TF-FAS: Twofold-Element Fine-Grained Semantic Guidance for Generalizable Face Anti-Spoofing

Xudong Wang[12*], Ke-Yue Zhang[2*], Taiping Yao[2†], Qianyu Zhou[3],

Shouhong Ding[2], Pingyang Dai[1†], and Rongrong Ji[1]

[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
[2] Youtu Lab, Tencent
[3] Shanghai Jiao Tong University
wangxd@stu.xmu.edu.cn {pydai, rrji}@xmu.edu.cn
{zkyezhang, taipingyao, ericshding}@tencent.com; zhouqianyu@sjtu.edu.cn

Abstract. Generalizable Face anti-spoofing (FAS) approaches have recently garnered considerable attention due to their robustness in unseen scenarios. Some recent methods incorporate vision-language models into FAS, leveraging their impressive pre-trained performance to improve the generalization. However, these methods only utilize coarse-grained or single-element prompts for fine-tuning FAS tasks, without fully exploring the potential of language supervision, leading to unsatisfactory generalization ability. To address these concerns, we propose a novel framework called TF-FAS, which aims to thoroughly explore and harness twofold-element fine-grained semantic guidance to enhance generalization. Specifically, the Content Element Decoupling Module (CEDM) is proposed to comprehensively explore the semantic elements related to content. It is subsequently employed to supervise the decoupling of categorical features from content-related features, thereby enhancing the generalization abilities. Moreover, recognizing the subtle differences within the data of each class in FAS, we present a Fine-Grained Categorical Element Module (FCEM) to explore fine-grained categorical element guidance, then adaptively integrate them to facilitate the distribution modeling for each class. Comprehensive experiments and analysis demonstrate the superiority of our method over state-of-the-art competitors. Code:https://github.com/xudongww/TF-FAS

Keywords: Face anti-spoofing · Vision-language model · Domain generalization

1 Introduction

Face recognition techniques have gained significant traction in diverse applications, such as smartphone login, access control, and electronic payments. Never-

* Equal contributions.
† Corresponding authors.

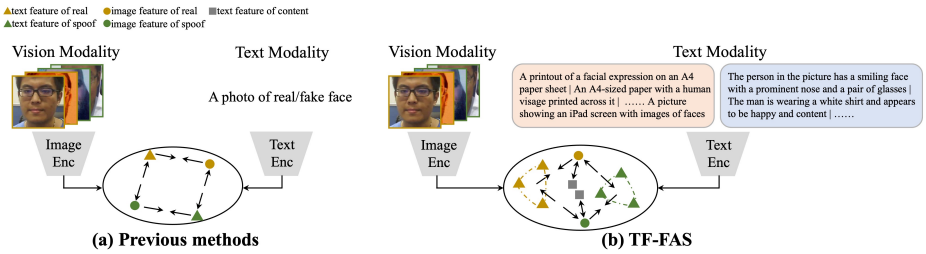


Fig. 1: (a) Previous methods depend exclusively on coarse-grained prompts during the fine-tuning process, failing to fully harness the potential of language supervision. Conversely, (b) TF-FAS is committed to exploring language guidance specifically tailored for FAS, including two elements: content and category, to effectively align the vision and language components, thereby promoting better generalization.

theless, face recognition techniques are constantly confronted with a range of potential threats posed by various presentation attacks, such as printed photos [4], masks [19], and video replays [68]. To mitigate these attacks, researchers propose various face anti-spoofing (FAS) methods that rely on either hand-crafted features [33,34,67,87,97] or deeply-learned features [76,81,90,94,105] for detection. Although existing methods have shown promising performance in intra-dataset scenarios, they encounter difficulties in effectively generalizing to unseen domains due to the inherent domain gap between the source and target distributions.

To address this challenge, domain generalization (DG) [9, 10, 79] methods have been incorporated into FAS tasks to learn content-agnostic discriminative features from multiple source domains allowing for better generalization to unseen domains. Adversarial learning-based methods [30,45,82] and meta-learning-based methods [18, 35, 47] are commonly used in DG. While these techniques indeed enhance the performance of cross-domain FAS, it is important to note that the inherent gap between different domains in the visual modality of FAS tasks still has an impact on the generalization abilities.

As visual-language methods gain prominence, researchers increasingly explore the use of universal language modalities to mitigate the inherent gap present within the visual modality of cross-domain FAS tasks. FLIP [69] employs the coarse-grained heuristic categorical prompts to fine-tune pre-trained models, enabling them to align images with the given prompts and enhance overall performance. VL-FAS [20] employs content-related prompts to provide supervision for models, guiding them to focus on specific facial regions. Supervised by the text modality, these methods indeed achieve remarkable performance. However, their limitation lies in the fact that they solely rely on coarse-grained or single-element prompts during the fine-tuning process, without fully exploring the potential of language supervision. Consequently, this might hurt the generalization abilities, as shown in Fig 1(a).

