67.62	47.99
Ours	67.99	48.32	68.86	49.63
Ours++	**68.74**	**49.51**	**69.03**	**49.92**

The bold value denotes the best results

Table 3 Accuracy of detecting Seq-DeepFake compared to deepfake detection methods

Methods	Face Components Manipulation		Face Attributes Manipulation	
	Fixed-Acc	Adap-Acc	Fixed-Acc	Adap-Acc
DRN (Wang et al., 2019)	66.06	45.79	64.42	43.20
MA (Zhao et al., 2021a)	71.31	52.94	67.58	47.48
TS (Luo et al., 2021)	71.92	53.89	66.77	46.38
Ours	72.65	55.30	68.86	49.63
Ours++	**72.81**	**56.12**	**69.03**	**49.92**

The bold value denotes the best results

tion with fixed N-length ($N = 5$). To keep the same length between predictions and annotation sequences for evaluation, we pad 'no manipulation' class into the annotation of manipulation sequences as well, denoted as 'Eyebrow-Hair-Lip-NM-NM'. Given this, we compare each manipulation class in the predicted sequences with its corresponding annotation to calculate the evaluation accuracy.

– **Adaptive Accuracy (Adaptive-Acc/Adap-Acc):** Moreover, since the proposed method detects facial manipulation sequences in auto-regressive mechanism, detection will be automatically stopped once predicting the `[EOS]` token. As such, the proposed method can detect facial manipulation sequences with adaptive lengths. To conduct the evaluation in this scenario, as illustrated in Fig. 9, the second type of evaluation is devised, which compares predicted manipulations and corresponding annotations within the maximum steps of manipulations ($N = 3$ in Fig. 9 and we just pad one 'no manipulation' class into prediction sequence) between them. This makes the evaluation more concentrated on the accuracy of manipulations.

7.2 Benchmark for Seq-DeepFake

We tabulate the first benchmark for detecting sequential facial manipulation based on facial components manipulation and facial attributes manipulation in Tables 1, 2 and 3. We note that, both baselines and the proposed method yield much higher performance under evaluation metric Fixed-Acc than Adaptive-Acc. This validates that detecting sequential facial manipulation with adaptive lengths is much harder than its simplified version with fixed length. It can be observed from Tables 1 and 2, that the proposed SeqFakeFormer obtains the best performance of detecting facial manipulation sequences compared to all considered baselines in both facial components manipulation and facial attributes manipulation. In addition, SeqFakeFormer also performs better than other baselines with both CNNs (ResNet-34 and ResNet-50), indicating the compatibility of the proposed method with different feature extractors. Specifically, the proposed method has achieved about 3–4% improvement in facial components sequential manipulation and 1–2% improvement in facial attributes sequential manipulation under two evaluation metrics. In particular, there exists a larger performance gap between SeqFakeFormer and other

baselines under evaluation metric Adaptive-Acc than Fixed-Acc, suggesting SeqFakeFormer is more effective in the more challenging case.

In addition, we tabulate the comparison between three SOTA deepfake detection methods and our method in Table 3. SeqFakeFormer also outperforms all SOTA deepfake detection methods in both manipulation types. Since all the baselines treat detecting Seq-DeepFake as a multi-label classification problem, only spatial information of manipulated images are extracted. In contrast, SeqFakeFormer is capable of exploiting both spatial and sequential manipulation traces and thus more useful sequential information can be modeled, which is the key to enhance the performance of Seq-DeepFake Detection.

