VL-FAS: DOMAIN GENERALIZATION VIA VISION-LANGUAGE MODEL FOR FACE ANTI-SPOOFING

Hao Fang\textsuperscript{1,2} Ajian Liu\textsuperscript{2} Ning Jiang\textsuperscript{3} Quan Lu\textsuperscript{3} Guoqing Zhao\textsuperscript{3} Jun Wan\textsuperscript{1,2,⋆}

\textsuperscript{1}School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
\textsuperscript{2}MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China
\textsuperscript{3}Mashang Consumer Finance Co., Ltd., Chongqing, China

ABSTRACT

Recent approaches have demonstrated the effectiveness of Vision Transformer (ViT) with attention mechanisms for domain generalization of Face Anti-Spoofing (FAS). However, current attention algorithms highlight all the salient objects (e.g., background objects, hair, glasses), which results in the feature learned by the model containing face-irrelevant noisy information. Inspired by existing Vision-language works, we propose the \textbf{VL-FAS} to extract more generalized and cleaner discriminative features. Specifically, we leverage fine-grained natural language descriptions of the face region to act as a task-oriented teacher, directing the model’s attention towards the face region through top-down attention regulation. Furthermore, to enhance the domain generalization ability of the model, we propose a Sample-Level Vision-Text optimization module (SLVT). SLVT uses sample-level image-text pairs for contrastive learning, allowing the visual coder to comprehend the intrinsic semantics of each image sample, thereby reducing the dependence on domain information. Extensive experiments show that our approach significantly outperforms the state-of-the-art and improves the performance of the ViT by about twice.

Index Terms— Face Anti-spoofing, Attention Regulation, Vision Language, Domain Generalization

1. INTRODUCTION

FAS \cite{1, 2} is critical in protecting face recognition systems from various presentation attacks \cite{3, 4}, e.g., printed photos, 3D masks, etc. In recent years, with the introduction of hand-craft descriptors based methods \cite{5, 6} and deep learning based methods \cite{7, 3}, FAS methods have been able to achieve impressive performance in intra-domain scenarios where the train and test distributions are similar. However, variations due to camera sensors, illumination conditions, and image resolutions result in a large domain gap between the source and target domains. Consequently, existing FAS methods exhibit limited generalization capacity when adapting to diverse unseen domains.

Recently, several studies have introduced domain generalization (DG) methods to FAS tasks. DG methods aim to learn domain-agnostic discriminative features from multiple source domains, which can be well generalized to unseen domains such as adversarial learning-based methods \cite{8, 9} and meta-learning-based method \cite{10}. In addition, several studies \cite{11, 12} have demonstrated the effectiveness of vision transformer (ViT) for cross-domain FAS tasks. Specifically, ViTs capture long-range dependencies between different patches with the self-attention mechanism, so they can extract local information in these patches and aggregate it globally to make well-considered discrimination. However, these approaches have two major limitations: (1) \textbf{The widely used self-attention mechanism is entirely stimulus-driven} \cite{13}. It highlights all salient objects in the image that are not task-oriented, as shown in the upper part of Fig. 1. This results
Fig. 2. The modules of our VL-FAS. (1) The Language Teacher module utilizes MiniGPT-4 to generate task-oriented captions and encode them into text features. (2) The Attention Regulation module incorporates the textual information into the visual feature $x_{\text{forward-1}}$ of the first forward propagation and decodes it into $x_{\text{decode}}$. In the second forward propagation, $x_{\text{decode}}$ is added to the value matrix of the attention layer. (3) The Sample-Level VL Optimization module combines task-oriented captions with category text as sample-level pairs. It enables the visual encoder to understand sample-level intrinsic semantics through contrastive learning.

2. RELATED WORK

2.1. Attention mechanism.

Some recent work [11, 12] have demonstrated the effectiveness of attention mechanism in Face Anti-spoofing (FAS) tasks. For humans, attention in visual tasks is naturally directed towards objects relevant to the task [18]. For example, in determining the authenticity of a face image, our attention is primarily focused on the face to gather crucial information for decision-making. Several studies have found that top-down visual attention is vital in improving task-specific performance [19, 20]. Therefore, in the FAS task with faces as the primary target, we introduce a top-down attention regulation mechanism to improve the robustness and generalization of the model. Specifically, we used fine-grained textual descriptions of face regions as teachers to guide the attention mechanisms in performing top-down regulation. In this way, the visual encoder will focus on the face region, thus capturing more generalized and cleaner discriminative features. Besides, we construct sample-level vision text pairs for contrastive learning. Sample-level text information can help the visual coder comprehend the intrinsic semantics of each image. In this way, the visual coder will break the limitation of domain labeling and extract sample-level discriminative features through the intrinsic differences of the samples.