To cope with this limitation, we propose a novel framework called TF-FAS, which is committed to exploring language guidance specifically tailored for FAS,

as shown in Fig 1(b). Specifically, TF-FAS proposes a twofold-element fine-grained semantic guidance approach to effectively align the vision and language components, thereby promoting better generalization. Firstly, a Content Element Decoupling module (CEDM) is proposed to systematically explore the content-related semantic elements present in each image. These semantic elements are then utilized to guide the model in decoupling content-related information via orthogonalizing the image features and content element features. By doing so, the model places greater emphasis on learning content-agnostic discriminative features from multiple source domains, thereby enhancing cross-domain performance. Secondly, to better describe each class for FAS, we propose a Fine-Grained Categorical Element Module (FCEM). Instead of using coarse-grained text to represent the data in FAS, we leverage numerous granular categorical texts for each subclass of attacks and real in FAS. This approach enables us to effectively capture the diverse forms within each category, thereby improving our ability to model the distribution accurately. Moreover, considering that each granular categorical text contributes differently to the subclass, we propose an adaptive integration strategy. This strategy automatically parameterizes the weight of each text during the training process to provide a comprehensive representation of a specific class.

Our main contributions can be summarized as follows:

- We propose a novel framework named TF-FAS, which introduces a twofold-element fine-grained semantic guidance approach to explore language guidance specifically tailored for FAS tasks.
- TF-FAS introduces two novel modules, the Content Element Decoupling Module (CEDM) and Fine-Grained Categorical Element Module (FCEM). They investigate fine-grained prompts for content and category independently, aiming to acquire content-agnostic discriminative features and effectively capture the diversity within each category. They enhance the generalization abilities of the system as a whole.
- Extensive experiments and analysis demonstrate the superiority of TF-FAS over state-of-the-art competitors on widely-used benchmark datasets.

2 Related Work

2.1 Face Anti-Spoofing

Conventional methods mainly utilize various handcrafted features such as LBP [6, 14, 22], HoG [36, 62, 88], and SIFT [2, 7, 57], to differentiate real and fake faces. However, the performance of these methods is underwhelming due to the shallow structure. With the advent of deep learning, many deep architectures are employed to extract more discriminative features. This evolution included the integration of auxiliary signals like depth maps [63], r-ppg signals [54], or reflection map [89] to enhance detection capabilities. Despite advancements in intra-dataset settings, substantial performance degradation is observed in target

domains due to pronounced domain shifts. FAS techniques employ domain adaptation (DA) [31, 38, 49–51, 55, 76, 98, 107] to mitigate the distribution disparities between source and target domains. However, the acquisition of a sufficient volume of unlabeled target data often poses significant challenges and incurs high costs. Domain generalization (DG) methods [9, 10, 79, 99, 100] have been incorporated into FAS tasks to facilitate the learning of content-agnostic features via adversarial learning [30, 45, 82], meta-learning [18, 35, 47, 106], test-time generalization [104] and instance whitening [105], thereby enhancing generalization to unseen domains. Recently, Vision Transformers (ViT)-based approach [23,28,43] posits that ViT can discern long-range dependencies for superior generalization. However, relying only on image data can limit its generalization capabilities in unseen domains. The emergence of visual-language methods offers new potential to address the aforementioned issues.

2.2 Vision-Language Models

Guided by natural language supervision, vision-language pretraining has recently surfaced as a promising approach for image [29, 39, 59, 80, 93] and video understanding [12, 58, 77, 78, 84]. These approaches diverge from the conventional method of utilizing discrete labels, offering a novel paradigm for recognition based on the alignment of visual and text features. It is inherently suited for zero-shot transfer across various downstream tasks [16, 25, 53, 66, 101]. Several studies have investigated the application of the transferable knowledge from pre-trained models to address tasks such as visual question answering (VQA) [40, 41, 56], zero-shot object detection [53, 66, 85], and image captioning [21, 27, 102], etc. Recent efforts have sought to leverage visual-language methods to bolster the cross-dataset generalization of FAS tasks [20, 52, 69]. These studies posit that text, rich in content-invariant information, can enhance model generalization. However, these methods only utilize coarse-grained or single-element prompts for fine-tuning FAS tasks, without fully exploring the potential of language supervision, leading to unsatisfactory generalization ability. In contrast, we propose a novel TF-FAS framework to thoroughly explore and harness twofold-element fine-grained semantic guidance to enhance generalization.

3 Methodology

In this section, we will provide an elaboration of each component of the proposed method. Section 3.1 outlines the overall architecture of our approach. Section 3.2 revisits the CLIP model for the FAS task. Sections 3.3 and 3.4 detail the implementation specifics of the CEDM and FCEM modules, respectively. Finally, Section 3.5 presents the overall training and optimization strategies.

3.1 Overview

Figure 2 shows the overview of the proposed TF-FAS method, which includes two key components: a Content Element Decoupling Module (CEDM), and a Fine-

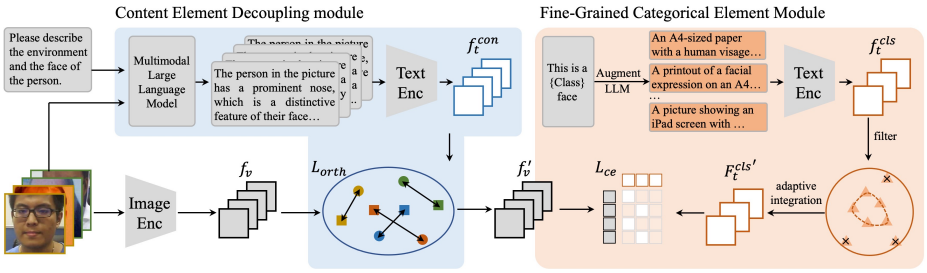


Fig. 2: The framework of TF-FAS explores language guidance specifically tailored for FAS, including two elements: content and category. Specifically, the Content Element Decoupling Module (CEDM) is proposed to systematically explore the content-related semantic elements to guide the model in decoupling content-related features from the essential features of FAS task. Then, the Fine-Grained Categorical Element Module (FCEM) leverages numerous granular categorical texts for each subclass of attacks and real in FAS to model the distribution accurately.

