Contrastive Learning for DeepFake Classification and Localization via Multi-Label Ranking

Supplementary Material

A. Dataset details

A.1. Sequential deepfake datasets

The Seq-DeepFake dataset. Recently introduced in the study by Shao et al. [32], the Seq-DeepFake dataset is a comprehensive collection of sequentially manipulated images. It encompasses two distinct subsets: Sequential Facial Components Manipulation (Seq-FaceComp) and Sequential Facial Attributes Manipulation (Seq-FaceAttr).

Seq-FaceComp subset. The Seq-FaceComp subset is an assembly of images where facial components are transplanted from a source image to a target image. This process results in composite images that exhibit clearly distinguishable facial parts, each with a designated sequence order. The subset contains a total of 35,166 images, a blend of both manipulated and original unaltered images. These images are annotated to reflect 28 different manipulation sequences. The sequence lengths vary, ranging from 1 to 5, with their distribution frequencies being approximately 20.48%, 20.06%, 18.62%, 20.88%, and 19.96% respectively.

Seq-FaceAttr subset. In contrast, the Seq-FaceAttr subset focuses on the direct modification of specific facial attributes in the target image, independent of any source image. This subset is composed of 49,920 images. Each image in this collection is categorized under one of the 26 identified manipulation sequence types. The distribution across the various sequence lengths is maintained evenly, ensuring a balanced representation in the dataset.

Common features and application. Noteworthy is the fact that both subsets, Seq-FaceComp and Seq-FaceAttr, feature a maximum sequence length of five, denoted as $L = 5$ in the context of our methodology (as detailed in Section 3). This characteristic is pivotal for our analysis. Moreover, these datasets present a unique opportunity for evaluating models in a multi-label classification scenario, with a particular focus on the sequences but independent of their specific order. This aspect is critical in understanding and detecting sequential manipulations in images, which has significant implications in the field of computer vision, particularly in the domain of deepfake detection and analysis.

A.2. Binary deepfake datasets

Overview of benchmark datasets. The field of deepfake detection is underpinned by several benchmark datasets, each playing a critical role in advancing binary classification research. Prominent among these are FaceForensics++ (FF++) [30], Celeb-DF (CDF) [22], WildDeepfake (WDF) [44], DeepFakeDetection (DFD) [30], and the DeepFake Detection Challenge (DFDC) [12].

FaceForensics++ (FF++). FF++ is renowned for its diverse dataset, comprising 1,000 videos for each of its four distinct manipulation techniques. This variety presents a comprehensive challenge in detecting a range of deepfakes.

Celeb-DF (CDF). CDF is known for its high-quality forgeries, offering a dataset that includes 5,639 manipulated videos and 590 original videos. The high fidelity of these deepfakes makes CDF a stringent test for detection models.

WildDeepfake (WDF). Reflecting real-world scenarios, WDF provides a balanced dataset with 3,509 fake and 3,805 genuine face sequences. This dataset is crucial for training and evaluating models under realistic conditions.

DeepFakeDetection (DFD). DFD contributes with its collection of 1,000 deepfake videos, adding to the diversity of available test data.

DeepFake Detection Challenge (DFDC). The DFDC dataset stands out with its substantial size, comprising an equal distribution of 2,500 authentic and 2,500 manipulated videos in its public test set.

Adaptation to binary classification. Our proposed model is adeptly tailored for the binary classification of deepfakes. By excluding the ranking loss component, $\mathcal{L}_{\text{Rank}}$, from the total loss equation (15) and limiting the number of learnable class tokens to one, the model becomes streamlined for this specific task. This configuration maintains its effectiveness by focusing exclusively on distinguishing between genuine and synthetic content, negating the need for sequence ranking. This simplification is particularly advantageous for efficiently addressing the binary nature of deepfake detection.

B. Implementation details

Feature extraction and tokenization. Our implementation begins with feature extraction using ResNet-34 and ResNet-50 as backbone networks. The extracted spatial features are then transformed into a sequence of tokens. These tokens, along with L learnable class tokens, form the input for a single-layer transformer.