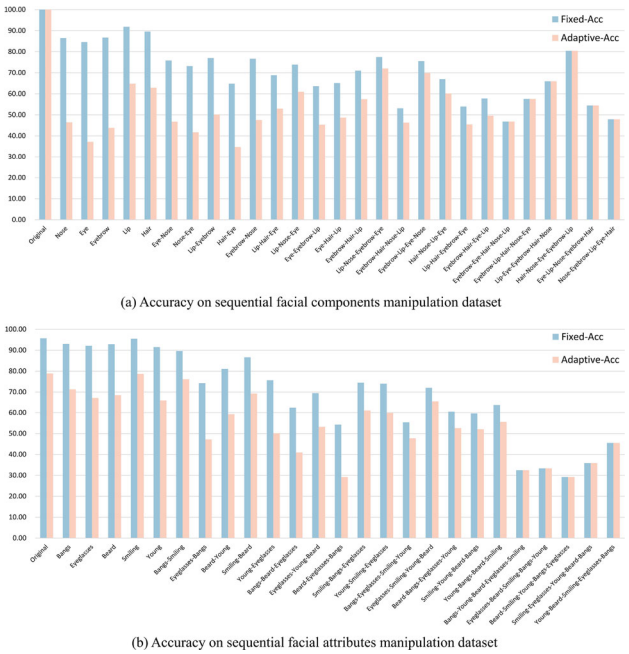
Furthermore, it can be seen from Tables 1, 2 and 3 that the upgraded SeqFakeFormer++ (Ours++) also exceeds other baselines and further improves the performance of SeqFakeFormer under all settings. This demonstrates the effectiveness of Seq-FakeFormer++ on original Seq-DeepFake dataset without perturbations.

7.3 Ablation Study

In this sub-section we investigate the impact of two key components in SeqFakeFormer, auto-regressive mechanism and Spatially Enhanced Cross-Attention module (SECA), to the overall performance. The considered components and the corresponding results obtained for each case are tabulated Tables 4 and 5. As evident from Tables 4 and 5, removing either auto-regressive mechanism or SECA will degrade the overall performance. This validates that auto-regressive mechanism facilitates the sequential relation modeling and SECA benefits the cross-attention. These components complement each other to produce better performance for detecting Seq-DeepFake.

7.4 Accuracy for Each Manipulation Sequence

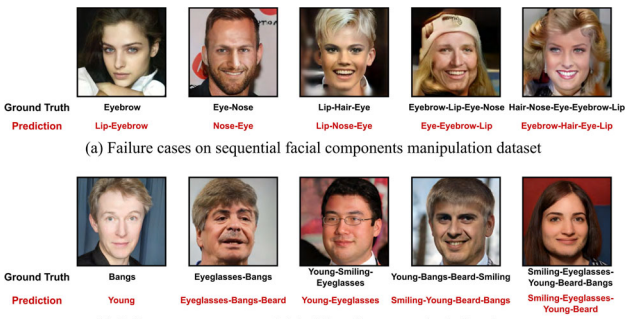
As mentioned in Sect. 3, we generate 28 types of manipulation sequences based on facial components manipulation while 26 types of manipulation sequences based on facial attributes manipulation. To provide a more detailed analysis, in this section, we plot accuracy for each manipulation sequence in both facial manipulation methods as shown in Fig. 10. It can be observed that diverse accuracy performance are achieved for different manipulation sequences, ranging from 46.81 to 100% under Fixed-Acc and 34.69 to 100% under Adaptive-Acc in sequential facial components manipulation, while ranging from 29.25 to 95.75% under Fixed-Acc and 29.21 to 78.88% under Adaptive-Acc in sequential facial attributes manipulation. This indicates a large portion of manipulation sequences are difficult to be detected and there exist some extremely hard cases. Therefore, we should further improve our method to cope with all types of manipulation sequences in the future. Furthermore, it can be seen from Fig. 10 that the accuracy gap between two evaluation metrics, Fixed-Acc and Adaptive-Acc, decreases along with the length of sequence increases. This is because the padded 'no manipulation' class is fewer in the longer manipulation sequence when evaluating under Adaptive-Acc, which is closer to the evaluation under Fixed-Acc.