Grained Categorical Element Module (FCEM). Firstly, inspired by the success of multimodal vision-language pre-trained (VLP) in the zero-shot across various downstream tasks, we adopt CLIP as our backbone, where the image encoder is a 12-layer visual transformer ViT-B/16 [3] and the text encoder consists of a 6-layer transformer [71]. The Large Language Model is GPT-4 [1] and the Multimodal Large Language Model is LLAVA [44]. During the training, to prevent the disruption of the joint text-image space, we froze the text encoder and trained only the image encoder during the generalization process. Initially, the given image is fed into the visual encoder to obtain the corresponding visual representation.

First, CEDM is proposed to systematically explore the content-related semantic elements present in each image, which guides the model in decoupling content-related information via orthogonalizing the image features and content element features. As such, the model places greater emphasis on learning content-agnostic discriminative features from multiple source domains, thereby enhancing the generalizability. As for FCEM, instead of using coarse-grained text to represent the data in FAS, we leverage numerous granular categorical texts to represent each subclass of attacks and real faces, which effectively capture the diverse forms within each category, thus improving the ability to model the distribution accurately. Moreover, considering that each granular categorical text contributes differently to the subclass, we propose an adaptive integration strategy, which automatically parameterizes the weight of each text during the training process, achieving efficient inference.

3.2 Revisiting CLIP Baseline for FAS

Contrastive Language-Image Pre-Training (CLIP) [60] extracts dual-modality features by a visual encoder and text encoder that is trained on the WebIm-

ageText, consisting of 400 million image-text pairs collected from a variety of publicly available sources on the Web. The visual encoder VisEnc() extracts features and projects them to a global feature $V \in \mathbb{R}^D$. The text encoder TextEnc() generates a global text representation $T \in \mathbb{R}^{D \times K}$ for K categories. Classification scores $S \in \mathbb{R}^K$ are computed as:

$$T = \text{TextEnc}\left(P_K\right); \quad V = \text{VisEnc}(I); \quad S = T^{\text{T}}V \tag{1}$$

where T and V are $L2$ normalized features, and P_K are the category prompts to describe the K categories. The matrix multiplication computes cosine similarity. For the FAS task, the number of classes is 2, corresponding to real and spoof, with the respective prompts being *The photo of a real face* and *The photo of a spoof face*. The training objective of CLIP is to maximize the cosine similarity $\text{sim}(\cdot, \cdot)$ of the paired image and text embedding while minimizing the cosine similarity of the unpaired ones. Hence, we employ cross-entropy loss to bring matching pairs closer and separate non-matching pairs in feature space, which is defined as:

$$\begin{aligned} &L_{ce}(x, y) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(sim(V_{x_i}, T_{y_i})), \\ &with \quad \text{sim}(V_{x_i}, T_{y_i}) = V_{x_i}^{\text{T}} T_{y_i} / \|V_{x_i}\| \|T_{y_i}\|. \end{aligned} \tag{2}$$

where N is the batch size, and V_{x_i} and T_{y_i} represent visual features of the image and text features of its corresponding category prompt, respectively. During the inference stage, the predicted class $\hat{y}$ is determined by the class description having the highest cosine similarity score $\text{sim}(\cdot, \cdot)$ with the given image I.

3.3 Content Element Decoupling Module

The pre-trained CLIP model extracts not only detailed contour information but also captures semantic content features from images, including environmental context (background, lighting) and facial attributes (age, gender, expressions), which are not directly relevant to the FAS tasks. Taking inspiration from previous research [86, 94], decoupling task-irrelevant features has the potential to enhance the overall ability to generalize. To this end, we propose CEDM, which comprehensively explores the semantic elements related to content and is employed to supervise the decoupling of categorical features from content-related features, thereby enhancing generalization abilities.

Content Prompts Generation. Recently, we have observed the remarkable success of Multimodal Language Models (MLLMs), which can comprehend the relationship between visual content and linguistic descriptions, generating descriptive text by combining text prompts with corresponding images. Inspired by this, we utilize MLLMs to generate content prompts for images in the FAS dataset. To ensure the text focuses on content irrelevant to the FAS task, such as environmental factors (lighting and background) and facial attributes (gender, age, and expression), we have designed the prompts as follows:
Please describe the environment (background, lighting intensity, etc.) and the face (age, gender, expression, etc.) of the person. Additionally, to prevent the

content prompts from reflecting real deceptive characteristics and to avoid conflicts between the decoupling objective and the task objective, we remove the sentences containing category-related keywords (such as "print" and "display") from the generated content prompts, ensuring that useful features are retained.

Content-related Decoupling. Inspired by several methods to disentangle visual features from the textual information, such as contrastive learning, mutual information minimization [5,13], or orthogonal loss [61], we utilize the generated content prompts P_i^{con}, to supervise the decoupling of visual features extracted. Specifically, we feed images I_i to an image encoder to obtain visual feature f_{vi}, and input the corresponding content prompts P_i^{con} to a text encoder to obtain content feature f_{ti}^{con}. Afterward, we calculate the dot product between the two sets of features and utilize the resulting value as a loss function to constrain the model. The objective is to ensure that the extracted visual features are independent of the content features. The loss of orthogonality L_{orth} is defined as:

$$L_{\text{orth}} = \sum_{i=1}^{N} f_{vi} \cdot f_{ti}^{con}, \tag{3}$$

where $f_{vi} = \text{VisEnc}(I_i), f_{ti}^{con} = \text{TextEnc}(P_i^{con})$.

