Exploring Unbiased Deepfake Detection via Token-Level Shuffling and Mixing

Xinghe Fu[1*], **Zhiyuan Yan**[2*], **Taiping Yao**[2†], **Shen Chen**[2], **Xi Li**[1†]

[1]College of Computer Science and Technology, Zhejiang University
[2]Youtu Lab, Tencent
xinghefu@zju.edu.cn, {zhiyuanyan, taipingyao, kobeschen}@tencent.com, xilizju@zju.edu.cn

Abstract

The generalization problem is broadly recognized as a critical challenge in detecting deepfakes. Most previous work believes that the generalization gap is caused by the differences among various forgery methods. However, our investigation reveals that the generalization issue can still occur when forgery-irrelevant factors shift. In this work, we identify two biases that detectors may also be prone to overfitting: position bias and content bias, as depicted in Fig. 1. For the position bias, we observe that detectors are prone to "lazily" depending on the specific positions within an image (*e.g.*, central regions even no forgery). As for content bias, we argue that detectors may potentially and mistakenly utilize forgery-unrelated information for detection (*e.g.*, background, and hair). To intervene in these biases, we propose two branches for shuffling and mixing with tokens in the latent space of transformers. For the shuffling branch, we rearrange the tokens and corresponding position embedding for each image while maintaining the local correlation. For the mixing branch, we randomly select and mix the tokens in the latent space between two images with the same label within the mini-batch to recombine the content information. During the learning process, we align the outputs of detectors from different branches in both feature space and logit space. Contrastive losses for features and divergence losses for logits are applied to obtain unbiased feature representation and classifiers. We demonstrate and verify the effectiveness of our method through extensive experiments on widely used evaluation datasets.

Introduction

Deepfake technology has become prominent due to its ability to create impressively realistic visual content. However, this technology can also be exploited for harmful purposes, such as violating personal privacy, disseminating false information, and undermining trust in digital media. Considering these potential threats, there is an urgent need to develop a reliable deepfake detection system.

Most of earlier deepfake detectors (Li and Lyu 2018; Yang, Li, and Lyu 2019; Qian et al. 2020; Gu et al. 2021) demonstrate effectiveness in the scenario of within-dataset but often falter in cross-dataset scenarios where there is a distribution gap between the training and testing data. The prevalent explanation for this generalization problem, as suggested in previous works (Luo et al. 2021; Yan et al. 2023a), is models' overfitting to specific forgery. These works argue that generalization failure primarily occurs because the forgery methods applied in training and testing data are not identical, leading many subsequent studies (Yan et al. 2024a; Chen et al. 2022; Yan et al. 2024b) to address this issue from various methodological perspectives.

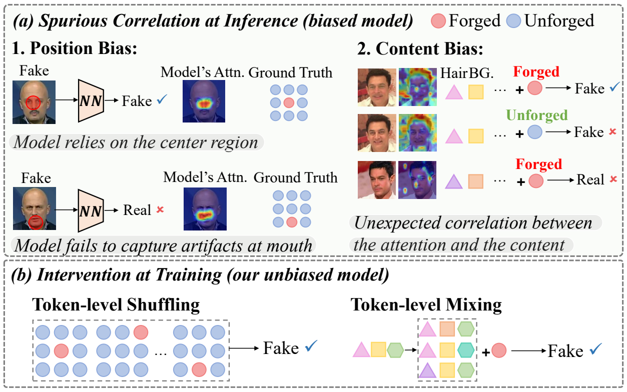
(a) Spurious Correlation at Inference (biased model)
Forged
Unforged
1. Position Bias:
Fake
NN
Fake ✓
Model's Attn.
Ground Truth
Model relies on the center region
Fake
NN
Real ×
Model's Attn.
Ground Truth
Model fails to capture artifacts at mouth
2. Content Bias:
Hair
BG.
Forged
Fake ✓
Unforged
Fake ×
Forged
Real ×
Unexpected correlation between the attention and the content
(b) Intervention at Training (our unbiased model)
Token-level Shuffling
Fake ✓
Token-level Mixing
Fake ✓

Figure 1: We present two identified biases in deepfake detection: (1) position bias and (2) content bias. (a) For position bias, we discover that the detector may focus more on the image center, regardless of whether the forgery is present. For content bias, we observe that detectors mistakenly concentrate on the forgery-irrelevant features (*e.g.,* background and hair). These biases can cause spurious correlations and lead to a biased detector; (b) Our method intervenes in the biases and helps establish a more robust model.

In this paper, we discover that the generalization problem persists even when identical forgery techniques are applied. Through our initial investigations, we identify two biases in deepfake detection: position bias and content bias. As illustrated in Fig. 1, we observe that, for position bias, detectors tend to "lazily" rely on the central region of the image for detection. Concerning content bias, we find that the detector might mistakenly and inadvertently use specific content combinations for detection, *e.g.*, background, hair, or

[*]These authors contributed equally.
[†]These are corresponding authors.

clothes. The two key observations inspire us to develop an unbiased detector capable of relying *less on biased information*, thereby creating more robust deepfake detectors.

To address these two biases, we introduce two plug-and-play strategies, namely the *shuffling branch* and *mixing branch*, which simply operate at the token level within ViTs. **Firstly**, in the shuffling branch, we implement a shuffling operation to rearrange the latent token order of a given image and shuffle its corresponding position embedding, to obtain the final shuffled representation. This method aims to disrupt the biased information (*e.g.,* ID (Dong et al. 2023)) by reorganizing the spatial position relationship. **Secondly**, in the mixing branch, we exchange a portion of the latent tokens between two images that have the same label within the same mini-batch. The rationale is that swapping certain tokens between two such images (with the same label) should not alter the original decision. Specifically, when only a limited number of tokens are interchanged, the remaining tokens should retain the key discriminative features while the content information is recombined.