(a) Accuracy on sequential facial components manipulation dataset

(b) Accuracy on sequential facial attributes manipulation dataset

Fig. 10 Accuracy for each manipulation sequence. Detection of a large portion of manipulation sequences is challenging and some extremely hard cases exist



(a) Failure cases on sequential facial components manipulation dataset

(b) Failure cases on sequential facial attributes manipulation dataset

Fig. 11 Examples of failure cases. Successful Seq-DeepFake detection is non-trivial as it demands the correct prediction in terms of manipulation types, orders and lengths simultaneously from hyper-realistic face images with subtle manipulation traces

7.5 Visualization of Attention Maps

To demonstrate SeqFakeFormer++ detects the manipulation operations based on their sequence semantics, following

Table 4 Ablation study of detecting Seq-DeepFake based on sequential facial components manipulation

Components		ResNet-34		ResNet-50	
Auto-reg	SECA	Fixed-Acc	Adap-Acc	Fixed-Acc	Adap-Acc
✗	✗	70.64	52.19	71.22	53.43
✗	✓	70.77	51.71	70.99	52.66
✓	✗	71.88	53.84	72.18	54.64
✓	✓	**72.13**	**54.80**	**72.65**	**55.30**

The bold value denotes the best results

Table 5 Ablation study of detecting Seq-DeepFake based on sequential facial attributes manipulation

Components		ResNet-34		ResNet-50	
Auto-reg	SECA	Fixed-Acc	Adap-Acc	Fixed-Acc	Adap-Acc
✗	✗	66.98	45.87	68.14	48.49
✗	✓	67.36	47.22	68.77	49.54
✓	✗	66.70	46.56	68.17	48.81
✓	✓	**67.99**	**48.32**	**68.86**	**49.63**

The bold value denotes the best results

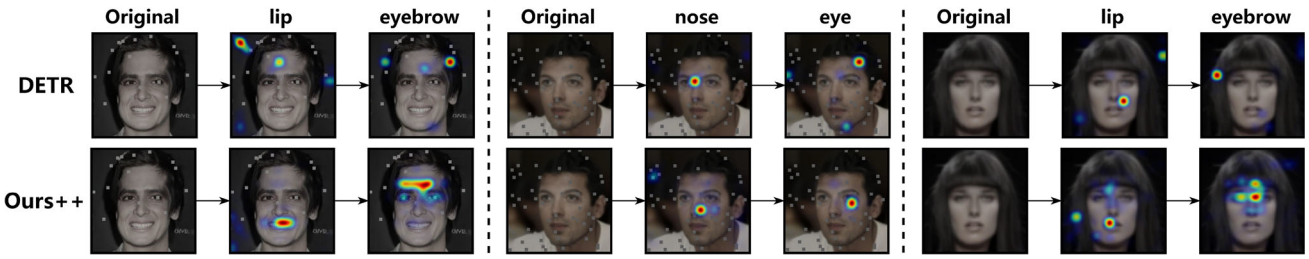


Fig. 12 Grad-CAM visualizations on the cross-attention maps corresponding to manipulation operations

ALBEF (Li et al., 2021a), we compute and visualize Grad-CAM (Selvaraju et al., 2017) on the cross-attention maps in the first layer of Sequence Decoder for each manipulation operation in the sequence. We show some samples in Fig. 12 as follows. As displayed in this figure, SeqFakeFormer++ is able to detect manipulation operations by attending to semantically relevant facial regions, such as lip, nose, and eyebrow. In contrast, DETR randomly pays attention to some semantically irrelevant facial regions during the detection process. This indicates that SeqFakeFormer++ is more able to capture sequence semantics when detecting each manipulation operation, suggesting the feasibility of this application.

