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

Cheng-Yao Hong Yen-Chi Hsu Tyng-Luh Liu

Institute of Information Science, Academia Sinica, Taiwan

{sensible, yenchi, liutyng}@iis.sinica.edu.tw

Abstract

We propose a unified approach to simultaneously addressing the conventional setting of binary deepfake classification and a more challenging scenario of uncovering what facial components have been forged as well as the exact order of the manipulations. To solve the former task, we consider multiple instance learning (MIL) that takes each image as a bag and its patches as instances. A positive bag corresponds to a forged image that includes at least one manipulated patch (i.e., a pixel in the feature map). The formulation allows us to estimate the probability of an input image being a fake one and establish the corresponding contrastive MIL loss. On the other hand, tackling the component-wise deepfake problem can be reduced to solving multi-label prediction, but the requirement to recover the manipulation order further complicates the learning task into a multi-label ranking problem. We resolve this difficulty by designing a tailor-made loss term to enforce that the rank order of the predicted multi-label probabilities respects the ground-truth order of the sequential modifications of a deepfake image. Through extensive experiments and comparisons with other relevant techniques, we provide extensive results and ablation studies to demonstrate that the proposed method is an overall more comprehensive solution to deepfake detection.

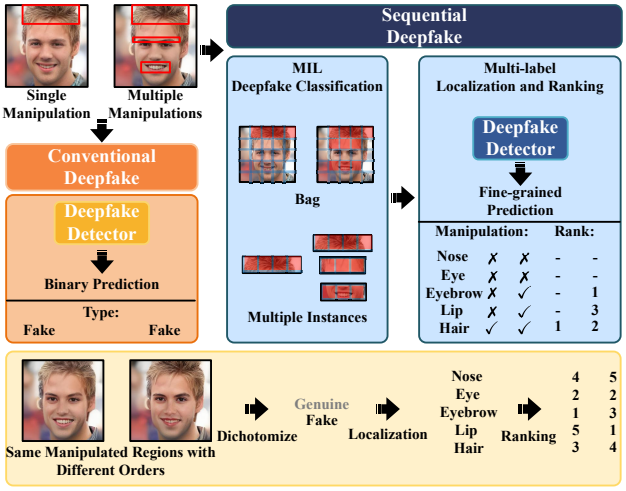

Figure 1. **Overview.** While addressing the conventional binary deepfake detection that dichotomies the images into genuine/fake, this work also focuses on the subtle scenario that forged images via deepfake mechanisms may be locally manipulated by one or more than one facial component/attribute. We introduce a multi-label ranking approach to tackling the "fine-grained" deepfake task (*i.e.*, to localize the modified facial components and to identify the order of manipulations) and develop a contrastive multi-instance learning (MIL) framework to solve the binary classification. It is noteworthy to mention that manipulating the same regions in different orders could result in distinct manipulated images.

1. Introduction

With the rapid growth of face-swapping techniques, deep forgery has become a concern on social media. An effective solution to address the matter is to utilize neural network-based approaches to decide the authenticity of given images. The task of deepfake classification is usually formulated as a binary classification problem. Recent research efforts on deepfake classification have delivered saturated performances [2, 5, 34, 40, 41]. Nevertheless, owing to the impressive development of generative networks, *e.g.*, StyleGAN [17] and diffusion models [18, 19, 29], deep forgery is no longer limited to face-to-face interchange. In particular, Shao et al. [32] propose a sequential facial manipulation dataset, Seq-DeepFake, in which the fake facial images are manipulated with the requested sequential constraints from the source (*e.g.*, components and attributes) by StyleMapGAN [19]. Take, for example, in Figure 1, the annotation of "Eyebrow-Hair-Lip" indicates that the resulting facial image has been successively manipulated with the eyebrow, hair and lip in the specified order. The sequential manipulations can be treated as a multi-label "localization" problem to decide not only which facial components have been forged but also what the manipulation order is. The latter task further complicates the localization scenario into a ranking problem,

which poses significant challenges and opens a new frontier for deepfake-related research.

Detecting sequential facial manipulations is more challenging than conventional deepfake classification. This causes most of the existing deepfake solutions to be no longer applicable. For example, the success of Face X-ray [20] is based on the observation that a fake facial image must have an essential blending operator to smooth the face boundary to make the forged image more natural during the face-swapping process. The particular method then focuses on learning how to capture the blending region from the paired source and target images. However, the tactic apparently does not work well on the sequential facial manipulation dataset, SeqDeepFake [32, 33]. The inefficiency is caused by two main factors. First, the paired source and target information of each manipulated image in SeqDeepFake is not available. Second, the resulting classifier from adversarial learning is often highly related to the generator. Therefore, it is hard to generalize the method to distinguish sequentially manipulated images without completely updating the generative model in [20]. The inadequacy in handling component-wise deepfake is indeed a common issue across many related methods,*e.g.*, [4, 5, 13, 34, 41]. After all, they are developed to solve a binary classification problem, rather than dealing with the sequential deepfake manipulations.

Aiming to establish a unified approach to deepfake detection, we decompose the underlying problem into three subtasks, including *deepfake classification*, *deepfake localization*, and *manipulation order*. In resolving the first subtask, we propose contrastive *multiple instance learning* (MIL) that treats an image as a bag and the spatial features as instances to tackle deepfake classification via minimizing a contrastive MIL loss. We then establish a multi-label ranking formulation to address the other two subtasks. Concerning the ability to identify which facial components have been forged, we loosely term the process as deepfake localization. In addition, it is reasonable to incorporate ranking reasoning into the stage so that the ranked list of multi-label probabilities can reflect the sequential modification order. As such, training the network model can be done via multi-task learning, and results in an effective deepfake detection model capable of accomplishing the three aforementioned tasks. We characterize our main contributions as follows.