Our solution presents two potential advantages compared to previous unbiased learning works (Liang, Shi, and Deng 2022; Yan et al. 2023a). **First**, these methods identify the part of the content bias problem and propose a disentanglement framework to overcome this bias through an implicit reconstruction learning process. However, they do not address the alleviation of position bias, which may occur even when no face is present in the image. In contrast, we propose an explicit unbiased learning method featuring two operations (*i.e.,* shuffling and mixing) designed to disrupt the biased context. **Second**, since our proposed methods specifically operate on latent tokens and are lightweight. they can be easily extended to any advanced ViT-based models, including recent state-of-the-art approaches (*e.g.,* CLIP). On the other hand, previous works (Liang, Shi, and Deng 2022; Yan et al. 2023a) are based on a fixed disentanglement framework with extra decoders for reconstruction, which is effective but may not be flexible enough for extension.

Our contributions are summarized as follows:

- We identify two critical factors except the forgery specificity that contribute to the generalization problem in deepfake detection: position bias and content bias.
- We propose an unbiased deepfake detection approach: UDD, to address position bias and content bias, which involves shuffling and mixing branch and alignment strategies.
- Extensive experiments show that our framework can outperform the performance of existing state-of-the-art methods in unseen testing datasets, demonstrating its effectiveness in generalization.

Related Work

General Deepfake Detection

The task of deepfake detection presents significant challenges, primarily in capturing the subtle traces of manipulation and enhancing the generalizability of detection models. Prior work in this area has focused on extending the data view, such as analyzing the frequency domain (Qian et al. 2020; Li et al. 2021, 2022) and leveraging specialized network modules (Zhao et al. 2021a; Dang et al. 2020; Song et al. 2022) to capture detailed forgery traces.

While promising detection performance is achieved, no constraints are presented in these methods that prevent model overfitting and enable the learning of generalized forgery information. To address the issue of generalization, researchers have employed synthesis and blending techniques in RGB images (Li et al. 2020a; Chen et al. 2022; Shiohara and Yamasaki 2022; Larue et al. 2023; Li and Lyu 2018). These methods (*e.g.*, SBIs (Shiohara and Yamasaki 2022) and SLADD (Chen et al. 2022)) reduce the content disparity between real and fake samples and encourage the model to learn common forgery artifacts for better generalization. Some methods (Sun et al. 2020; Yan et al. 2024a) also utilize augmentation and synthesis in the latent space to enrich the forgery samples and attain a more robust detection model. Other methods like UCF (Yan et al. 2023a) and IID (Dong et al. 2023) leverage generalized forgery features (like ID difference) in the detection with a specially designed learning framework.

Unlike previous methods, which focus on general forgery artifacts, we identify the position and content bias for the detection model and directly reduce the bias in the training data. Specifically, we design the random shuffling and mixing operations and apply them to both real and fake samples to obtain unbiased forgery representation in the proposed learning framework.

Unbiased Representation Learning

The goal of unbiased representation learning methods is to learn feature representations that are invariant to the biases presented in the training data. Data augmentation plays a crucial role in unbiased representation learning by introducing variability into the training process, which helps to prevent the model from overfitting to spurious correlations in the data. Some work (Mitrovic et al. 2020; Ilse, Tomczak, and Forré 2021) theoretically analyzes the effect of data augmentation in learning invariant features from the causal perspective. Recent advancements in this field have also explored the use of contrastive learning (Williams 1992; Oord, Li, and Vinyals 2018), where the model is trained to distinguish between similar (positive) and dissimilar (negative) pairs of data points. By doing so, the model learns to focus on the most discriminative features of the data, which are often less susceptible to bias. This approach has been particularly effective in learning representations that are general for various tasks (Rizve et al. 2021; Ba et al. 2024).

For general deepfake detection, we propose a unified unbiased learning framework to reduce the bias toward spurious correlations like position and content factors.

Method

Overview

The deepfake detection task is commonly modeled as a binary image classification problem. The sampling of an image is associated with latent variables of forgery, content,