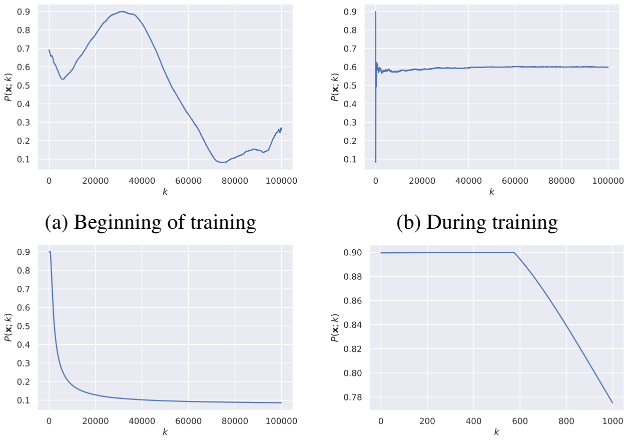
Model training. The training of the model adheres to the loss function specified in the total loss equation (15). The hyperparameters—λ_1, λ_2, τ, and α—are set to values of 1, 1, 0.2, and 1, respectively. The training extends over 200 epochs, employing the AdamW optimizer in conjunction with a cosine annealing learning rate schedule. We set the initial learning rate at 10^{-4}, reducing the learning rate for the feature extractor by a factor of ten.

Stronger backbone for further improvement. In addition to the ResNet-based models, we have conducted experiments using the Swin Transformer [25] as an alternative backbone. For adapting the Swin-based model, we followed the training procedures detailed in [25], extending the training duration to 400 epochs. These experiments were facilitated by Nvidia V100 GPUs.

Data augmentation techniques. Our approach includes two innovative data augmentation techniques inspired by SBIs [34], aiming to enhance the model's robustness and generalization capabilities. The first technique, dubbed "*strong augmentation*", involves generating a pseudo-fake image from an authentic facial image. This process distorts the facial landmarks of the original image to craft a modified version that appears inauthentic. The second technique, "*weak augmentation*", incorporates standard image classification methods such as horizontal flipping, random cropping, color jittering, and Gaussian blurring. In our training process, strong augmentation is initially applied to create a pseudo-fake image. Following this, we randomly select a fake image from the dataset and pair it with genuine and pseudo-fake images. This approach ensures a balanced representation of each image type within the training dataset, crucial for effective model training and performance.

Overview of the model architecture Illustrated in Figure 2, our proposed framework comprises two distinct branches: the patch branch and the class branch. We employ two types of tokens within these branches—patch tokens and learnable class tokens—to capture the features of the image's patches and its overall class, respectively. This concept is reminiscent of the Visual Transformer architecture, where a CLS token encapsulates the class feature of an image, while the patch tokens represent the local features. Our model leverages these tokens to embody two levels of feature granularity: patch tokens for capturing local details, and class tokens for identifying the type of manipulation. In the context of binary deepfake detection, our methodology adopts a multiple instance learning (MIL) perspective, utilizing a sorted list of similarity scores between patch tokens. This approach is regulated by the loss function (12) ($\mathcal{L}_{\text{CLS}}$).

For multi-label classification tasks, we employ distinct loss functions, (17) ($\mathcal{L}^{U}_{\text{BCE}}$ and $\mathcal{L}^{V}_{\text{BCE}}$) , to fine-tune the two types of tokens, each tailored to enhance multi-label outcome predictions. Additionally, we introduce a specialized loss term, (13) ($\mathcal{L}_{\text{Rank}}$) , designed to address the nuances of multi-label ranking challenges. Our training regimen aims to enable the network to perform deepfake classification and localization concurrently. This is achieved by considering a cumulative loss function, (15) ($\mathcal{L}_{\text{Total}}$). Overall, our work adopts the MIL framework for the specified task and proposes a novel strategy for managing multi-label ranking, setting our approach apart from prior methodologies.



(a) Beginning of training

(b) During training

(c) End of training

(d) Zoom in from (c)

Figure A1. **The distribution variation of the proposed $P(\mathbf{x}; k)$** from the beginning training to the end on a fake sample.

C. Additional results

Effectiveness of contrastive MIL. The efficacy of our contrastive Multiple Instance Learning (MIL) approach is thoroughly investigated by observing the changes in the probability distribution function $P(\mathbf{x}; k)$ over the course of training on deepfake datasets. Initially, as depicted in Figure A1 (a), there is a tendency for similarity values between genuine and fake samples to congregate around certain pivotal points. The implementation of contrastive MIL specifically targets the separation of these similarity values associated with genuine-fake pairings from those observed in genuine-genuine and fake-fake combinations.