7.6 Face Recovery

After detecting facial manipulation sequences, we are able to perform more challenging tasks, like recovering the original face from the manipulated face image. Specifically, we formulate the Face Recovery task as: given a sequentially manipulated face image, reverse the manipulation process

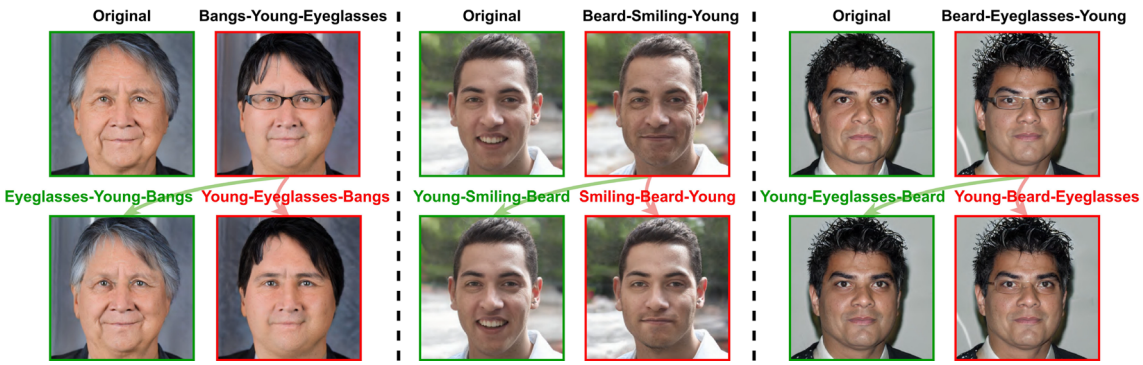


Fig. 13 Face recovery based on correct and wrong facial manipulation sequences

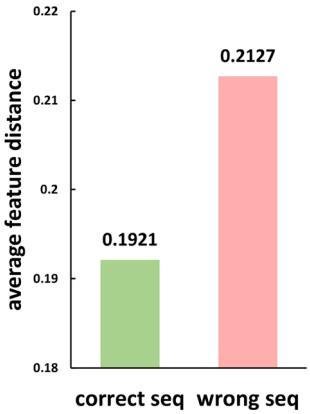


Fig. 14 Identity preservation

to get an image as close as possible to the original image. For example, in the facial attributes manipulation case, given an image generated by sequential manipulations on different attributes on the original face, we want to recover the original image. In fact, this task can be seen as an inverse sequential facial attribute manipulation problem, which can be effectively solved by the data generation pipeline described in Sect. 3 in an inverse manner. Specifically, as can be observed in Fig. 13, once we detect the correct facial manipulation sequence, i.e., correct manipulations ordered with correct manipulation steps, we can recover the original face by performing face attribute manipulation based on the inverse order of the detected sequence (process with green arrow). Comparatively, recovering the face image with wrongly ordered manipulation sequences may encounter different problems, such as incomplete recovery of age, smile, glasses, and *etc*. (process with red arrow). Figure 14 evaluates the results using identity preservation metrics as in Jiang et al. (2021), where smaller feature distance means identity is better preserved. The average feature distance between randomly selected 100 original faces and recovered faces using correct manipulation sequences is clearly smaller than that of the wrongly ordered sequence, implying the identity can be better recovered with correct manipulation sequence. Based on the above analysis and experiments, we prove that the detection of facial manipulation sequences is highly useful for face recovery, and we hope it can be applied to more meaningful tasks in the future.

7.7 Failure Cases

To provide a deeper understanding for our novel task and method, we display some failure cases produced by the proposed method as illustrated in Fig. 11. From Fig. 11, it can be seen that there exist diverse failure cases, including wrong predictions with respect to manipulation type, sequence order, sequence length, and *etc*. This validates that it is quite challenging for our novel research problem as we need to detect facial manipulation sequences in terms of correct manipulation types, orders and lengths simultaneously from hyper-realistic face images with subtle manipulation traces. This motivates us to continually develop detection methods to tackle such a novel yet challenging research problem.