3.4 Fine-Grained Categorical Element Module

In the real world, the types of attacks in liveness detection are diverse, such as print, video replay, *etc.*. Furthermore, within the same category of attack, there are multiple forms, such as A4 paper and kraft paper in print attacks. Simple and singular binary classification task prompts are insufficient to model the full spectrum of attack types. Therefore, we propose the FCEM to explore fine-grained categorical element guidance, and then adaptively integrate them to facilitate the distribution modeling for each class.

Fine-grained Categorical Prompt Generation. In the FAS task, attack types include print, replay, and 3D attacks, *etc.* Print attacks focus on high-level visual semantics, while replay attacks involve low-level textural features, such as moiré patterns. It is unreasonable to conflate all the attacks. This paper utilizes the fine-grained task prompts to subdivide and guide the attack types in the classification. These detailed prompts provide the model with more deceptive information, enabling it to learn the subtle differences between different attack types and thereby enhancing its ability to discriminate between them. Additionally, diversifying category prompts enhances the robustness of the content decoupling process. Specifically, motivated by [17], we propose employing GPT-4 [1] to automatically generate a diverse set of prompts, rather than manually designing prompts, since it might introduce subjective bias and is cumbersome. Concretely, we query GPT with requirement prompts such as:

Paraphrase the sentence: {The photo of a {class} face.} with similar semantics.

For each class in FAS tasks, we generate 64 categorical prompts, which are displayed in the supplementary materials.

Categorical Prompt Filtering. The inherent randomness in the generation of 64 prompts for each class may inevitably introduce some noise prompts, which is not appropriate for the FAS datasets. Using them directly could potentially hinder performance. To address this issue, we propose a filtering mechanism to automatically select the most effective prompts for the tasks. Specifically, we evaluate the compatibility of generated prompts P_j^{cls} with the dataset distribution based on the similarity between the textual features of the prompts f_{ti}^{cls} and the image features f_{vi}. We then filter out generated prompts that deviate significantly from the dataset distribution using a threshold θ. The filtering is formulated as follows:

$$\text{Filtered Prompts} = \left\{ P_j^{cls} \mid (\sum_{i=1}^{N} \text{sim}(P_j^{cls}, I_i))/N \geq \theta \right\} \tag{4}$$

where P_j^{cls} is jth prompt, I_i is the ith image of corresponding class, sim(P_j^{cls}, I_i) is the function that calculates the similarity of the features between the prompt P_j^{cls} and the image I_i, and N is the total number of this class in the dataset.

Adaptive Prompt Integration. To comprehensively represent one specific class, individual prompt alone is insufficient. It is necessary to integrate all the prompts belonging to the same category in order to obtain the final features. A intuitive approach would be to average the features of prompts belonging to the same type, representing the corresponding class. However, we observe that each prompt fits the dataset to varying degrees, and thus, using a simple average for aggregation is unreasonable. Therefore, we set the prompt weights as learnable parameters, adaptively adjusting the weight coefficients during training to better model the overall data distribution. Specifically, the integration process is formulated as:

$$F_t^{cls} = \sum_{i=1}^{N} \sigma(w_i) \cdot f_{ti}^{cls}, \quad \sigma(w_i) = \frac{e^{w_i}}{\sum_{j=1}^{K} e^{w_j}} \tag{5}$$

where $f_{ti}^{cls} = \text{TextEnc}(P_i^{cls})$ is the feature of corresponding category prompts P_i^{cls}, w_i is the learnable weight with the summarization of all $\sigma(w_i)$ equals 1.

3.5 Overall Training and Optimization

To further enhance the model's robustness against data variations, we fellow FLIP [69] to employ a simCLR loss for auxiliary training. This approach generates two views (I_{v_1} and I_{v_2}) of a given image I through distinct transformations. The features of the two transformed images are extracted by the image encoder VisEnc() and subsequently projected via a non-linear projection network $\mathcal{H}$. A contrastive loss is then applied to the projected features. $\boldsymbol{f}_{v_1} = \text{VisEnc}(I_{v_1}), \boldsymbol{f}_{v_2} = \text{VisEnc}(I_{v_2})$. $\boldsymbol{h}_1 = \mathcal{H}(\boldsymbol{f}_{v_1}), h_2 = \mathcal{H}(\boldsymbol{f}_{v_2}), \boldsymbol{h}_1, \boldsymbol{h}_2 \in \mathbb{R}^{d_h}$.

$$L_{simCLR} = \texttt{simCLR}(\boldsymbol{h}_1, \boldsymbol{h}_2)$$

Overall, we formulate the joint optimization objective as:

$$L = L_{ce} + \lambda_1 L_{orth} + \lambda_2 L_{simCLR} \tag{6}$$

where λ_1 and λ_2 is hyper-parameters.