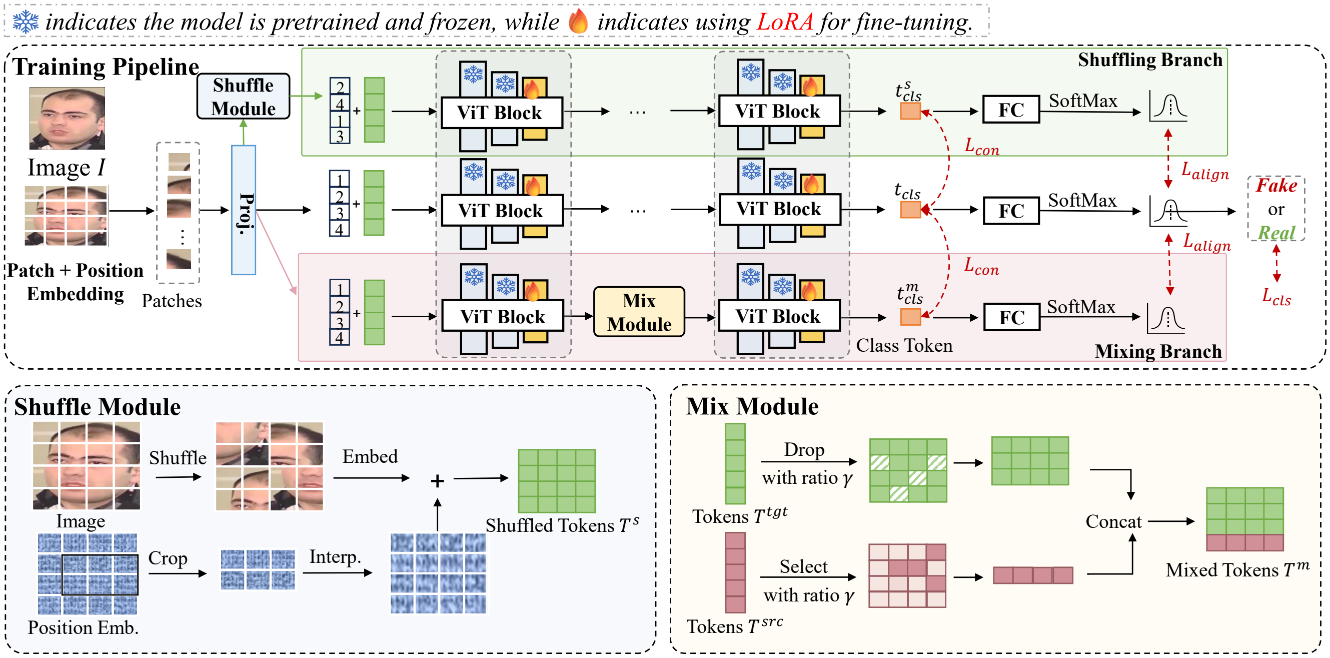

Figure 2: The overall pipeline of the proposed framework. The input image is sent to the original, shuffling and mixing branch during training. The shuffling branch (S-Branch) introduces the random intervention on position information with the shuffle module at the embedding layer. The mixing branch (M-Branch) introduces the intervention on content information with the mix module between randomly selected blocks. Both operations are applied to token-level representations in the latent space. All branches share the parameters of the network. Contrastive loss $\mathcal{L}_{con}$ and alignment loss $\mathcal{L}_{align}$ are applied over branches to attain unbiased forgery representation and classifier.

and position, while the sampling of labels is solely related to the forgery latent variable. Formally, let X denote an image, and Y its corresponding label, where $Y = 1$ indicates a forged image and $Y = 0$ indicates an authentic one. The sampling of the training set can be formulated as follows:

$$Y \sim p(Y|Z_f), X \sim p(X|Z_f, Z_b), \tag{1}$$

where Z_f represents the forgery variable, and $Z_b = (Z_c, Z_p)$ represents the content and position variables. The label Y is conditioned on the forgery latent variable Z_f, while the image X is conditioned on all three latent variables. This sampling and dataset construction approach may lead to the model capturing spurious correlations (in Fig. 1) between content Z_c, position Z_p, and labels Y, as the distribution of Z_c and Z_p are biased in the training set (*e.g.*, specific forged identity and cropping in the pre-processing). To mitigate the model's bias towards position and content in the learned forgery-related representations, we have designed two branches of augmentation operations based on the token-level latent space of transformer models. Specifically, the operations involve shuffling operations at the embedding layer and mixing operations in the forwarding process. These operations disrupt and recombine the position information and image content, thereby enhancing the randomness in the sampling over these two latent variables and blocking the spurious correlations between them and labels.

The overall framework includes the token-shuffling branch and the token-mixing branch during forwarding, and the feature and logit level alignment loss to attain the unbiased representation and classifier. A causal analysis of the framework is also provided.

Token-Shuffling Branch

To enhance the randomness of the forged position distribution within images and reduce the model's reliance on specific locations, we introduce a shuffling module at the embedding layer of vision transformers (ViTs) to form the token-shuffling branch (in Fig. 2). This module comprises two parts: one that performs random rectangular sampling and interpolation on the patch token position encodings, and another that executes blockwise shuffling of the correspondence between patch tokens and their position encodings.

Random interpolated position embeddings pos'_i. To introduce randomness of absolute position and scaling into the position embeddings pos_i (Kim, Angelova, and Kuo 2023; Yuan et al. 2023), we reshape the patch position embeddings into a rectangle (*e.g.*, 14×14) and perform the following operations. 1) We first sample the aspect ratio r, area S, and location (x, y) of the rectangle from a uniform distribution. The area is kept larger than 30% of the whole rectangular area of position embeddings. 2) After sampling, we crop the position embeddings according to the obtained rectangle. 3) The cropped local position embeddings are then interpolated and flattened as the random interpolated position embeddings pos'_i.

Shuffled patch tokens $e_{\pi(i)}$. To enhance the relative position randomness of forgery traces, we introduce random shuffling of patch tokens. Since forgery traces depend on the content of local regions, the shuffling is performed in a blockwise manner and the local correlation within each block is maintained. We reshape the patch tokens into a rectangle and divide it into $s \times s$ blocks. The blocks are then randomly permuted, creating a mapping from the original index i of each patch to a new index $\pi(i)$.

After performing these operations, we add the shuffled patch tokens to the new position embeddings to obtain the token embeddings $t_i \in \mathbb{R}^D$,

$$t_i = pos'_i + e_{\pi(i)}. \tag{2}$$

The shuffling module approximates a $do(\cdot)$-operator on the positional latent variable Z_p and output the token set $T^s = \{t_i\}_{i=1}^{N+1}$ (class token included). The shuffle module introduces randomness in both relative and absolute position information while preserving the locality of forgery information. This helps the model reduce the bias towards positions.

Token-Mixing Branch

To enhance the randomness of sample content and enable the model to extract unbiased forgery features across different contexts, we designed a token-level content mixing module for the forward process. The mixing module consists of two steps: random dropping target token sets and mixing source and target token sets.