2.2. Cross-domain FAS.

Recently, some methods introduced domain adaptation techniques [21, 22] into FAS. This technique requires access to some data of the target domain for training. However, the picture of the target domain is often inaccessible in practical applications. Therefore, domain-generalizable FAS (DG-FAS) [8, 23] becomes the research focus. Among them, domain-adversarial learning [24] utilizes generators and discriminators to learn domain invariant feature representations. The contrast learning [9, 25, 3] utilizes specific properties as contrast (anchor points) to learn irrelevant features. However, all these methods rely on manual domain labeling of image data, are limited to coarse-grained and subjective data distributions, and cannot reflect the actual domain distribution. In this paper, we construct sample-level image-text pairs. Using image-text pairs for contrastive learning, the visual coder
can learn the intrinsic semantics in each image by utilizing the differences between text features in the feature space. In this way, the model jumps out of the limitations of domain labeling and reflects the domain distribution through the differences between samples.

3. METHOD

3.1. Task-oriented text.

As shown in the language teacher module of Fig. 2, we first need to obtain fine-grained descriptions of facial regions as 'teachers' to guide attention modulation. Specifically, we employ MiniGPT-4 [26], which is powered by Vicuna-7B and is resource-efficient to generate descriptions. For example, using MiniGPT-4, in order to obtain rich, instructive knowledge about facial regions, we use the question: 'please describe the [face] in this image in detail.' In subsequent training, we use Clip’s text encoder to obtain features of the language modality, and all its parameters are updated to adapt to the FAS task.

3.2. Attention regulation.

ABSViT [13] has demonstrated its effectiveness in top-down attention regulation. We draw inspiration from the versatility of this method. Specifically, assume that the training set of face images contains M samples and is denoted as $S = \{I_i, T_i\}_{i=0}^{M-1}$. During each training iteration, attention regulation is divided into four steps. (1) The visual coder $V(·)$ performs one forward propagation to encode $I_i$ into the visual feature $x_i^f$. Simultaneously, the text encoder $T(·)$ also propagates one forward to encode $T_i$ into $t_i^f$. (2) We derive the cosine similarity matrix $\text{sim}(·)$ between the textual feature $t_i^f$ and the visual feature $x_i^f$. We truncate the values of the elements in this similarity matrix between 0 and 1. We then use the visual features $x_i^f$ and the cosine similarity matrix $\text{sim}(·)$ to obtain the decoded image features $x_i^d = \text{sim}(x_i^f)$. This step aims to enhance the parts of the visual feature similar to the text and diminish the unrelated parts, thereby obtaining a task-guided visual representation. (3) We feed $x_i^d$ to the top-down decoder $D_t$ to obtain the representation of the middle layer $D_t(x_i^d)$. We save this representation in order, where $l$ represents the serial numbers of the middle decoder and the middle layer. (4) We perform the forward propagation again, unlike the first time, this time $D_t(x_i^d)$ is added to the value matrix $V$ of self-attention layer $\text{Att}(·)$. Where an additional reconstruction loss needs to be optimized for the decoder $D_t$ to better reconstruct the features $D_{t+1}(x_i^d)$ in layer $l+1$ as follows:

$$L_{re} = \frac{1}{T} \sum_{t=1}^{T} \left\| z_t - D_t(z_{t+1}) \right\|^2$$

where the $z$ represents the feature of the second forward propagation, $l$ represents the layer index, and $D$ represents the decoder.