4 Experiment

4.1 Experimental Setting

Datasets and DG Protocols: Our evaluation encompasses two protocols. Strictly following the [28], we adopt a leave-one-domain-out approach for the two protocols, treating each dataset as a distinct domain to gauge cross-domain capabilities on the remaining domain. **Protocol 1** tests our method on established cross-domain FAS benchmarks: MSU-MFSD **(M)** [83], CASIA-MFSD **(C)** [96], Idiap Replay Attack **(I)** [15], and OULU-NPU **(O)** [8], with scenarios like **OCI** $\rightarrow$ **M** indicating **O**, **C**, and **I** as sources and **M** as the target. **Protocol 2**, strictly following [92], is a single-source-to-single-target setup using **M**, **C**, **I**, and **O** datasets, yielding 12 scenarios. To fairly compare with FLIP [69], we also conduct the above experiments with the auxiliary dataset the CelebA-Spoof [95]. In addition, to better simulate the real-world scenarios without large pre-trained datasets, we also conduct the experiments without CelebA-Spoof.
Implementation Details: The image encoder and the text encoder are the dual-stream CLIP where the image encoder adopts the ViT-B/16 structure. The LLMs used for expanding category prompts utilizes GPT-4 [1] and the MLLMs is LLAVA [44] with 13 billion parameters. Face images are preprocessed to a resolution of $224 \times 224 \times 3$ and segmented into patches measuring 16×16. The maximum length of the textual token sequence L is set to 77. Our method is implemented with PyTorch and trained with Adam optimizer, with both the learning rate and weight decay initialized at 10^{-6}. During training, batch sizes are set to 3. For testing, the batch size is set to 10 across all protocols. Each variant of our model undergoes training for a total of 4000 iterations. λ_1 and λ_2 are set to 1. The text encoder is frozen and only the image encoder and the parameters of the category prompt are trained.
Evaluation Metrics: Following [28], we use two metrics: Half Total Error Rate (HTER) and Area Under the Receiver Operating Characteristic Curve (AUC). HTER, the average of False Acceptance and False Rejection Rates, indicates error balance, with lower values being better. AUC measures discrimination ability, with values closer to 1 being superior and 0.5 indicating random chance. These metrics together offer a comprehensive evaluation of the model's performance.

4.2 Cross-domain FAS Performance

The MCIO dataset, being smaller compared to CelebA-Spoof [95], benefits significantly from the addition of it in bridging the domain gap between different domains. To comprehensively investigate the impact of the proposed method

Table 1: Evaluation of cross-domain performance in Protocol 1, between MSU-MFSD (**M**), CASIA-MFSD (**C**), Replay Attack (**I**) and OULU-NPU (O) with the assessment metrics being HTER and AUC. The * indicates using the CelebA-Spoof [83] as the supplementary source dataset.

Method	OCI → M		OMI → C		OCM → I		ICM → O		Avg.
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER
MADDG (CVPR' 19) [64]	17.69	88.06	24.50	84.51	22.19	84.99	27.98	80.02	23.09
MDDR (CVPR' 20) [74]	17.02	90.10	19.68	87.43	20.87	86.72	25.02	81.47	20.64
NAS-FAS (TPAMI' 20) [91]	16.85	90.42	15.21	92.64	11.63	96.98	13.16	94.18	14.21
RFMeta (AAAI' 20) [65]	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16	16.97
D^2AM (AAAI' 21) [11]	12.70	95.66	20.98	85.58	15.43	91.22	15.27	90.87	16.09
DRDG (IJCAI' 21) [48]	12.43	95.81	19.05	88.79	15.56	91.79	15.63	91.75	15.66
Self-DA (AAAI' 21) [76]	15.40	91.80	24.50	84.40	15.60	90.10	23.10	84.30	19.65
ANRL (ACM MM' 21) [47]	10.83	96.75	17.85	89.26	16.03	91.04	15.67	91.90	15.09
FGHV (AAAI' 21) [46]	9.17	96.92	12.47	93.47	16.29	90.11	13.58	93.55	12.87
SSDG-R (CVPR' 20) [30]	7.38	97.17	10.44	95.94	11.71	96.59	15.61	91.54	11.28
SSAN-R (CVPR' 22) [82]	6.67	98.75	10.00	96.67	8.88	96.79	13.72	93.63	9.80
PatchNet (CVPR' 22) [72]	7.10	98.46	11.33	94.58	13.40	95.67	11.82	95.07	10.90
GDA (ECCV' 22) [107]	9.20	98.00	12.20	93.00	10.00	96.00	14.40	92.60	11.45
AMEL (ACM MM' 22) [106]	10.23	96.62	11.88	94.39	18.60	88.79	11.31	93.96	13.00
IADG (CVPR' 23) [105]	5.41	98.19	8.70	96.44	10.62	94.50	8.86	97.14	8.40
GAC-FAS (CVPR' 24) [37]	5.00	97.56	8.20	95.16	4.29	98.87	8.60	97.16	6.52
DiVT-M (WACV' 23) [42]	**2.86**	99.14	8.67	96.62	3.71	99.29	13.06	94.04	7.07
VL-FAS (ICASSP' 24) [20]	3.13	99.31	4.00	98.64	5.00	98.90	7.92	97.05	5.01
TF-FAS (Ours)	3.44	**99.42**	**0.81**	**99.92**	**2.24**	**99.67**	**2.26**	**99.48**	**2.19**
ViT* (ECCV' 22) [28]	1.58	99.68	5.70	98.91	9.25	97.15	7.47	98.42	6.00
FLIP-MCL* (ICCV' 23) [69]	4.95	98.11	**0.54**	99.98	4.25	99.07	2.31	99.63	3.01
TF-FAS* (Ours)	**1.49**	**99.80**	0.58	**99.99**	**1.56**	**99.89**	**1.43**	**99.93**	**1.27**

Table 2: Evaluation of cross-domain performance in Protocol 2, for all the 12 different combinations between MSU-MFSD (**M**), CASIA-MFSD (**C**), Replay Attack (**I**) and OULU-NPU (O) with the assessment metrics being HTER. The * indicates using the CelebA-Spoof [83] as the supplementary source dataset.