Random token dropout. For a given set of tokens T_l at layer l during the forward pass, we randomly drop a proportion γ of the patch tokens. Let the feature representation of the target image X at layer l be $T_l^{tgt} = \{t_i\}_{i=1}^{N+1}$, with the class token denoted as $t_{cls} = t_{N+1}$. We sample a subset of indices $ID = \{a_i\}$, where $1 \leq a_i \leq N$ and $p(i \in ID) = 1 - \gamma$. The feature representation after token dropout is $T_l^{tgt'} = \{t_i | i \in ID \vee i = N+1\}$. Assuming that forgery features are local and that tokens have undergone global feature exchange in the latent space, the remaining tokens still represent partial forgery features of the image.

Source & target mixing. For the target features $T_l^{tgt'}$ after dropout, we further select source features T_l^{src} of another sample ($Y^{tgt} = Y^{src}$) from the batch for the mixing operation. As tokens have a global receptive field in the latent space, they can capture global content information. By mixing tokens from different sources in the latent space, we recombine the contextual information of the features, alleviating the risk that forgery features do not rely on some specific facial context. Specifically, we randomly select tokens from the source feature set T_l^{src} to form a subset $T_l^{src'}$, where $|T_l^{src'}| = \gamma N$. We then merge $T_l^{src'}$ and $T_l^{src'}$ to form the mixed token set $T_l^m = T_l^{tgt'} \cup T_l^{src'}$ and use this token set for the subsequent forwarding process.

We implement the token-mixing branch by inserting the mixing module at randomly selected layers in the forwarding process. In the token-mixing branch, we achieve a random combination of content contexts, approximating a $do(\cdot)$-operator on the content latent variable Z_c, introducing content randomness.

Overall Framework

Building upon the two operation branches for position and image content intervention, we have designed an unbiased forgery representation learning framework including the tuning architecture and loss functions. In the framework, vision transformers are trained to capture the causal correlation between input images $X \in \mathbb{R}^{3 \times H \times W}$ and forgery labels $Y \in \{0, 1\}$ with less bias towards positions and image content in the three-branch data flows.

Architecture. It is important to retain and utilize the knowledge for deepfake detection tasks. Therefore, we freeze the backbone parameters and introduce only a small number of learnable parameters (*e.g.*, LoRA (Hu et al. 2021)). For forged images, the model is required to focus on the forged regions within the attention mechanism and encode forgery-related information during the forward process. Hence, we incorporate learnable parameters in both the multi-head attention and MLP layers. Taking the multi-head attention in transformers as $MSA(\cdot; \{W_a\})$ and the MLP layer in transformers as $MLP(\cdot; \{W_m\})$, where W_a and W_m represent arbitrary pre-trained linear projection matrix in MSA and MLP, the forwarding process of a transformer block with additional learnable parameters is as follows:

$$T'_l = MSA(T_l; \{W_a + \Delta W_a\}) + T_l, \tag{3}$$

$$T_{l+1} = MLP(T'_l; \{W_m + \Delta W_m\}) + T'_l, \tag{4}$$

where $T_l \in \mathbb{R}^{N \times L \times D}$ denotes the input tensor of the l-th transformer block, ΔW_a and ΔW_m denote the low-rank matrix. Suppose that the rank of $\Delta W \in \mathbb{R}^{d_1 \times d_2}$ (ΔW_a or ΔW_m) is r, we decompose ΔW as $\Delta W = AB^T$ and take low-rank matrices $A \in \mathbb{R}^{d_1 \times r}$ and $B \in \mathbb{R}^{d_2 \times r}$ (as in LoRA) as learning parameters during training. It turns out that the design not only achieves parameter-efficient and stable training but also alleviates the overfitting risk in the detection.

Learning Objective. For the loss function design, we align the output from the original branch with the token-shuffling and token-mixing branches on two levels (**feature level** and **logit level**) to obtain unbiased feature representations and classifiers. At the feature level, we employ the contrastive loss function (as in SimCLR (Chen et al. 2020)), aligning the class token t_{cls} with t_{cls}^s and t_{cls}^m as positive sample pairs after mapping through a three-layer MLP projector $g(\cdot)$. The feature-level contrastive loss is as follows:

$$\mathcal{L}_c(t', t'^{+}) = -\log \frac{e^{\mathrm{sim}(t', t'^{+})/\tau}}{e^{\mathrm{sim}(t', t'^{+})/\tau} + \sum_{t'^{-}} e^{\mathrm{sim}(t', t'^{-})/\tau}}, \tag{5}$$

$$\mathcal{L}_{con} = \mathcal{L}_c(g(t_{cls}), g(t_{cls}^s)) + \mathcal{L}_c(g(t_{cls}), g(t_{cls}^m)), \tag{6}$$

where $t' = g(t_{cls})$ and $\mathrm{sim}(x, y) = x^T y / ||x||_2 ||y||_2$. We take other samples within a batch after $g(\cdot)$ as negative samples in the contrastive loss. Since t_{cls}, t_{cls}^s, and t_{cls}^m contain similar forgery features, this loss enables aligned feature representations. At the logit level, we impose constraints on the prediction probabilities using Jensen–Shannon divergence here and align the predicted logits after the classifier:

$$\mathcal{L}_{align} = D_{\mathrm{JS}}(P_{Y|t_{cls}} || P_{Y|t_{cls}^s}) + D_{\mathrm{JS}}(P_{Y|t_{cls}} || P_{Y|t_{cls}^m}) \tag{7}$$

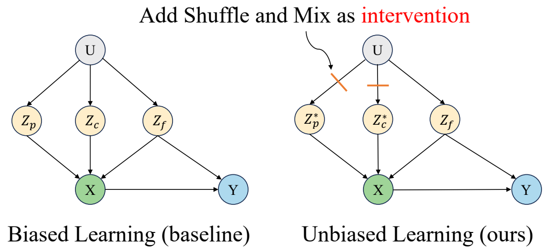

Figure 3: Causal graph for illustrating the proposed framework (the right) versus the baseline (the left). Within the graph, X and Y are observed. The unobserved confounder U causes a backdoor path from the position (Z_p) and content (Z_c) variables to the label Y. Differing from the baseline, our proposed unbiased learning method performs an intervention that blocks the backdoor paths for training an unbiased detector.