- We decompose the general deepfake problem into three parts, *deepfake classification*, *deepfake localization*, and *manipulation order* which leads to a systematic view of solving the deepfake problem comprehensively.
- We propose a contrastive multi-instance learning formulation for binary deepfake classification. The synergy between the two learning paradigms improves the model learning effectively, and more importantly, it gives rise to a well-established concept of how to define the probability of a given image being deepfake.
- We develop a multi-label ranking approach to coupling multi-label predictions with ranking reasoning. In inference, the sequential order of deepfake manipulations can be readily obtained from the rank order of the output multi-label probabilities.
- We establish a unified approach to deepfake classification and localization, and achieve state-of-the-art performances on popular benchmark datasets.

2. Related work

Deepfake detection. Owing to the active development of face manipulation technology and the upsurge of people's awareness about multimedia security, more research efforts have been paid to develop all sorts of deepfake detection methods in recent years. Deepfake detection can be categorized into two types of approaches based on the underlying data format: *image-based* [2, 5, 13, 20, 24, 34, 40, 41, 43] and *video-based* [8, 14, 21, 42]. For image-based deepfake detection, Zhu et al. [43] propose a two-stream architecture to enrich the face feature for detection. One is a conventional network, and the other is a 3D decomposition framework that aims to find more clues and details on the face image. Chen et al. [5] fuse the RGB and frequency features with a cross-attention module to learn an artifact mask decoder from the fake images. The decoder uses the source and target information from the manipulated image to generate the mask as a ground-truth label. Cao et al. [2] regard the detection problem as anomaly detection and utilize an encoder-decoder framework for real-fake representation learning. Liu et al. [24] determine the forgery image from the phase spectrum variation between the original and up-sampled images. Zhao et al. [40] introduce multiple attention modules to capture different discriminative locations and insert a texture enhancement block into the backbone to extract high-frequency features. Several other methods attempt to capture the artifacts generated by swapping faces from two images. Li et al. [20] propose Face X-ray to find the blended region from the forgery image. Moreover, Zhao et al. [41] exploit the fact that the forgery faces are manipulated from two different sources and propose an inconsistent image generator to support the classifier in learning the consistency mask. Based on a similar entry point, Dong et al. [13] utilize the self-attention mechanism to form an identity consistency transformer to detect a forgery image. To extend the above concepts, Shiohara and Yamasaki [34] introduce a self-blended framework that can learn the blended clues from the proposed augmentation technique.

For video-based deepfake detection, Cozzolino et al. [8] use a three-dimensional morphable model to generate deepfake video and learn a temporal network to embed the sequence features for the video classifier. Zhou and Lim [42] present a two-plus-one joint detection model for tackling both manipulated visual and auditory modalities.

More recently, Shao et al. [32, 33] generalize the image-based deepfake detection from a binary classification problem to a multi-label classification problem. Specifically, the image is manipulated from sequential components/attributes, dramatically increasing the detection challenge.

Multiple instance learning. Following [11], the *multiple instance learning* (MIL) paradigm defines a "bag" as *positive* if it contains at least one positive instance. In other words, all instances in a negative bag are assumed to be negative. An earlier approach by Chen et al. [6] transforms each sample bag into a high-dimensional feature space and adopts the Support Vector Machine (SVM) to determine the essential features and construct the classifier simultaneously. Ilse et al. [16] introduce MIL attention pooling that leverages neural networks to parameterize the distribution of instances in a bag to detect predefined positive instances. In medical imaging, several approaches regard MIL-related tasks on histopathology datasets as weakly supervised learning. Zhang et al. [39] introduce the pseudo-bag concept to enrich the sample bags to address the insufficiency of whole slide images. Furthermore, Thandiackal et al. [36] present Zoom-MIL which utilizes multi-level zooming to fuse multiple magnifications and reduce the computation cost.

Ranking mechanism. A ranking scheme is designed to find the optimal sorting function that can rank the sequential input. Although early efforts [1, 23] propose the bitonic sorting network to solve the rank issue, most current techniques rely on the neural network to achieve the differential ranking operation. Petersen et al. [27] first present Differentiable Sorting Networks and take it as an extension by enforcing monotonicity and limiting the bound of approximation error. They subsequently introduce a differential top-k network [28] to address the multi-class problem via the ranking mechanism.

3. Method

We consider a general formulation of deep-fake detection in which the underlying photorealistic manipulations can be applied to either the whole facial region or some of the predefined facial components. For the sake of discussion, we categorize the former task as *deepfake classification* and the latter as *deepfake localization*, where in this scenario we also need to recover the sequential order of the component-wise deepfake manipulations as described in [32].

Problem formulation. Suppose there are totally L facial components to which photorealistic manipulations can be applied. Since the exact order of modifying the facial components does matter, we cast the task of deepfake localization as a *multi-label ranking* problem [10]. Consider now a deepfake dataset $\mathcal{D} = \{(\mathbf{x}, Y)\}$, where $\mathbf{x}$ is an image and $Y = \{l_j\}_{j=1}^{k}$ with $k \leq L$ is an ordered subset of $\{1, 2, \ldots, L\}$, indicating that the jth ($j \leq k$) deepfake modification has been performed on the l_jth facial component. When Y is an empty set, it implies that $\mathbf{x}$ is a genuine facial image. It is convenient to generate from Y two L-dimensional vectors $\mathbf{y} = (y_i)$ and $\mathbf{r} = (r_i)$ by

$$y_i = \begin{cases} 1, & \text{if } i = l_j \in Y; \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

and

$$r_i = \begin{cases} j, & \text{if } i = l_j \in Y; \\ L, & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathbf{y}$ is the standard multi-label binary vector and $\mathbf{r}$ is the corresponding rank vector. We realize the above definitions with a hands-on example. Assume that totally five facial components can be modified, *i.e.*, $L = 5$, and a deepfake image has been created by first manipulating facial component 4 and then facial component 2. Our definitions imply that $Y = \{4, 2\}$, $\mathbf{y} = (0, 1, 0, 1, 0)$ and $\mathbf{r} = (5, 2, 5, 1, 5)$.