8 Benchmark for Seq-DeepFake-P

We tabulate the benchmark for detecting sequential facial manipulation with perturbations (Seq-DeepFake-P) based on facial components manipulation and facial attributes manipulation in Tables 6, 7 and 8. Compared to benchmark results regarding Seq-DeepFake dataset in Tables 1, 2 and 3, Tables 6, 7 and 8 show that the performance obtained by all the baselines and proposed method degrades when dealing with Seq-DeepFake-P, especially in the detection of sequential facial components manipulation. This demonstrates it is much more challenging to detect Seq-DeepFake in the more realistic scenario under various perturbations.

Facing such a harder scenario, it can be seen from Tables 6 and 7 that the proposed SeqFakeFormer++ (Ours++) obtains the best detection performance for both facial components manipulation and facial attributes manipulation. It surpasses all considered baselines by 1–4% under two evaluation metrics. Particularly, we can clearly observe that SeqFakeFormer++ substantially outperforms SeqFakeFormer (Ours) in Tables 6 and 7. This implies the efficacy of two supplemented image-sequence reasoning modules for tackling Seq-DeepFake-P data. In addition, we further incorporate one more image feature extraction backbone, ResNet-18, into the proposed framework, as shown in Tables 6 and 7. It can be seen that SeqFakeFormer and SeqFakeFormer++ with ResNet-18 (He et al., 2016) backbone also yield better performance compared to other baselines, presenting similar trends as the other backbones. This further indicates the versatility of the proposed method that SeqFakeFormer and SeqFakeFormer++ are compatible with various image feature extraction backbones.

Meanwhile, we tabulate the comparison with three SOTA deepfake detection methods, including our conference version, in Table 8. SeqFakeFormer++ also exceeds all SOTA deepfake detection methods in both sequential manipulation detection tasks. Notably, SeqFakeFormer++ is able to improve the performance with respect to Adaptive-Acc by at least 3% compared to these three SOTA deepfake detection methods.

To further validate the superiority of the proposed method over some relevant multimodal learning methods, we finetune representative multimodal models such as BLIP (Li et al., 2022) and ALBEF (Li et al., 2021a) with Seq-DeepFake-

Table 6 Accuracy of detecting Seq-DeepFake-P based on sequential facial components manipulation with perturbations

Methods	ResNet-18		ResNet-34		RestNet-50	
	Fixed-Acc	Adap-Acc	Fixed-Acc	Adap-Acc	Fixed-Acc	Adap-Acc
Multi-Cls	60.90	40.44	63.45	40.90	64.58	42.21
DETR (Carion et al., 2020)	60.86	38.81	63.13	40.78	61.17	38.19
Ours (Shao et al., 2022)	60.97	40.90	63.51	42.01	65.19	44.34
Ours++	**63.28**	**42.56**	**65.23**	**44.45**	**66.11**	**45.35**

The bold value denotes the best results

Table 7 Accuracy of detecting Seq-DeepFake-P based on sequential facial attributes manipulation with perturbations

Methods	ResNet-18		ResNet-34		RestNet-50	
	Fixed-Acc	Adap-Acc	Fixed-Acc	Adap-Acc	Fixed-Acc	Adap-Acc
Multi-Cls	62.95	41.16	64.60	43.14	64.77	43.47
DETR (Carion et al., 2020)	63.08	40.69	65.65	44.93	65.62	45.23
Ours (Shao et al., 2022)	63.12	43.56	65.57	44.86	65.62	45.43
Ours++	**65.60**	**45.11**	**66.56**	**46.42**	**66.64**	**47.30**

The bold value denotes the best results

Table 8 Accuracy of detecting Seq-DeepFake-P compared to deepfake detection methods

Methods	Face components manipulation		Face attributes manipulation	
	Fixed-Acc	Adap-Acc	Fixed-Acc	Adap-Acc
DRN (Wang et al., 2019)	56.23	30.07	61.23	38.33
MA (Zhao et al., 2021a)	65.41	43.88	65.59	44.66
TS (Luo et al., 2021)	64.76	42.89	65.56	44.72
Ours (Shao et al., 2022)	65.19	44.34	65.62	45.43
Ours++	**66.11**	**45.35**	**66.64**	**47.30**

The bold value denotes the best results

P data and tabulate their comparison results in Table 9. The experimental results demonstrate that SeqFakeFormer and SeqFakeFormer++ surpass both multimodal baselines by a large margin for the task of Seq-DeepFake, suggesting the efficacy of the proposed method.