Method	C → I	C → M	C → O	I → C	I → M	I → O	M → C	M → I	M → O	O → C	O → I	O → M	Avg.
ADDA (CVPR' 17) [70]	41.8	36.6	-	49.8	35.1	-	39.0	35.2	-	-	-	-	39.6
DRCN (ECCV' 16) [24]	44.4	27.6	-	48.9	42.0	-	28.9	36.8	-	-	-	-	38.1
DupGAN (CVPR' 18) [26]	42.4	33.4	-	46.5	36.2	-	27.1	35.4	-	-	-	-	36.8
KSA (TIFS' 18) [38]	39.3	15.1	-	12.3	33.3	-	9.1	34.9	-	-	-	-	24.0
DR-UDA (TIFS' 20) [75]	15.6	9.0	28.7	34.2	29.0	38.5	16.8	3.0	30.2	19.5	25.4	27.4	23.1
MDDR (CVPR' 20) [74]	26.1	20.2	24.7	39.2	23.2	33.6	34.3	8.7	31.7	21.8	27.6	22.0	26.1
ADA (ICB' 19) [73]	17.5	9.3	29.1	41.5	30.5	39.6	17.7	5.1	31.2	19.8	26.8	31.5	25.0
USDAN-Un (PR' 21) [31]	16.0	9.2	-	30.2	25.8	-	13.3	3.4	-	-	-	-	16.3
GDA (ECCV' 22) [107]	15.10	5.8	-	29.7	20.8	-	12.2	2.5	-	-	-	-	14.4
CDFTN-L (AAAI' 23) [92]	**1.7**	8.1	29.9	11.9	9.6	29.9	8.8	**1.3**	25.6	19.1	5.8	6.3	13.2
TF-FAS	10.82	**3.44**	**4.16**	**1.51**	**3.19**	**4.50**	**0.69**	3.88	**3.53**	**1.28**	**5.68**	**2.09**	**3.73**
FLIP-MCL* (ICCV' 23) [69]	10.57	7.15	3.91	**0.68**	7.22	4.22	0.19	5.88	3.95	**0.19**	5.69	8.40	4.84
TF-FAS*	**3.06**	**1.59**	**3.78**	0.69	**1.34**	**2.50**	**0.11**	**2.31**	**1.40**	1.5	**4.02**	**1.59**	**1.99**

on domain generalization, all protocols were conducted both with and without CelebA-Spoof [95]. Tables 1 and 2 detail the zero-shot cross-domain performance under **Protocols 1** and **2**, respectively. The results and analyses are as follows.

Table 3: Ablation studies on each proposed component: CPDM and FCEM

Baseline	FCEM	CPDM	C → I		C → M		C → O		Avg.
			HTER	AUC	HTER	AUC	HTER	AUC	HTER
✓			20.75	86.77	11.68	95.61	19.24	89.38	17.22
✓	✓		14.61	91.83	5.98	97.43	7.59	97.49	9.39
✓	✓	✓	**10.82**	**95.54**	**3.44**	**98.85**	**4.16**	**99.07**	**6.14**

Cross-domain performance in Protocol 1. The proposed framework attained optimal performance, compared to the state-of-the-art (SOTA) methods, in three-quarters of the settings (C=+3.19, I=+1.47, O=+5.66), with an average performance increase of +2.82 without CelebA-Spoof [95] even surpassing the FLIP [69] train with CelebA-Spoof [95] 1.74 (in terms of average HTER). Likewise, with the inclusion of CelebA-Spoof [95], optimal performance was achieved in three-quarters of the settings (M=+0.09, I=+2.69, O=+0.88), yielding an average enhancement of 1.74. This demonstrates the effectiveness of the proposed TF-FAS in modeling live data and effectively bridging the domain gap.
Cross-domain performance in Protocol 2. In single-source to single-target settings, the proposed TF-FAS surpasses current SOTA methods by a considerable margin of +9.47 and +2.85 in terms of average HTER without and with the inclusion of CelebA-Spoof [95], respectively. Specifically, for the target domain O, there are substantial improvements of +25.74, +25.4, and +22.07 when selecting C, I, and O as the source domains, respectively, without CelebA-Spoof [95]. When including CelebA-Spoof [95], in comparison to FLIP-MCL [69], the proposed method achieves a maximum increase of +7.51 (C → I). These findings confirm that TF-FAS is capable of learning robust generalizable features and adept at navigating challenges posed by limited data and domain gaps.