To train a deepfake detection model with the training data $\mathcal{D}$, we consider a CNN-transformer network, as illustrated in Figure 2. For each training sample $(\mathbf{x}, Y)$, the CNN+FPN module transforms $\mathbf{x}$ into feature maps of size $\mathbb{R}^{w \times h \times d}$, which can be reshaped and row-wise ℓ_2-normalized into a token vector $T \in \mathbb{R}^{N \times d}$ and $N = w \times h$ is the spatial size.

We then form two vectors of tokens, including the patch tokens $U \in \mathbb{R}^{N \times d}$ and the learnable class tokens, $V \in \mathbb{R}^{L \times d}$. The two sets of tokens are passed through the transformer encoder ϕ, which performs self-attention to correlate their features by

$$U \xrightarrow{\phi} \widetilde{U} \in \mathbb{R}^{N \times d}, \quad V \xrightarrow{\phi} \widetilde{V} \in \mathbb{R}^{L \times d}. \tag{3}$$

We compute the similarity values of each patch token to all other tokens by

$$S = \max(\widetilde{U}\widetilde{U}^{\top}, 0) \in \mathbb{R}^{N \times N}, \tag{4}$$

where S is rectified into a nonnegative matrix such that all of its elements are in $[0, 1]$. Since the similarity matrix is symmetric and we are concerned only with the correlations of each token to all other tokens, it suffices to focus on the upper triangular part of S, excluding those on the diagonal. We arrange these entries of interest in ascending order of similarity value and denote them by

$$\mathbf{u} = (u_1, u_2, \ldots, u_n), \tag{5}$$

where $n = N(N-1)/2$, the size of upper triangle of S.

MIL deepfake classification. With the sorted list $\mathbf{u}$ of similarity responses between patch tokens, we can consider the task of deepfake detection from the multiple instance learning (MIL) viewpoint. That is, we consider a face image $\mathbf{x}$ as a bag and the positive label 1 indicates that $\mathbf{x}$ is indeed

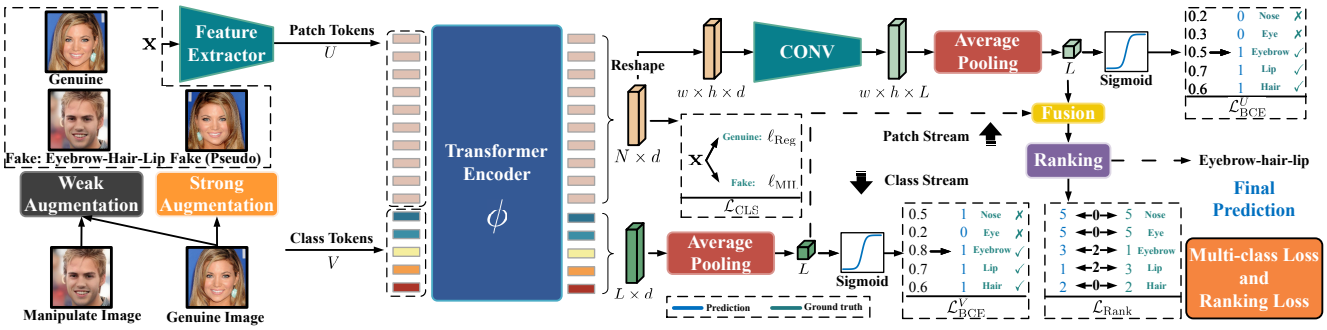

Figure 2. **The model architecture of our method.** There are two types of input tokens: *patch tokens* extracted from CNN+FPN and *learnable class tokens*. The stage of model training is driven by three loss terms: $\mathcal{L}_{\text{CLS}}$, $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{Rank}}$ to achieve contrastive multiple instance learning, multi-label localization and ranking, respectively. In the inference stage, the sequential order of deepfake manipulations can be readily obtained from the rank order of the output multi-label probabilities.

fabricated as a deepfake one. In terms of the elements in $\mathbf{u}$, if $\mathbf{x}$ is a deepfake image, we expect to uncover that there exists at least one u_i (starting from the front end of $\mathbf{u}$) with a small value close to 0. On the other hand, a negative bag (*i.e.*, $\mathbf{x}$ is not a deepfake image) implies all u_i are close to 1. To incorporate the above observations into the model learning process, we introduce a *contrastive* formulation to realize the MIL concept for deepfake detection. Assume that a deepfake image $\mathbf{x}$ results in the k smallest similarity responses on the front end of the sorted list $\mathbf{u}$. We propose to compute its probability of being deepfake by contrasting the average responses from the positive and negative distributions:

$$P(\mathbf{x};k) = 2 \times \frac{\exp(u^+(k)/\tau)}{\exp(u^+(k)/\tau) + \exp(u^-(k)/\tau)} - 1 \quad (6)$$

where τ is the temperature parameter,

$$u^+(k) = \frac{1}{k}\sum\nolimits_{i=1}^{k}(a - u_i), \quad (7)$$

$$u^-(k) = \frac{1}{n-k}\sum\nolimits_{i=k+1}^{n}(a - u_i), \quad (8)$$

and a is a scalar that is set to 1 in our implementation. The contrastive ratio in (6) is expected to be close to 1 when $\mathbf{x}$ is fake and $1/2$, otherwise. After shifting and scaling as in (6), it falls within $[0, 1]$ and can be used to approximate the probability of a given image $\mathbf{x}$ being a deepfake one by

$$P(\mathbf{x}) = \max_{1 \le k \le n} P(\mathbf{x};k), \quad (9)$$

where the reason for seeking a maximum is supported by the existence of at least one positive/fake instance. We thus define the contrastive MIL loss for each $(\mathbf{x}, Y) \in \mathcal{D}$ as