8.1 Ablation Study

In this sub-section we study the impact of two proposed image-sequence reasoning modules in SeqFakeFormer++, Image-Sequence Contrastive Learning (ISC) and Image-Sequence Matching (ISM), to the overall performance. We tabulate ablation studies in terms of ISM and ISC in Tables 10 and 11. As observed from Tables 10 and 11, the performance regarding two Seq-DeepFake detection tasks under all evaluation metrics degenerates once removing one of these reasoning modules, indicating the effectiveness and complementarity of them. It should be noted from Tables 10 and 11 that adding one of the reasoning modules in SeqFakeFormer++ will obtain better performance than SeqFakeFormer. This further verifies the importance of each reasoning module for detecting Seq-DeepFake under perturbations.

Table 9 Accuracy of detecting Seq-DeepFake-P compared to multimodal learning methods

Methods	Face components manipulation		Face attributes manipulation	
	Fixed-Acc	Adap-Acc	Fixed-Acc	Adap-Acc
BLIP (Li et al., 2022)	55.19	37.50	57.78	39.97
ALBEF (Li et al., 2021a)	60.78	38.15	60.23	42.02
Ours (Shao et al., 2022)	65.19	44.34	65.62	45.43
Ours++	**66.11**	**45.35**	**66.64**	**47.30**

The bold value denotes the best results

Table 10 Ablation study of SeqFakeFormer++ in detecting Seq-DeepFake-P based on sequential facial components manipulation

Components		ResNet-34		ResNet-50	
ISM	ISC	Fixed-Acc	Adap-Acc	Fixed-Acc	Adap-Acc
✗	✗	63.51	42.01	65.19	44.34
✗	✓	63.72	42.13	65.22	44.50
✓	✗	63.66	42.27	65.34	44.52
✓	✓	**65.23**	**44.45**	**66.11**	**45.35**

The bold value denotes the best results

Table 11 Ablation study of SeqFakeFormer++ in detecting Seq-DeepFake-P based on sequential facial attributes manipulation

Components		ResNet-34		ResNet-50	
ISM	ISC	Fixed-Acc	Adap-Acc	Fixed-Acc	Adap-Acc
✗	✗	65.57	44.86	65.62	45.43
✗	✓	66.17	46.36	65.87	46.21
✓	✗	66.47	46.31	66.48	46.24
✓	✓	**66.56**	**46.42**	**66.64**	**47.30**

The bold value denotes the best results

9 Conclusion

This paper studies a novel research problem—Detecting Sequential DeepFake Manipulation, aiming to detect a sequential vector of multi-step facial manipulation operations. We also introduce the first Seq-DeepFake dataset to provide sequentially manipulated face images. Supported by this new dataset, we cast detecting Seq-DeepFake manipulation as a specific image-to-sequence task and propose a Seq-DeepFake Transformer (SeqFakeFormer). Two modules, Spatial Relation Extraction and Sequential Relation Modeling with Spatially Enhanced Cross-Attention, are integrated into SeqFakeFormer, complementing each other. We further construct a more challenging Seq-DeepFake-P dataset to mimic the more realistic deepfake data distributions under perturbations. A dedicated Seq-DeepFake Transformer with Image-Sequence Reasoning (SeqFakeFormer++) is built to deal with such a noisy scenario. Extensive experimental results demonstrate the superiority of SeqFakeFormer and SeqFakeFormer++ and valuable observations pave the way for future research in broader deepfake detection.

Acknowledgements This study is supported by National Natural Science Foundation of China (Grant No. 62306090); Natural Science Foundation of Guangdong Province of China (Grant No. 2024A1515010147); This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221-0012), NTU NAP, and under the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

Data Availability The Seq-DeepFake dataset analysed during this study is publicly available for the research purpose - Seq-DeepFake dataset.