The overall loss function is as follows (given that $\mathcal{L}_{ce}$ is the cross-entropy loss for classification):

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{con} + \lambda_2 \mathcal{L}_{align} \tag{8}$$

Causal Analysis

In this subsection, we aim to perform an analysis from the view of causal learning to illustrate the effectiveness of the proposed strategies. In the previous sections, we show the existence of *spurious correlations* (see Fig. 1) when naively training a model to detect deepfakes. From the view of causal learning, our goal is to design an intervention to mitigate such spurious correlations in the original data distribution. In this work, we observe two main factors that contribute to this spurious correlation (*i.e.,* position bias and content bias). Thus, it is important to disentangle these forgery-unrelated biases (confounders) from the forgery-related causal features.

The *causal graph* is shown in Fig. 3 for unbiased deepfake detection. We denote the unobserved *confounder* as U, which produces forgery-irrelevant nuisance factors (*i.e.,* Z_p and Z_c) and forgery-related (causal) feature (*i.e.,* Z_f). For convenience, we use Z_b to represent the forgery-irrelevant bias (including Z_p and Z_c in our case). The correlation between Z_b and Y is mediated through U, forming a backdoor[1] path $X \leftarrow Z_b \leftarrow U \rightarrow Z_f \rightarrow Y$. The existence of this backdoor path leads to a possible bias towards Z_b during training, which hinders the learning of correct causal path $X \leftarrow Z_f \rightarrow Y$. To achieve unbiased deepfake detection, we need to intervene on Z_b and break the backdoor path.

Here, we use the *"do"-operator* [2] (formally expressed as $do(X = x)$) to denote performing a treatment z on the variable Z_b which makes Z_b independent from the causal variable Z_f ($Z_b \perp\!\!\!\perp Z_f$) (Pearl 2009). We denote the causality from X to Y predicted by networks to be $P_\theta(Y|X)$, which is the treatment effect of an input image X on label Y. Removing the backdoor, the correlation learned in the framework is now equal to the causality, *i.e.,* $P_\theta(Y|X) = P_\theta(Y|Z_f, Z_b) = P_\theta(Y|Z_f, do(Z_b))$. The proposed shuffling and mixing operations can be regarded as two different interventions added at the token level to the image. This enables the model to remove the spurious correlation (position bias and content bias) and learn more general forgery features for deepfake detection.

Experiments

Setup

Datasets. For comprehensively assess the proposed method, we utilize five widely-used public datasets FaceForensics++ (FF++) (Rossler et al. 2019), Celeb-DF (CDF) (Li et al. 2020b), DFDC (Dolhansky et al. 2020), DFDC-Preview (DFDCP) (Dolhansky et al. 2019), and DFD (Deepfakedetection 2021) in our experiments, following previous works (Mohseni et al. 2020; Yan et al. 2023c). FF++ dataset is the most widely used dataset in deepfake detection tasks. We use the training split of FF++ (Rossler et al. 2019) as the training set. For the cross-dataset evaluation, we test our model on datasets other than FF++. For the robustness evaluation, we test our model on the test split of FF++.

Implementation details. We utilize the ViT-B model as the backbone network of detectors. The backbone is initialized with the pre-trained weights from the vision encoder of CLIP (Radford et al. 2021) by default. We evenly sample 8 frames (Mohseni et al. 2020) from the training videos of FF++ (c23) to form the training set. For S-Branch, we divide the token map into 2×2 blocks in shuffling. For M-Branch, the mixing ratio r is set to 0.3. The rank for ΔW is set to 4. The hyperparameters τ, λ_1, and λ_2 in loss functions are set to 0.1, 0.1, and 0.1 in the training. S-Branch and M-Branch are not applied at inference.

Metrics. We utilize area under receiver operating characteristic curve (AUC) scores for empirical evaluation in experiments. Frame-level AUC is calculated based on the predicted scores of frame inputs. We calculate the video-level AUC with averaged predicted scores of 32 frames sampled from a video. Video-level AUC scores are reported in experiments unless otherwise specified.

Comparison with Existing Methods

Generalization evaluation. In the cross-dataset evaluation part, models are trained on the FF++ (c23) dataset and evaluated on CDF, DFDCP, DFDC, and DFD datasets. We report both frame-level and video-level AUC scores in Table 1 and compare our methods with previous state-of-the-art methods, *e.g.*, LSDA (Yan et al. 2024a) and TALL++ (Xu et al. 2024). To achieve great improvement in generalization, we focus on reducing the overfitting or bias to forgery-irrelated information like position or content. Our method achieves state-of-the-art results in both frame-level and video-level AUC scores. For frame-level AUC scores, our method

[1] With a backdoor, classifiers trained directly on X and Y will not be causal with Z_f.

[2] The "do"-operator formalizes the process of intervening in a system. In contrast to conventional statistical methods that demonstrate correlation, the "do"-operator enables us to model the consequences of actively manipulating a variable.