$$\ell_{\text{MIL}}(\mathbf{x}) = -J(Y)\log P(\mathbf{x}) - (1 - J(Y))\log(1 - P(\mathbf{x})) \quad (10)$$

where $J(Y) = 1$ if a sample $(\mathbf{x}, Y)$ is a deepfake image, and 0, otherwise. In addition, for an authentic image $\mathbf{x}$, it is reasonable to expect that all the similarity responses u_i should be close to 1. The useful observation motivates the inclusion of the following regularization loss:

$$\ell_{\text{Reg}}(\mathbf{x}) = \sum_{i=1}^{n} \|1 - u_i\|_2, \quad (11)$$

to ensure proper similarity responses for a real $\mathbf{x}$. We can then express the loss function for deepfake classification as

$$\mathcal{L}_{\text{CLS}} = \sum_{(\mathbf{x},Y)\in\mathcal{D}} \ell_{\text{MIL}}(\mathbf{x}) + (1 - J(Y))\,\ell_{\text{Reg}}(\mathbf{x}). \quad (12)$$

We are now ready to solve the multi-label ranking problem. To begin with, we average the patch-token and the class-token logits to obtain $\mathbf{f} = (f_i) = (\mathbf{f}^U + \mathbf{f}^V)/2$. The fusion between the two streams gives rise to multi-label predictions $\{P_i(\mathbf{x})\}_{i=1}^{L}$, analogous to those from (16). The main idea behind our formulation is as follows: by constructing a rank-aware loss term, the learned network model is expected to output multi-label predictions $\{P_i(\mathbf{x})\}_{i=1}^{L}$ that respect the rank order $\mathbf{r} = (r_i)$, implied by the given sample $(\mathbf{x}, Y) \in \mathcal{D}$. In other words, if $i, j \in Y$ and $r_i < r_j$ (*i.e.*, facial component i is modified before facial component j is manipulated), the network is trained to make multi-label predictions with $P_i(\mathbf{x}) > P_j(\mathbf{x})$. To this end, we design the following loss term for tackling multi-label ranking,

$$\mathcal{L}_{\text{Rank}} = \sum\nolimits_{(\mathbf{x},Y)\in\mathcal{D}} \mathbf{w}(\mathbf{r}, \{P_i(\mathbf{x})\}) \cdot \ell(\mathbf{x}), \quad (13)$$

where $\ell(\mathbf{x}) \in \mathbb{R}^L$ is analogously defined as in (18) but with multi-label predictions $\{P_i(\mathbf{x})\}_{i=1}^{L}$ based on the fused logits $\mathbf{f}$. To complete the explanation of (13), it remains to elaborate how the rank-aware weight vector $\mathbf{w}(\mathbf{r}, \{P_i(\mathbf{x})\})$ is designed. As our aim is to preserve the rank order $\mathbf{r}$ in the multi-label predictions $\{P_i(\mathbf{x})\}_{i=1}^{L}$, we let $\mathbf{o} = (o_i) \in \mathbb{R}^L$ to encode the rank order (nonincreasing order of probability

values) among the multi-label predictions. We then define the weight vector $\mathbf{w}(\mathbf{r}, \{P_i(\mathbf{x})\}) = (w_i) \in \mathbb{R}^L$ by

$$w_i = \begin{cases} \alpha, & \text{if } i \notin Y \ \wedge \ r_i > |Y|; \\ \alpha \times |o_i - r_i|, & \text{otherwise,} \end{cases} \tag{14}$$

where α is a hyperparameter to our method. We now justify the definition of $\mathbf{w}$. Given a deepfake sample $(\mathbf{x}, Y) \in \mathcal{D}$, there are $|Y| \leq L$ components that have been modified. The first condition in (14) indicates that the facial component i is genuine and its corresponding prediction $P_i(\mathbf{x})$ is not among the $|Y|$ largest outputs of $\{P_i(\mathbf{x})\}_{i=1}^L$. Such an outcome is preferable, and thus w_i is uniformly set to α. The second condition includes two scenarios. The first is that $i \notin Y$ and $r_i \leq |Y|$. This implies that the network model predicts a high-rank deepfake probability to a genuine facial component, which should be penalized with $\alpha \times |o_i - L|$. (Note that from (2), when $i \notin Y$, we set $r_i = L$.) The second scenario concerns the case that $i \in Y$, *i.e.*, facial component i has been changed. We thus formulate the definition of w_i to enforce reducing the difference between o_i and r_i. We conclude that by adding $\mathcal{L}_{\text{Rank}}$ to our formulation, the learned network model can output multi-label predictions $\{P_i(\mathbf{x})\}_{i=1}^L$ to detect which facial components have been manipulated, and also the order of modifications, which is implied by the resulting order of probability magnitudes.

Total loss. To train the proposed network model for simultaneously carrying out deepfake classification and localization, our method considers the following total loss:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CLS}} + \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{Rank}}, \tag{15}$$

where λ_1 and λ_2 are parameters to weigh the effects of specific loss terms, and $\mathcal{L}_{\text{BCE}} = \mathcal{L}_{\text{BCE}}^U + \mathcal{L}_{\text{BCE}}^V$. Note that the two sets of multi-label probability predictions $\{P_i^U\}$ and $\{P_i^V\}$ are computed only in the training stage so that $\mathcal{L}_{\text{BCE}}^U$ and $\mathcal{L}_{\text{BCE}}^V$ can be utilized to achieve effective model training. In inference, the multi-label prediction is provided solely from the $\mathcal{L}_{\text{Rank}}$ head, as shown in Figure 2.

Finally, we emphasize that the proposed approach provides a unified solution to the deepfake problem. When dealing with a classical task of binary deepfake classification, it is convenient to exclude the $\mathcal{L}_{\text{Rank}}$ term from the total loss in (15) and simply set the number of learnable class tokens to one to achieve binary classification.