References

Bello, I., Zoph, B., Vaswani, A., Shlens, J., & Le, Q. V. (2019). Attention augmented convolutional networks. In *CVPR* (pp. 3286–3295).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *ECCV* (pp. 213–229).

Chen, G., Shen, L., Shao, R., Deng, X., & Nie, L. (2024). Lion: Empowering multimodal large language model with dual-level visual knowledge. In *CVPR* (pp. 26540–26550).

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *CVPR* (pp. 248–255).

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). The deepfake detection challenge (DFDC) preview dataset. arXiv preprint arXiv:1910.08854

Durall, R., Keuper, M., Pfreundt, F. J., & Keuper, J. (2019). Unmasking deepfakes with simple features. arXiv preprint arXiv:1911.00686

Dzanic, T., Shah, K., & Witherden, F. (2020). Fourier spectrum discrepancies in deep network generated images. *NeurIPS, 33*, 3022–3032.

Gao, P., Zheng, M., Wang, X., Dai, J., & Li, H. (2021). Fast convergence of DETR with spatially modulated co-attention. In *CVPR* (pp. 3621–3630).

Gu, S., Bao, J., Chen, D., & Wen, F. (2020). GIQA: Generated image quality assessment. In *ECCV* (pp. 369–385).

Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021). Lips don't lie: A generalisable and robust approach to face forgery detection. In *CVPR* (pp. 5039–5049).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).

He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., & Liu, Z. (2021). ForgeryNet: A versatile benchmark for comprehensive forgery analysis. In *CVPR* (pp. 4360–4369).

Jiang, L., Li, R., Wu, W., Qian, C., & Loy, C. C. (2020). Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR* (pp. 2889–2898).

Jiang, Y., Huang, Z., Pan, X., Loy, C. C., & Liu, Z. (2021). Talk-to-edit: Fine-grained facial editing via dialog. In *ICCV* (pp. 13799–13808).

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *CVPR* (pp. 4401–4410).

Kim, H., Choi, Y., Kim, J., Yoo, S., & Uh, Y. (2021). Exploiting spatial dimensions of latent in GAN for real-time image editing. In *CVPR* (pp. 852–861).

Lee, C. H., Liu, Z., Wu, L., & Luo, P. (2020). Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*.

Lee, W., Kim, D., Hong, S., & Lee, H. (2020). High-fidelity synthesis with disentangled representation. In *ECCV* (pp. 157–174).

Li, J., Dongxu, L., Caimingm, X., & Steven, H. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML* (pp. 12888–12900).

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS, 34*, 9694–9705.

Li, J., Xie, H., Li, J., Wang, Z., & Zhang, Y. (2021). Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *CVPR* (pp. 16317–16326).

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face X-ray for more general face forgery detection. In *CVPR* (pp. 5001–5010).

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). CELEB-DF: A large-scale challenging dataset for deepfake forensics. In *CVPR* (pp. 3207–3216).

Li, Z., Xie, Y., Shao, R., Chen, G., Jiang, D., & Nie, L. (2024). Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. In *NeurIPS*.

Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., & Yu, N. (2021). Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *CVPR* (pp. 772–781).

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *CVPR* (pp. 3730–3738).

Luo, Y., Zhang, Y., Yan, J., & Liu, W. (2021). Generalizing face forgery detection with high-frequency features. In *CVPR*.

Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. In *ICML* (pp. 4055–4064).

Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV* (pp. 86–103).

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *CVPR* (pp. 1–11).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). GRAD-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV* (pp. 618–626).

Shao, R., Lan, X., Li, J., & Yuen, P. C. (2019). Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR* (pp. 10023–10031).

Shao, R., Lan, X., & Yuen, P. C. (2017). Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3D mask face anti-spoofing. In *IJCB* (pp. 748–755).