Type	Method	CDF	DFDCP	DFDC	DFD	Type	Method	CDF	DFDCP	DFDC	DFD
Frame-Level	Xception	73.7	73.7	70.8	81.6	Video-Level	Xception	81.6	74.2	73.2	89.6
	Efficient-b4	74.9	72.8	69.6	81.5		Efficient-b4	80.8	68.0	72.4	86.2
	FWA	66.8	63.7	61.3	74.0		LipForensics	82.4	-	73.5	-
	Face X-ray	67.9	69.4	63.3	76.7		FTCN	86.9	74.0	71.0	94.4
	RECCE	73.2	74.2	71.3	81.2		RECCE	82.3	73.4	69.6	89.1
	F3-Net	73.5	73.5	70.2	79.8		F3-Net	78.9	74.9	71.8	84.4
	SPSL	76.5	74.1	70.4	81.2		PCL+I2G	90.0	74.4	67.5	-
	SRM	75.5	74.1	70.0	81.2		SBIs*	90.6	87.7	75.2	88.2
	UCF	73.5	73.5	70.2	79.8		UCF	83.7	74.2	77.0	86.7
	IID	83.8	81.2	-	-		SeeABLE	87.3	86.3	75.9	-
	ICT	85.7	-	-	84.1		UIA-ViT	82.4	75.8	-	94.7
	LSDA	83.0	81.5	73.6	88.0		TALL++	92.0	-	78.5	-
	ViT-B (IN21k)	75.0	75.6	73.4	86.4		ViT-B (IN21k)	81.7	77.7	76.3	89.5
	ViT-B (CLIP)	81.7	80.2	73.5	86.6		ViT-B (CLIP)	88.4	82.5	76.1	90.0
	UDD (Ours)	**86.9**	**85.6**	**75.8**	**91.0**		UDD (Ours)	**93.1**	**88.1**	**81.2**	**95.5**

Table 1: Comparison with previous methods. We report both frame-level and video-level AUC (%) of our models trained on FF++ (c23) and compare the results with previous SOTA methods. Methods with * are our reproduction results using the released models. **Bold** and underline indicate the best and the second-best results. We report and cite the results of other methods from their original papers or DeepfakeBench (Yan et al. 2023b).

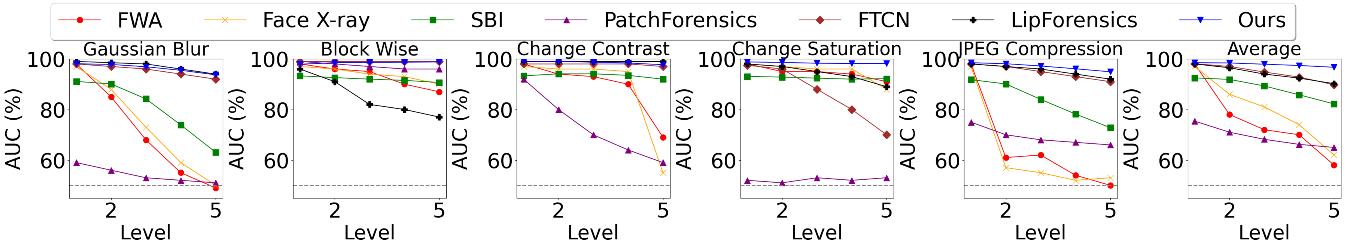

Figure 4: The results of robustness evaluation on the test set of FF++ (c23). Video-level AUC (%) is reported under five different types of perturbations following (Jiang et al. 2020). Our method is more robust than previous methods across corruptions.

achieves 86.9% on CDF and outperforms ViT-based state-of-the-arts method ICT (Dong et al. 2022) and frequency-based methods (Qian et al. 2020; Li et al. 2021; Luo et al. 2021). Our method also achieves improvement on DFDCP by 4.4% (from 81.2% to 85.6%), DFDC by 2.3% (from 73.5% to 75.8%), and DFD (from 88.0% to 91.0%), outperforming RECCE (Cao et al. 2022), IID (Huang et al. 2023) and LSDA (Yan et al. 2024a). For video-level AUC scores, our method surpasses previous state-of-the-art methods with image augmentation and synthesis *i.e.*, PCL+I2G (Zhao et al. 2021b) and SeeABLE (Larue et al. 2023) on CDF (from 90.0% to 93.1%) and DFDCP (from 87.7% to 88.1%). On DFDC and DFD datasets, our method outperforms the second-best method TALL++(from 78.5% to 81.2%) and UIA-ViT (Zhuang et al. 2022) (from 94.7% to 95.5%). The comparison results demonstrate the generalization capability of the proposed method and verify the importance of unbiased forgery representation learning.

Robustness evaluation. To verify the robustness of the proposed method, different types of perturbations from (Jiang et al. 2020) are applied to the test set of FF++ (c23). Video-level AUC scores are reported at different perturbation levels (level 5 is the most severe) in Fig. 4. Previous methods (Haliassos et al. 2021; Zheng et al. 2021; Chai et al. 2020) achieve great detection performance with full-training. However, these methods may capture abundant low-level forgery clues that are sensitive to perturbations like noise and compression Furthermore, CNN-based methods (Tan and Le 2019) are obstructed by some specific perturbations like blockwise noise from Fig. 4, while transformer models can easily overlook these noises in the attention mechanism. Our method achieves the highest averaged AUC scores in the robustness evaluation and is less sensitive to severe perturbations of different types than previous state-of-the-art methods.