Multi-label localization and ranking. The contrastive MIL formulation leads to a new loss term specified in (12) for learning deepfake classification. To extend our method for deepfake localization, we consider multi-label ranking to uncover which facial components have been modified as well as the underlying order of manipulations. The Transformer encoder ϕ generates, for each sample $(\mathbf{x}, Y)$, two sets of features from the patch tokens, $U \in \mathbb{R}^{N \times d}$ and the class tokens, $V \in \mathbb{R}^{L \times d}$ as in (3). Our network model applies convolutions to U and then carries out average pooling to obtain the patch-token logits $\mathbf{f}^U = (f_i^U) \in \mathbb{R}^L$. In a similar way, we have the class-token logits $\mathbf{f}^V = (f_i^V) \in \mathbb{R}^L$. By independently applying a sigmoid function σ to each logit, we obtain two sets of multi-label predictions as

$$P_i^{\mathcal{X}}(\mathbf{x}) = \sigma(f_i^{\mathcal{X}}) \in [0, 1], \quad i = 1, \ldots, L, \tag{16}$$

where $\mathcal{X}$ can be replaced by U or V to respectively imply that the predictions are based on the features from patch tokens or class tokens. Recall that the ground-truth label vector Y yields the corresponding multi-label binary vector $\mathbf{y} = (y_i)$ and the rank vector $\mathbf{r} = (r_i)$, which are both L-dimensional. With the multi-label predictions given by (16), we define the multi-label BCE loss as

$$\mathcal{L}_{\text{BCE}}^{\mathcal{X}} = \sum\nolimits_{(\mathbf{x}, Y) \in \mathcal{D}} \mathbf{1} \cdot \ell^{\mathcal{X}}(\mathbf{x}), \tag{17}$$

where "$\cdot$" denotes inner product, $\mathbf{1}$ is all-ones vector, and the ith element of $\ell^{\mathcal{X}}(\mathbf{x}) \in \mathbb{R}^L$ is defined by

$$\ell_i^{\mathcal{X}}(\mathbf{x}) = -y_i \log P_i^{\mathcal{X}}(\mathbf{x}) - (1 - y_i) \log(1 - P_i^{\mathcal{X}}(\mathbf{x})). \tag{18}$$

It is worth mentioning that both the multi-label predictions P^U and P^V from (16) are computed only during the training stage. Including the two loss terms $\mathcal{L}_{\text{BCE}}^U$ and $\mathcal{L}_{\text{BCE}}^V$ helps regulate model training and more critically align the class-wise logits from the patch-token and class-token streams.

4. Experiments

We begin by detailing the experimental outcomes on the sequential deepfake dataset [32]. Next, we present extensive results within a multi-label context as well as in conventional deepfake classification settings to illustrate the versatility of our method. In addition, we conduct a thorough ablation study to ascertain the contribution of each pivotal component within our methodology. For clarity in comparative analysis, techniques from other research that we include in our evaluation will be highlighted in bold throughout the discussion. Comprehensive dataset details are deferred to Appendix A.

Implementation details. For a fair comparison with the SeqFakeFormer approach detailed by Shao et al. [32], regarding the challenge of sequential facial manipulation detection, we have configured our method to utilize ResNet-34 and ResNet-50 architectures [15] as the convolutional neural network (CNN) backbones for feature extraction. In addressing the conventional deepfake classification task, facial frames are first extracted from the source videos via RetinaFace [9] and subsequently resized to a uniform resolution of 384×384 pixels. The training regimen for this task parallels that of the sequential facial manipulation, with Swin Transformer as the chosen backbone; however, we adjust the hyperparameters

Method	Seq-FaceComp Acc.		Seq-FaceAttr Acc.	
	Multi-label (%)	Ranking (%)	Multi-label (%)	Ranking (%)
Multi-Cls [32]*	78.32	69.66	85.14	66.99
DETR [37]*	-	69.87	-	67.93
SeqFakeFormer [32]*	-	72.13	-	67.99
Ours*	**82.31** ↑ 3.99	**73.72** ↑ 4.06	**86.42** ↑ 1.28	**68.82** ↑ 1.83
Multi-Cls [32]†	79.54	69.75	88.23	66.66
DRN [37]†	-	66.06	-	64.42
DETR [3]†	-	69.75	-	67.62
MA [37]†	-	71.31	-	67.58
Two-Stream [26]†	-	71.92	-	66.77
SeqFakeFormer [32]†	-	72.65	-	68.86
MMNet [38]†	-	73.93	-	69.27
Ours†	**84.12** ↑ 4.58	**74.54** ↑ 4.79	**90.45** ↑ 2.22	**69.58** ↑ 2.92
Ours‡	**84.36** ↑ 4.82	**74.97** ↑ 5.22	**90.74** ↑ 2.51	**70.02** ↑ 3.36

For feature extractor: * : ResNet-34 † : ResNet-50 ‡ : Swin Transformer as the backbone.

Table 1. **The experimental results with multi-label and ranking scenarios** on the Seq-FaceComp and Seq-FaceAttr datasets. Bold texts denote the best results. Note that the performance gains by ours are based on the baseline, Multi-Cls.

to $L = 1$ and $\lambda_2 = 0$, underscoring the adaptability of our method across different tasks. A crucial point to note is the computational scale of n—representing the size of the upper triangle of S in (5) which is on the order of $\mathcal{O}(N^4)$, with N signifying the feature map dimensions. Consequently, a brute-force approach in determining k across the range of 1 to n as shown in (9) and (10) could markedly impede training efficiency. To sidestep this computational bottleneck, we employ a strategy of uniformly sampling k values between 1 to n at 100 distinct points, in lieu of an exhaustive enumeration. For additional details on the training process, we refer the readers to Appendix B.