4.3 Ablation Studies

Due to the significant domain gap between dataset C and other datasets, transferring knowledge learned from source domain C to other domains results in a considerable performance drop. Furthermore, incorporating CelebA-Spoof [95] as supplementary data for the source domain helps to bridge the gap between the source and target domains. Therefore, to convincingly demonstrate the feasibility of the proposed method for domain generalization, all ablation experiments are conducted in the settings of C→I, C→M, and C→O without using CelebA-Spoof [95] as additional source domain data.
Effect of CEDM and FCEM. To explore the impact of each proposed module on the generalization of FAS, we conducted ablation experiments on the proposed modules, using a dual-stream CLIP structure with category prompts: *The photo of the real face* and *The photo of the spoof face* as the baseline. As shown in Table 3, the inclusion of the FCEM module resulted in a 7.83% improvement in the average HTER, indicating that the fine-grained FCEM can better model attack

Table 4: Effect of category prompt generation and integration.

Coarse	Fine-grain	Coop	Augment	Filter	Adaptive	C → I HTER	C → I AUC	C → M HTER	C → M AUC	C → O HTER	C → O AUC	Avg. HTER
✓						20.75	86.77	11.68	95.61	19.24	89.38	17.22
	✓					15.37	91.67	10.84	95.25	15.41	92.76	13.87
	✓	✓				17.54	89.77	13.52	93.92	12.31	95.09	14.56
	✓		✓			16.26	89.5	8.23	99.74	10.17	96.76	11.55
	✓		✓	✓		15.37	90.08	7.94	96.29	9.15	96.78	10.82
	✓		✓	✓	✓	**14.61**	**91.83**	**5.98**	**97.43**	**7.59**	**97.49**	**9.39**

Table 5: Effect of different disentangling functions.

Prompt	C → I HTER	C → I AUC	C → M HTER	C → M AUC	C → O HTER	C → O AUC	Avg. HTER
none	14.61	91.83	5.98	97.43	7.59	97.49	9.39
club	13.1	93.16	3.76	98.65	6.8	97.91	7.89
mine	11.57	94.41	3.44	97.66	7.72	97.96	7.58
contrast	12.31	93.34	5.29	97.8	5.41	98.20	7.67
orthogonality	**10.82**	**95.54**	**3.44**	**98.85**	**4.16**	**99.07**	**6.14**

types, discern subtle differences between different attack types, and enhance spoof detection capabilities. The addition of the CEDM module led to a 3.25% increase in average HTER, suggesting that content semantics related to the domain do not facilitate domain generalization, and decoupling such content from the features can promote cross-domain generalization.

Effect of category prompt generation and aggregation: To thoroughly investigate the impact of each component within the FCEM, we conducted comprehensive ablation experiments. The results, as shown in Table 4, reveal that coarse-grained category prompts struggle to model the diverse types of live attacks, resulting in a suboptimal performance with an average HTER of only 17.22%. The introduction of fine-grained prompts enhances the model's ability to discern subtle differences between various attack types, improving the average HTER by 3.38%. Furthermore, we observed that fine-grained prompts learned through the CoOp [103] approach did not outperform manually designed prompts, suggesting that prompt tuning may lead to overfitting on the source dataset, which is detrimental to domain generalization, as evidenced by a 0.69% decrease in average HTER. The expansion of category prompts via GPT-4 [1] enables more comprehensive modeling of sample categories, thereby improving generalization performance and increasing the average HTER by 2.32%. The filtering operation helps to eliminate augmentation samples that deviate from the dataset, enhancing the fit of category samples and increasing the average HTER by 0.73%. Finally, adaptive weight parameter learning, which adaptively balances each sample, further promotes the process of category prompt modeling across all samples, leading to a 1.43% increase in the average HTER.

Effect of different disentangling loss: Table 5 presents the effects of different disentanglement functions. We found that for the FAS task, all tested

Table 6: Effect of different content semantics.

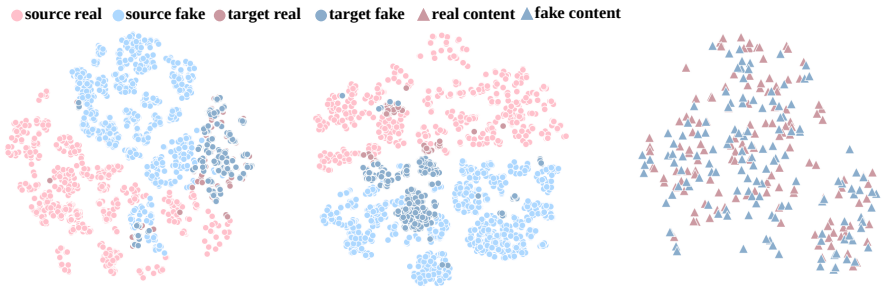
Prompt	C → I		C → M		C → O		Avg.
	HTER	AUC	HTER	AUC	HTER	AUC	HTER
none	14.61	91.83	5.98	97.43	7.59	97.49	9.39
environment	12.24	94.85	5.29	97.53	5.6	98.33	7.71
face semantics	**10.07**	94.56	3.44	97.91	5.98	98.06	6.50
all	10.82	**95.54**	**3.44**	**98.85**	**4.16**	**99.07**	**6.14**

disentanglement functions have a certain degree of efficacy, with the orthogonal function [61] yielding the best disentanglement performance increasing the average HTER by 3.25%. The club [13] and mine [5] methods, which estimate mutual information through neural networks, aim to minimize the mutual information between the model-extracted visual features and content features. However, these complex mutual information estimation methods are not the most effective for feature disentanglement in FAS tasks. In contrastive learning [32], we treated category prompts and content prompts as positive and negative for images, respectively, to promote independence between image features and content features. However, in this function, the attraction of positive samples plays a dominant role, and the optimization process for the independence of image features and content features is not complete, resulting in unsatisfactory disentanglement performance. The more direct orthogonal method, which enforces orthogonality between image features and content features in high-dimensional space, effectively renders the extracted visual features content-agnostic and enhances cross-dataset generalization capabilities.