As the training progresses, notable shifts in the probability distribution function $P(\mathbf{x}; k)$ occur. These shifts are highlighted in Figure A1 (b), illustrating the progressive separation and clarity achieved through our approach. At the culmination of the training phase, there is a significant congregation of genuine-fake instances trending towards zero, as clearly demonstrated in Figure A1 (c). For a more detailed understanding, Figure A1 (d) provides a closer examination of the probability distribution function across various values of k. This analysis is vital in understanding the subtleties of deepfake detection within the framework of multi-instance learning.

Setting	λ_1 ($\mathcal{L}_{\text{bce}}$)	λ_2 ($\mathcal{L}_{\text{Rank}}$)	τ	α	Ranking Acc. (%)
I	1	1	0.2	1	**74.54 (+0.00%)**
II	0.5	1	0.2	1	73.42 (-1.12%)
III	1.5	1	0.2	1	73.22 (-1.32%)
IV	1	0.5	0.2	1	72.33 (-2.21%)
V	1	1.5	0.2	1	74.12 (-0.42%)
VI	1	1	0.5	1	74.55 (+0.01%)
VII	1	1	0.2	0.5	72.27 (-2.27%)

Table A1. **The ablation studies of hyperparameters in losses.**

Method	Ranking Acc. (%)	Intra-testing AUC (%) (FF++)	Cross-testing AUC (%) (CDF)
Ours (separate)	74.54	99.82	89.12
Ours (joint)	**74.62**	**99.89**	**90.23**

Table A2. **The results with simultaneously training** on the traditional binary and sequential manipulation datasets.

Ablation studies on loss hyperparameters. We present the ablation studies on the hyperparameters described in weight vector (14) and the total loss equation (15), as shown in TableA1. These investigations employ ResNet-50 as the underlying architecture and are carried out using the Seq-FaceComp dataset.

Simultaneous training within the unified framework. To fully utilize the integrated architecture of the proposed unified framework, we conducted exploratory experiments to evaluate the advantages of simultaneous training on both the Seq-FaceComp and FaceForensics++ datasets. The results, as detailed in Table A2, show a moderate improvement in sequential deepfake manipulation outcomes. However, we observed a more substantial improvement in binary deepfake classification tasks, especially in cross-testing scenarios.

D. More Qualitative Results

Grad-CAM visualization. We present qualitative results utilizing Grad-CAM [31] on the Seq-FaceComp dataset. Selected examples are shown in Figure A2. These heatmaps are derived from the backpropagation of logits associated with features such as ”Lip” and ”Nose”. The application of contrastive MIL and ranking mechanisms notably refines the heatmaps' focus and precision, a comparison clearly visible in Figure A2 (b) when contrasted with the baseline method. It's evident that areas showing lower similarity values align with the manipulated regions in the images, corroborating our expectations.

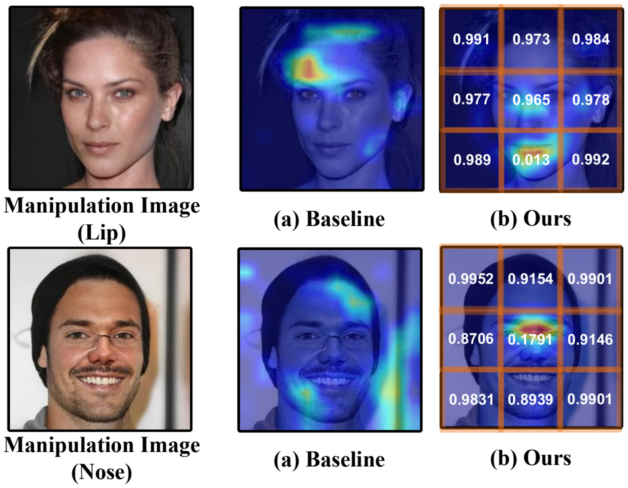


Figure A2. **Qualitative results** by Grad-CAM [31] between baseline and the proposed approach. Two test images from the Seq-FaceComp with “Lip” and “Nose” manipulation. (a) Although the heatmap from the baseline has noticed the accurate region sporadically, it still has a gap to improve. (b) The heatmap from the proposed approach has successfully focused on the manipulation region.

E. Limitations

Scope of applicability. While our contrastive multi-instance learning technique demonstrates high accuracy in distinguishing authentic from counterfeit instances, especially in the context of sequential manipulations, its application has been predominantly in analyzing human facial data. However, the realm of deep learning-based manipulations is rapidly expanding, not just in the visual domain but also encompassing audio and other digital media forms. This expansion underscores the need for a more versatile detection methodology capable of identifying multi-modal manipulations. Addressing this broader spectrum of deepfake phenomena remains a crucial challenge and an area for future development in our approach.