4.1. Sequential deepfake manipulation.

To address the challenge of sequential facial manipulation, our focus was on benchmarking the presented approach against SeqFakeFormer [32]. SeqFakeFormer integrates CNNs and transformers with an autoregressive mechanism to handle the sequential aspect of the problem. In a departure from this, the current method incorporates a ranking mechanism to manage the multi-label scenario, streamlining both the training and inference processes. To assess the efficacy of this novel approach, the fixed accuracy (Fixed-Acc) metric from [32] is employed on the Seq-FaceComp and Seq-FaceAttr datasets. Fixed-Acc quantitatively measures the sequence alignment of predictions and annotations, taking their rank-wise dependencies into account. The approach's performance is contrasted with several established methods, including simple multi-classifiers (**Multi-Cls**), **DRN** [37], **DETR** [3], **MA** [37], **Two-Stream** [26], SeqFakeFormer [32], and **MMNet** [38]. The comparative results are summarized in Table 1. The approach has proven not only practical but also superior in addressing the nuanced complexities of sequential deepfake detection, outperforming the referenced methodologies.

4.2. Binary deepfake classification.

In the domain of binary deepfake classification, the performance of the proposed method was assessed through both intra-testing and cross-testing scenarios. Contemporary research in deepfake detection can generally be divided into two primary categories. The initial category concentrates on classification tasks [3, 24, 40], employing exclusively authentic and counterfeit annotations for training. The latter category imposes no restrictions on the training process [13, 20, 34, 41]. Researchers within this group often integrate adversarial learning techniques to generate more sophisticated fake samples, thereby enhancing the robustness of the classifier. This study synthesizes the strengths of both aforementioned categories, leveraging a novel data augmentation strategy in conjunction with an end-to-end training framework. We also furnish a comparative analysis of the method's performance against several established works: **Multi-Att** [40], **SPSL** [24], **RECCE** [2], **Face X-Ray** [20], **LRL** [5], and **SBIs** [34]. These comparisons are drawn within the conventional scope of binary deepfake detection and span two distinct scenarios.

Intra-testing. The experiment entailed both the training and evaluation of a model on an identical dataset. As illustrated in Table 2 under the "Intra-testing" column, a majority of the methodologies have effectively addressed the deepfake classification challenge, with even fundamental models like

Method	Intra-testing AUC		Cross-testing (Train on FF++ only) AUC			
	FF++ (%)	CDF (%)	CDF (%)	WDF (%)	DFDC (%)	DFD (%)
Xception [7]	96.30	99.73	61.80	62.72	48.98	87.86
EifficientNet-B4 [35]	99.70	99.81	64.29	63.83	-	-
Multi-Att [40]†	99.29	99.94	67.44	59.74	-	-
SPSL [24]*	96.91	-	76.88	-	66.16	-
RECCE [2]*	99.32	99.94	68.71	64.31	69.06	-
Face X-Ray [20]*	99.17	-	80.58	-	**80.92**	95.40
LRL [5]*	99.46	-	78.26	-	76.53	89.24
SBIs [34]†	99.64	93.74	**93.18**	-	72.42	97.56
SBIs [34]‡	99.72	95.68	89.12	70.56	71.08	97.34
Ours‡	**99.82** ↑ 3.52	**99.98** ↑ 0.25	91.56 ↑ 29.76	**73.41** ↑ 10.69	75.17 ↑ 26.19	**97.88** ↑ 10.02

For feature extractor: * : Xception † : EifficientNet-B4 ‡ : Swin Transformer as the backbone.

Table 2. **The experimental results with intra-testing and cross-testing.** The model for cross-testing is only trained on the FF++ dataset. Bold texts denote the best results. Note that the performance gains in the last row are based on the baseline, Xception.

Model	$\mathcal{L}^U_{\text{BCE}}$	$\mathcal{L}^V_{\text{BCE}}$	$\mathcal{L}^U_{\text{CLS}}$	$\mathcal{L}^U_{\text{Rank}}$	Seq-FaceComp Acc. Ranking (%)
I	✔				52.43
II	✔		✔		54.21
III	✔			✔	72.52
IV	✔		✔	✔	73.43
V		✔	✔	✔	72.87
VI	✔	✔	✔	✔	**74.54**

Table 3. **Ablation study of the proposed losses** on the Seq-FaceComp with multi-label ranking setting (ResNet-50).

Xception [7] and EfficientNet-B4 [35] demonstrating impressive accuracy. The method under discussion here exhibits the highest accuracy, though the margin of improvement is slight. Echoing the sentiments presented in the introduction, it appears that the intra-testing performance is nearing a plateau. Consequently, the principal obstacle in deepfake detection now lies in generalizing to cross-testing scenarios.

Cross-testing. The standard protocol was adhered to by training each model exclusively on the FaceForensics++ (FF++) [30] dataset, followed by evaluating their performance on the test sets of Celeb-DF (CDF) [22], WildDeepfake (WDF) [44], DeepFakeDetection (DFD) [30], and DeepFake Detection Challenge (DFDC) [12].. The corresponding outcomes are presented in the "Cross-testing" column of Table 2. To ensure equitable comparisons, the SBIs approach was implemented using the Swin Transformer as a backbone, denoted as SBIs‡ in Table 2. Our method yields substantial performance gains, especially on the DFDC dataset, which can be attributed to the implemented augmentation strategy and the classification loss $\mathcal{L}_{\text{CLS}}$. This enhancement suggests that harnessing the fine-grained information amongst patch tokens is advantageous for deepfake detection.

Manipulation Components	Nose	Eye	Eyebrow	Lip	Hair
Baseline (Multi-Cls)	0.41	0.38	0.35	0.46	0.42
Ours	**0.72**	**0.61**	**0.66**	**0.75**	**0.74**

Table 4. **The correlation**, quantified by *coefficient of determination* R^2, between manipulation components prediction and ordering.