Shao, R., Lan, X., & Yuen, P. C. (2018). Joint discriminative learning of deep dynamic textures for 3D mask face anti-spoofing. *IEEE Transactions on Information Forensics and Security, 14*(4), 923–938.

Shao, R., Lan, X., & Yuen, P. C. (2020). Regularized fine-grained meta face anti-spoofing. In *AAAI* (Vol. 34, pp. 11974–11981).

Shao, R., Perera, P., Yuen, P. C., & Patel, V. M. (2020). Open-set adversarial defense. In *ECCV* (pp. 682–698).

Shao, R., Perera, P., Yuen, P. C., & Patel, V. M. (2022). Federated generalized face presentation attack detection. *IEEE Transactions on Neural Networks and Learning Systems, 35*(1), 103–116.

Shao, R., Perera, P., Yuen, P. C., & Patel, V. M. (2022). Open-set adversarial defense with clean-adversarial mutual learning. *International Journal of Computer Vision, 130*(4), 1070–1087.

Shao, R., Wu, T., & Liu, Z. (2022). Detecting and recovering sequential deepfake manipulation. In *ECCV* (pp. 712–728).

Shao, R., Wu, T., & Liu, Z. (2023). Detecting and grounding multi-modal media manipulation. In *CVPR* (pp. 6904–6913).

Shao, R., Wu, T., Nie, L., & Liu, Z. (2025). Deepfake-adapter: Dual-level adapter for deepfake detectiong. *International Journal of Computer Vision*.

Shao, R., Wu, T., Wu, J., Nie, L., & Liu, Z. (2024). Detecting and grounding multi-modal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(8), 5556–5574.

Shen, L., Chen, G., Shao, R., Guan, W., & Nie, L. (2024). MOME: Mixture of multimodal experts for generalist multimodal large language models. In *NeurIPS*.

Shen, Y., Gu, J., Tang, X., & Zhou, B. (2020a). Interpreting the latent space of GANs for semantic face editing. In *CVPR* (pp. 9243–9252).

Shen, Y., Yang, C., Tang, X., & Zhou, B. (2020b). Interfacegan: Interpreting the disentangled face representation learned by GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(4), 2004–2018.

Shen, Y., & Zhou, B. (2021). Closed-form factorization of latent semantics in GANs. In *CVPR* (pp. 1532–1540).

Voynov, A., & Babenko, A. (2020). Unsupervised discovery of interpretable directions in the GAN latent space. In *ICML* (pp. 9786–9796).

Wang, H., Liu, W., Bocchieri, A., & Li, Y. (2021). Can multi-label classification networks know what they don't know? *NeurIPS, 34*, 29074–29087.

Wang, S. Y., Wang, O., Owens, A., Zhang, R., & Efros, A. A. (2019). Detecting photoshopped faces by scripting photoshop. In *CVPR* (pp. 10072–10081).

Wang, W., Alameda-Pineda, X., Xu, D., Fua, P., Ricci, E., & Sebe, N. (2018). Every smile is unique: Landmark-guided diverse smile generation. In *CVPR* (pp. 31–39).

Yang, H., Huang, D., Wang, Y., & Jain, A. K. (2018). Learning face age progression: A pyramid architecture of GANs. In *CVPR* (pp. 31–39).

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. In *CVPR* (pp. 2185–2194).

Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., & Xia, W. (2021). Learning self-consistency for deepfake detection. In *ICCV* (pp. 15023–15033).

Zhu, P., Abdal, R., Qin, Y., & Wonka, P. (2020). Sean: Image synthesis with semantic region-adaptive normalization. In *CVPR* (pp. 5549–5558).

Zhu, X., Wang, H., Fei, H., Lei, Z., & Li, S. Z. (2021). Face forgery detection by 3D decomposition. In *CVPR* (pp. 2929–2939).

Zhuang, P., Koyejo, O., & Schwing, A. G. (2021). Enjoy your editing: Controllable GANs for image editing via latent space navigation. In *ICLR*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.