3.3. Sample-level vision text optimization.

In the field of vision language prompt tuning, Zhou et al. [15] have demonstrated that CLIP [14] can improve model generalization by using manually designed sample prompts. Our work further validates this finding. Although FAS is a binary classification task and cannot use category descriptions as text inputs, our content generation module effectively addresses this issue. Specifically, we first construct category-wise texts similar to CLIP, such as $T_i = "A photo of a [class] face.\)", where $i$ represents a specific image sample and $\text{class}$ represents real or fake. We then combine the content-wise texts generated by MiniGPT-4 with these category-wise texts to create sample-level texts with rich information. Finally, through contrastive learning like CLIP, we narrow the gap between visual and textual features in the feature space for each sample, which can be formulated as:

$$L_{con} = - \sum_t \log \frac{e^{\text{sim}(t_i^f,x_i^d)/\tau}}{\sum_{l=0}^{N-1} e^{\text{sim}(t_i^f,x_l^d)/\tau}},$$

where $t_i^f$ is the corresponding facial text description we generate using MiniGPT-4. In order to obtain rich, instructive knowledge about facial regions, we use the question: 'please describe the [face] in this image in detail.' In subsequent training, we use Clip’s text encoder to obtain features of the language modality, and all its parameters are updated to adapt to the FAS task.
where X contains N image samples, \( \tau \) is a learnable temperature parameter, \( t_f \) is the encoded feature of the sample-level text and y is the corresponding label of the image embedding. Therefore, under the optimization of the contrastive loss, we leverage the information from the text to enable the visual encoder to understand the intrinsic semantics of the samples. In this way, the FAS model can overcome manual domain labels’ limitations and generalize the sample-level features to unseen domains.

Finally, we use the cross-entropy loss to supervise the categorization features of ViT, and the overall loss is as follows:

\[
L_{\text{all}} = \lambda_1 L_{\text{re}} + \lambda_2 L_{\text{con}} + \lambda_3 L_{\text{cls}}
\]  

(3)

4. EXPERIMENTS

4.1. Datasets and Metrics

We evaluate on four widely used datasets: Oulu-NPU (O) [27], CASIA (C) [28], Idiap Replay attack (I) [29], and MSU-MFSD (M) [30]. Following prior works, we treat each dataset as one domain and apply the leave-one-out test protocol to evaluate their cross-domain generalization. In addition, we use Half Total Error Rate (HTER) and Area Under Curve (AUC) as evaluation metrics.

4.2. Experimental details

VL-FAS is implemented using PyTorch and trained with the Adam optimizer. We employ ViT-B as our visual network backbone and utilize the text encoder from CLIP to extract text features. The initial learning rate is set to 0.0001, which is reduced by a factor of 0.99 every five epochs. The total training period lasts for 3000 epoch generations. We set the batch size to 12. In particular, we used only the vision transformer for forward propagation during the testing phase.

4.3. Comparisons to the SOTA methods

We perform cross-domain generalization in four commonly used Leave-One-Out (LOO) settings for the FAS task. The comparison methods in Tab. 1 are divided into two parts: State-of-the-art methods and FAS methods utilizing vision transformers as backbones. From the table, we make the following observations. (1) Our VL-FAS achieves superior performance compared to all other methods. Our method exhibits the best performance on protocols of O&C&I to M, O&M&I to C, O&C&M to I, Average metric, and competitive performance on the protocol of I&C&M to O. Although the AUC metric on I&C&M to O protocol is 0.09% lower than the best result, the primary evaluation metric HTER is still the best. These results demonstrate the domain generalization capacity of our method. (2) Our approach showcases performance improvement in vision transformers. Both TransFAS [12] and ViT [11] utilize vision transformer networks pre-trained on the Imagenet dataset, while our method exhibits significant improvements compared to them. Specifically, compared to ViT [11], which is widely reported as a benchmark for generalization experiments, our average HTER is nearly three times better.

4.4. Ablation Study and visualization

As shown in Tab. 2, we demonstrate the performance improvement of each module of the proposed method compared to the baseline method. Re. and Con. represent reconstruction loss in the attention regulation module and contrastive loss in the sample-level visual text optimization module. These ablation experiments demonstrate that our module effectively improves the performance of benchmarks across cross-domain experiments.

![Visualization attention during training process.](image)

We visualize some of the attention during the training process, as shown in Fig. 3, where the second and third rows show the attention of the traditional ViT and the attention of our method, respectively. This figure indicates that the sensory field of the model is focused on the face region during training.

5. CONCLUSION

This paper provides a novel visual linguistic approach (VL-FAS) for generalizable FAS. Specifically, we introduce language modality into FAS to make the model pay more attention to the face region by directing the attention mechanism. Furthermore, we employ training with vision-text pairs at the sample level, which enables the model to comprehend the underlying semantics of the samples and overcome domain limitations. Extensive experiments and analysis on several bench-
mark datasets demonstrate the superiority of our method over state-of-the-art competitors.

6. REFERENCES