4.3. Ablation study and analysis

Effect of each loss. In contrast to the Multi-Cls model described in Table 1, the ResNet-50 model utilizing $\mathcal{L}^U_{\text{BCE}}$ in Table 3 exclusively generates multi-label predictions. These predictions are then ranked by their respective probabilities for each category before evaluation. Absent the $\mathcal{L}_{\text{Rank}}$, while the multi-label performance of ResNet-50 is noteworthy, there is a marked decrease in the Fixed-Acc metric, attributable to the missequenced order of predictions. Therefore, $\mathcal{L}_{\text{Rank}}$ plays an essential role in refining multi-label predictions into ordered sequence predictions. Moreover, the introduction of the contrastive Multiple Instance Learning (MIL) loss, $\mathcal{L}_{\text{CLS}}$, markedly enhances model performance.

Correlation analysis. As outlined in Section 3, it is posited that for any two facial components i and j included in Y, if $r_i < r_j$ (indicating that facial component i is modified prior to the manipulation of facial component j), the network should be trained to predict labels such that $P_i(\mathbf{x}) > P_j(\mathbf{x})$. To verify the association between the predictions and the order of component alterations, the *coefficient of determination*, denoted as R^2, is calculated. The results in Table 4 reveal a strengthened correlation between prediction and actual order. This outcome validates the effectiveness of the proposed method in addressing the ranking within multi-label classification tasks, particularly through the use of $\mathcal{L}_{\text{Rank}}$.

Qualitative results. The presentation of qualitative results continues with the use of Grad-CAM [31] on the Seq-FaceComp dataset, as depicted in Figure 3. The heatmaps

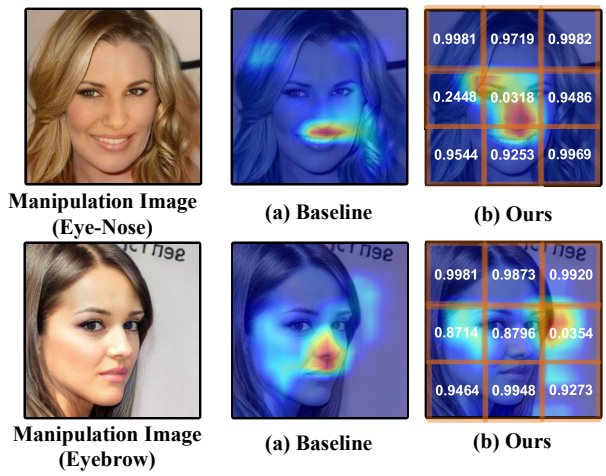

Figure 3. **Qualitative visualization.** Grad-CAM results of two test images from Seq-FaceComp.

Figure 4. **Qualitative visualizations comparison** between baseline (Multi-Cls), SeqFakeFormer, and ours.

are generated by backpropagating the logits for "Eyebrow" and "Eye-Nose" categories. Thanks to the integration of the contrastive and the ranking mechanisms, Figure 3(b) exhibits a heatmap that is both more concentrated and precise in comparison to the baseline. To further manifest the influence of the contrastive MIL loss, $\mathcal{L}_{\text{CLS}}$, mean self-similarity values as defined in (4) are provided for each respective region. As anticipated, a patch exhibiting a lower similarity score relative to its counterparts is indicative of the manipulated area. Also, a qualitative comparison is included to facilitate a visual assessment of heatmaps generated by the baseline (Multi-Cls), SeqFakeFormer [32], and our method in Figure 4. This comparison elucidates the enhanced localization precision by our contrastive MIL-infused formulation.

Effectiveness of $\mathcal{L}_{\text{CLS}}$. The evaluation of the contrastive MIL loss, denoted as $\mathcal{L}_{\text{CLS}}$, is a crucial aspect of the study. For visual clarification, Figure 5 features a histogram that represents the average distribution $\frac{1}{|\mathcal{D}|}\sum_{\mathcal{D}}\mathbf{u}$ across the FF++ test set. In Figure 5(a), the classifier demonstrates an ability to discern between authentic and fabricated facial imagery based on slight variations in the distribution—a task that is typically challenging for human observers. The integration of the $\mathcal{L}_{\text{CLS}}$ leads to a more distinct and simplified distribution demarcation between genuine and manipulated instances. An examination of Figure 5(b) from a different angle reveals that counterfeit facial images often arise from the combination of two authentic facial images. The regions that most commonly betray alteration are those at the facial

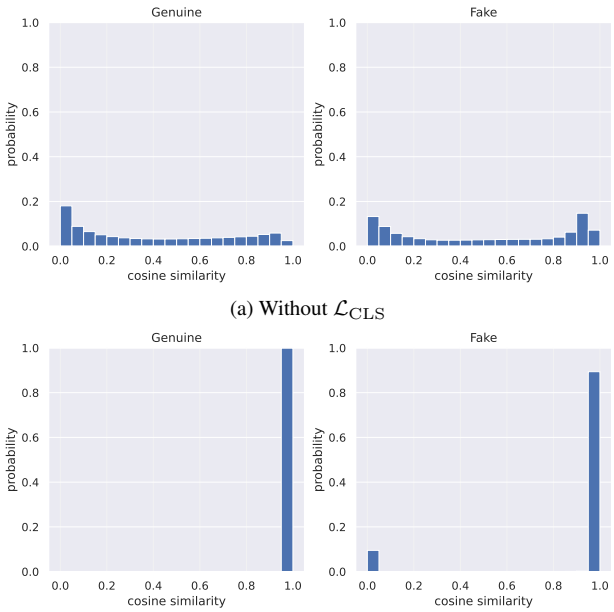

Figure 5. **The histogram of averaged distribution $\frac{1}{|\mathcal{D}|}\sum_{\mathcal{D}}\mathbf{u}$.** (a) The histograms from the baseline are like an "U" shape, no matter whether the images are genuine or fake. (b) With the contrastive MIL loss $\mathcal{L}_{\text{CLS}}$, we regularize the $\mathbf{u}$ close to 1 in genuine images and encourage the k values from $\mathbf{u}$ to approaching 0 in fake images.

boundaries or composite parts, with the central facial zones and extremities usually retaining their authenticity. Consequently, the extent of these altered regions is small in relation to the entirety of the image. This observation aligns with the MIL principle that a forged image will present falsification in a minimal number of key points, where $k \ll n$—meaning the number of these points is significantly less than the total number of points or regions in the image.