Ablation Study

Ablation of components. We study the generalization effect of the proposed components with cross-dataset evaluation in Table 2. Both the shuffling branch (S-Branch) and the mixing branch (M-Branch) play a key role in achieving state-of-the-art generalization results. Quantitatively, video-level AUC scores improve from 90.70% (without both) to 93.13% (with both) on CDF, and from 78.25% (without

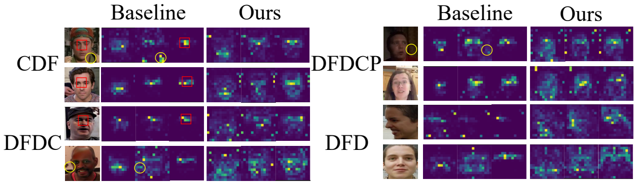

Figure 5: The visualization of multi-head attentions. For visualization, we select attention maps with clear activation in the last layer of ViT models. Squares and circles indicate the position bias and content bias in the baseline model.

Model	CDF	DFDCP	DFDC	DFD
Baseline	88.41	82.46	76.07	90.04
$+\Delta W$	90.70	85.97	78.25	93.08
w/o S-Branch	91.25	87.30	81.02	94.95
w/o M-Branch	90.51	87.68	80.33	92.25
w/o $\mathcal{L}_{con}$	91.20	86.35	79.74	94.03
w/o $\mathcal{L}_{align}$	90.52	87.01	80.18	94.62
Ours	**93.13**	**88.11**	**81.21**	**95.51**

Table 2: Ablation of the proposed components. Video-level AUC (%) is reported for comparison.

both) to 81.21% (with both) on DFDC. The results indicate that alleviating the position and content bias during training is important to the generalization capability of the detection model. Furthermore, The AUC scores of our method drop by over 1% on CDF, DFDCP, and DFDC when the contrastive loss $\mathcal{L}_{con}$ or the alignment loss $\mathcal{L}_{align}$ is not applied. Our method exhibits a significant advantage over the baseline model in the generalization capability when these components are applied (at least 4.72% improvement).

Ablation of pre-trained weights. We study the influence of pre-trained weights for the backbone networks in Table 3. The results demonstrate that the selection of the pre-trained model is also significant to the generalization capability in deepfake detection. Pre-trained weights from CLIP show great performance improvements on some datasets (*e.g.*, CDF) compared with weights trained on ImageNet (IN21k). However, the improvement brought by our method is insensitive to the selection of backbones from Table 3. The AUC scores increase from 81.66% (w/o ours) to 88.56% (w/ ours) on CDF and from 77.72% (w/o ours) to 83.84% (w/ ours) on DFDCP when using ImageNet pre-trained weights.

Ablation of mixing stages. To identify the appropriate position to insert mixing modules in the M-Branch, we conduct the ablation on the M-Branch and the results are shown in Table 4. We divide the forwarding process (from the 1st to the 2nd-last block) of the ViT-B model into three stages. The mixing module is randomly applied after one of the blocks in a specified stage during forwarding. Since the content information contained in image patches is limited, mixing before the first block (the first row in Table 4) cannot effectively introduce the content information from other images and achieves sub-optimal generalization performance (the lowest averaged AUC score 87.17%). In the forwarding process, tokens perceive the information over the whole image with the attention mechanism and capture high-level content semantics in the later blocks. However, mixing in the late stage shows less effect on the final output. Using blocks in the mid-stage for mixing achieves the best overall performance (from 87.17% to 89.07% in the avg. AUC).

Model	CDF	DFDCP	DFDC	DFD
ViT-B (IN21k)	81.66	77.72	76.30	89.82
+Ours	**88.56**	**83.84**	**76.50**	**92.73**
ViT-B (CLIP)	88.41	82.46	76.07	90.04
+Ours	**93.13**	**88.11**	**81.21**	**95.51**

Table 3: Ablation of pre-trained weights. Video-level AUC (%) is reported for comparison.

Model	CDF	DFDC	DFD	Avg.
Before First	92.44	76.82	92.25	87.17
Early	**92.55**	78.50	94.72	88.59
Mid	91.25	**81.02**	**94.95**	**89.07**
Late	91.02	79.06	93.44	87.84

Table 4: Ablation of mixing stages. Video-level AUC (%) is reported for comparison. "Before first, early, mid, and late" indicate different positions where the mix module is inserted in the token-mixing branch.

Visualization of attention maps. We visualize the attention map for fake samples (class tokens as queries) of the last attention layer in transformers for better viewing of the position and content bias in generalized deepfake detection. To better visually demonstrate, we selected attention maps with clear responses from three heads for display and compared our model with the baseline in Fig. 5. Red and yellow circles highlight the position and content bias in the attention maps of the baseline model. It is observed that the baseline model relies on some specific regions (*e.g.*, upper center) and content information (*e.g.*, clothes or background) for detection. At the same time, our method captures diverse forgery artifacts in the attention maps and relies less on biases.

Conclusion

This paper identifies position and content biases that can cause the generalization problem in deepfake detection. To mitigate these biases, we introduce UDD, a plug-and-play detection approach that learns unbiased forgery representations from a causal perspective. To achieve the intervention on the position and content, we propose the token-level shuffling and mixing branches for transformers. We also introduce feature-level contrastive loss and logit-level alignment loss to acquire unbiased feature representation and classifiers. Extensive experiments confirm the generalization capability and robustness of the proposed approach.

Acknowledgements

This work is supported in part by National Natural Science Foundation of China under Grant No.62441602, National Science Foundation for Distinguished Young Scholars under Grant 62225605, Zhejiang Provincial Natural Science Foundation of China under Grant LD24F020016, Project 12326608 supported by NSFC, "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C01020), and the Fundamental Research Funds for the Central Universities.

References

Ba, Z.; Liu, Q.; Liu, Z.; Wu, S.; Lin, F.; Lu, L.; and Ren, K. 2024. Exposing the deception: Uncovering more forgery clues for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 719–728.

Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; and Yang, X. 2022. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4113–4122.

Chai, L.; Bau, D.; Lim, S.-N.; and Isola, P. 2020. What makes fake images detectable? understanding properties that generalize. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, 103–120. Springer.

Chen, L.; Zhang, Y.; Song, Y.; Liu, L.; and Wang, J. 2022. Self-supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18710–18719.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; and Jain, A. K. 2020. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Deepfakedetection. 2021. https://ai.googleblog.com/2019/09/contributing-data-to-deepfakedetection.html Accessed 2021-11-13.

Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; and Ferrer, C. C. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

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

Dong, S.; Wang, J.; Ji, R.; Liang, J.; Fan, H.; and Ge, Z. 2023. Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3994–4004.

Dong, X.; Bao, J.; Chen, D.; Zhang, T.; Zhang, W.; Yu, N.; Chen, D.; Wen, F.; and Guo, B. 2022. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9468–9478.

Gu, Z.; Chen, Y.; Yao, T.; Ding, S.; Li, J.; Huang, F.; and Ma, L. 2021. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, 3473–3481.

Haliassos, A.; Vougioukas, K.; Petridis, S.; and Pantic, M. 2021. Lips Don't Lie: A Generalisable and Robust Approach To Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q.; Wang, Q.; and Ye, D. 2023. Implicit Identity Driven Deepfake Face Swapping Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4490–4499.

Ilse, M.; Tomczak, J. M.; and Forré, P. 2021. Selecting data augmentation for simulating interventions. In *International conference on machine learning*, 4555–4562. PMLR.

Jiang, L.; Li, R.; Wu, W.; Qian, C.; and Loy, C. C. 2020. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kim, D.; Angelova, A.; and Kuo, W. 2023. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11144–11154.

Larue, N.; Vu, N.-S.; Struc, V.; Peer, P.; and Christophides, V. 2023. SeeABLE: Soft Discrepancies and Bounded Contrastive Learning for Exposing Deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21011–21021.

Li, J.; Xie, H.; Li, J.; Wang, Z.; and Zhang, Y. 2021. Frequency-aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, J.; Xie, H.; Yu, L.; and Zhang, Y. 2022. Wavelet-enhanced weakly supervised local feature learning for face forgery detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1299–1308.

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020a. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, Y.; and Lyu, S. 2018. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020b. Celeb-df: A new dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liang, J.; Shi, H.; and Deng, W. 2022. Exploring Disentangled Content Information for Face Forgery Detection. In *Proceedings of the European Conference on Computer Vision*, 128–145. Springer.

Luo, Y.; Zhang, Y.; Yan, J.; and Liu, W. 2021. Generalizing Face Forgery Detection with High-frequency Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mitrovic, J.; McWilliams, B.; Walker, J.; Buesing, L.; and Blundell, C. 2020. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*.

Mohseni, S.; Pitale, M.; Yadawa, J.; and Wang, Z. 2020. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pearl, J. 2009. *Causality*. Cambridge university press.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of the European Conference on Computer Vision*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rizve, M. N.; Khan, S.; Khan, F. S.; and Shah, M. 2021. Exploring complementary strengths of invariant and equivariant representations for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10836–10846.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision*.

Shiohara, K.; and Yamasaki, T. 2022. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18720–18729.

Song, L.; Li, X.; Fang, Z.; Jin, Z.; Chen, Y.; and Xu, C. 2022. Face forgery detection via symmetric transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4102–4111.

Sun, H.; Li, C.; Liu, B.; Liu, Z.; Wang, M.; Zheng, H.; Feng, D. D.; and Wang, S. 2020. AUNet: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms. *Physics in Medicine & Biology*, 65(5): 055005.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, 6105–6114. PMLR.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *ML*, 8(3): 229–256.

Xu, Y.; Liang, J.; Sheng, L.; and Zhang, X.-Y. 2024. Learning Spatiotemporal Inconsistency via Thumbnail Layout for Face Deepfake Detection. *International Journal of Computer Vision*, 1–18.

Yan, Z.; Luo, Y.; Lyu, S.; Liu, Q.; and Wu, B. 2024a. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8984–8994.

Yan, Z.; Yao, T.; Chen, S.; Zhao, Y.; Fu, X.; Zhu, J.; Luo, D.; Wang, C.; Ding, S.; Wu, Y.; et al. 2024b. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*.

Yan, Z.; Zhang, Y.; Fan, Y.; and Wu, B. 2023a. UCF: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision*, 22412–22423.

Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; and Wu, B. 2023b. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. In Oh, A.; Neumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 4534–4565. Curran Associates, Inc.

Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; and Wu, B. 2023c. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. *Advances in Neural Information Processing Systems*.

Yang, X.; Li, Y.; and Lyu, S. 2019. Exposing Deep Fakes Using Inconsistent Head Poses. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Yuan, Y.; Fu, X.; Yu, Y.; and Li, X. 2023. DenseDINO: boosting dense self-supervised learning with token-based point-level consistency. *arXiv preprint arXiv:2306.04654*.

Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; and Yu, N. 2021a. Multi-attentional Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhao, T.; Xu, X.; Xu, M.; Ding, H.; Xiong, Y.; and Xia, W. 2021b. Learning Self-Consistency for Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision*.

Zheng, Y.; Bao, J.; Chen, D.; Zeng, M.; and Wen, F. 2021. Exploring Temporal Coherence for More General Video Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision*, 15044–15054.

Zhuang, W.; Chu, Q.; Tan, Z.; Liu, Q.; Yuan, H.; Miao, C.; Luo, Z.; and Yu, N. 2022. UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European Conference on Computer Vision*, 391–407. Springer.