5. Conclusion

This work aims to develop a unified framework that comprehensively addresses both sequential deepfake manipulations and binary deepfake classification. To this end, we propose to decompose the general deepfake problem into three parts: deepfake classification, deepfake localization, and manipulation order. The proposed approach introduces novel contrastive MIL learning and explores multi-label ranking to elegantly tackle all three subtasks. The extended experimental results demonstrate the effectiveness and flexibility of the proposed formulation in dealing with the various deepfake application scenarios. The provided analyses are also reasonable to support the usefulness of our method.

Acknowledgements. This work was supported in part by NSTC grants 111-2221-E-001-015-MY3 and 112-2634-F-007-002 of Taiwan. We thank National Center for High-performance Computing for providing computing resources.

References

[1] Kenneth E. Batcher. Sorting networks and their applications. In *American Federation of Information Processing Societies: AFIPS Conference Proceedings: 1968 Spring Joint Computer Conference, Atlantic City, NJ, USA, 30 April - 2 May 1968*, pages 307–314. Thomson Book Company, Washington D.C., 1968. 3

[2] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4103–4112. IEEE, 2022. 1, 2, 6, 7

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 213–229. Springer, 2020. 6

[4] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18689–18698. IEEE, 2022. 2

[5] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 1081–1088. AAAI Press, 2021. 1, 2, 6, 7

[6] Yixin Chen, Jinbo Bi, and James Ze Wang. MILES: multiple-instance learning via embedded instance selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):1931–1947, 2006. 3

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1800–1807. IEEE Computer Society, 2017. 7

[8] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 15088–15097. IEEE, 2021. 2

[9] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5202–5211. Computer Vision Foundation / IEEE, 2020. 5

[10] Lihi Dery. Multi-label ranking: Mining multi-label and label ranking data. *CoRR*, abs/2101.00583, 2021. 3

[11] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1-2):31–71, 1997. 3

[12] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton-Ferrer. The deepfake detection challenge (DFDC) preview dataset. *CoRR*, abs/1910.08854, 2019. 7, 1

[13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9458–9468. IEEE, 2022. 2, 6

[14] Zhihao Gu, Taiping Yao, Yang Chen, Shouhong Ding, and Lizhuang Ma. Hierarchical contrastive inconsistency learning for deepfake video detection. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XII*, pages 596–613. Springer, 2022. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 5

[16] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2132–2141. PMLR, 2018. 3

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228, 2021. 1

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12):4217–4228, 2021. 1

[19] Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, and Youngjung Uh. Exploiting spatial dimensions of latent in GAN for real-time image editing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 852–861. Computer Vision Foundation / IEEE, 2021. 1

[20] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5000–5009. Computer Vision Foundation / IEEE, 2020. 2, 6, 7

[21] Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. Sharp multiple instance learning for deepfake video detection. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1864–1872. ACM, 2020. 2

[22] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 3204–3213. Computer Vision Foundation / IEEE, 2020. 7, 1

[23] Cong Han Lim and Steve Wright. A box-constrained approach for hard permutation problems. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2454–2463. JMLR.org, 2016. 3

[24] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 772–781. Computer Vision Foundation / IEEE, 2021. 2, 6, 7

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. 2

[26] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16317–16326. Computer Vision Foundation / IEEE, 2021. 6

[27] Felix Petersen, Christian Borgelt, Hilde Kuehne, and Oliver Deussen. Differentiable sorting networks for scalable sorting and ranking supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8546–8555. PMLR, 2021. 3

[28] Felix Petersen, Christian Borgelt, Hilde Kuehne, and Oliver Deussen. Monotonic differentiable sorting networks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 3

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[30] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 1–11. IEEE, 2019. 7, 1

[31] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society, 2017. 7, 3

[32] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and recovering sequential deepfake manipulation. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIII*, pages 712–728. Springer, 2022. 1, 2, 3, 5, 6, 8

[33] Rui Shao, Tianxing Wu, and Ziwei Liu. Robust sequential deepfake detection. *CoRR*, abs/2309.14991, 2023. 2, 3

[34] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18699–18708. IEEE, 2022. 1, 2, 6, 7

[35] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 6105–6114. PMLR, 2019. 7

[36] Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew F. K. Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXI*, pages 699–715. Springer, 2022. 3

[37] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Detecting photoshopped faces by scripting photoshop. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 10071–10080. IEEE, 2019. 6

[38] Ruiyang Xia, Decheng Liu, Jie Li, Lin Yuan, Nannan Wang, and Xinbo Gao. Mmnet: Multi-collaboration and multi-supervision network for sequential deepfake detection. *CoRR*, abs/2307.02733, 2023. 6

[39] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. DTFD-MIL: double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18780–18790. IEEE, 2022. 3

[40] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2185–2194. Computer Vision Foundation / IEEE, 2021. 1, 2, 6, 7

[41] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 15003–15013. IEEE, 2021. 1, 2, 6

[42] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14780–14789. IEEE, 2021. 2

[43] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z. Li. Face forgery detection by 3d decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2929–2939. Computer Vision Foundation / IEEE, 2021. 2

[44] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 2382–2390. ACM, 2020. 7, 1