Effect of different content prompt: Table 6 illustrates the impact of disentangling environmental information (lighting and background) and facial semantics (age, gender, expression, and facial features) on performance. Disentangling environmental information increased the average HTER by 1.68%, while disentangling facial semantics led to a more significant improvement of 2.89%. Disentangling both aspects further enhanced generalization capabilities, yielding a 3.59% improvement in average HTER. These results suggest that both environmental and facial attributes are detrimental to the FAS task and hinder domain transfer. The greater impact of disentangling facial semantics is likely due to the face extraction pre-processing step, which emphasizes facial regions in the images, making facial attributes more influential than environmental factors.

4.4 Visualization and Analysis

T-SNE visualization of image feature distributions. To understand how TF-FAS models live data and learns common knowledge across different datasets, we used t-SNE to visualize feature distributions for each domain. Fig. 3 (a-b) illustrates these visualizations. Compared to the FLIP, our method shows clear segmentation boundaries on the source dataset, highlighting the effectiveness of



(a) FLIP (b) TF-FAS (c) Content

Fig. 3: Comparison results of t-SNE feature visualization.

FCEM in modeling live data. Additionally, on the target dataset, which was not used during training, our method also establishes clear decision boundaries, with similar distributions between the source and target datasets. This indicates that the CEDM effectively decouples features, enabling the model to learn commonalities across datasets and enhancing cross-dataset generalization performance.

T-SNE visualization of content feature distributions. To understand the role of the content prompt in the FAS task, we used t-SNE to reduce the dimensionality of the content feature space for visualization. As shown in Figure 3(c), the reduced content features are chaotically scattered, with both real and fake images uniformly dispersed. This suggests that image semantics are irrelevant to the FAS task, while environmental information and facial semantics are content-related and affect generalization. Therefore, decoupling content semantics during domain transfer is crucial. The visualization supports this rationale.

5 Conclusion

In this paper, we propose a novel framework of TF-FAS, which introduces a twofold-element fine-grained semantic guidance method to explore language guidance designed for FAS tasks. Concretely, we propose the Content Element Decoupling Module (CEDM) to conduct a thorough investigation of semantic elements associated with content. It plays a pivotal role in supervising the disentanglement of categorical features from those related to semantics, consequently fortifying the generalization capabilities of the model. Furthermore, acknowledging the intricate variations within the data of each class in FAS, we have devised the Fine-Grained Categorical Element Module (FCEM), which is tailored to scrutinize and harness fine-grained categorical element guidance, and adeptly integrates these insights to refine the modeling of distributions for each class, thereby capturing the subtle distinctions that are critical for effective anti-spoofing. Extensive experimental evaluations and in-depth analyses have been conducted, which collectively attest to the preeminence of our TF-FAS framework over the current state-of-the-art competitors.

Acknowledgements

This work was supported by National Science and Technology Major Project (No. 2022ZD0118201), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of Chine a (No. U21B2037, No. U22B2051, No. U23A20383, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No.2022J06001).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Agarwal, A., Singh, R., Vatsa, M.: Face anti-spoofing using haralick features. In: International Conference on Biometrics: Theory, Applications and Systems (BTAS). pp. 1–6 (2016)
3. Alexey, D.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929 (2020)
4. Anjos, A., Marcel, S.: Counter-measures to photo attacks in face recognition: a public database and a baseline. In: International Joint Conference on Biometrics (IJCB). pp. 1–7 (2011)
5. Belghazi, M.I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, R.D.: Mine: mutual information neural estimation. arXiv preprint arXiv:1801.04062 (2018)
6. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face anti-spoofing based on color texture analysis. In: IEEE international conference on image processing (ICIP). pp. 2636–2640 (2015)
7. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face antispoofing using speeded-up robust features and fisher vector encoding. IEEE Signal Processing Letters (SPL) pp. 141–145 (2016)
8. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: Oulu-npu: A mobile face presentation attack database with real-world variations. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG). pp. 612–618 (2017)
9. Cai, R., Cui, Y., Li, Z., Yu, Z., Li, H., Hu, Y., Kot, A.: Rehearsal-free domain continual face anti-spoofing: Generalize more and forget less. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8037–8048 (2023)
10. Cai, R., Yu, Z., Kong, C., Li, H., Chen, C., Hu, Y., Kot, A.C.: S-adapter: Generalizing vision transformer for face anti-spoofing with statistical tokens. IEEE Transactions on Information Forensics and Security (TIFS) (2024)
11. Chen, Z., Yao, T., Sheng, K., Ding, S., Tai, Y., Li, J., Huang, F., Jin, X.: Generalizable representation learning for mixture domain face anti-spoofing. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 1132–1139 (2021)

12. Cheng, F., Wang, X., Lei, J., Crandall, D., Bansal, M., Bertasius, G.: Vindlu: A recipe for effective video-and-language pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10739–10750 (2023)
13. Cheng, P., Hao, W., Dai, S., Liu, J., Gan, Z., Carin, L.: Club: A contrastive log-ratio upper bound of mutual information. In: International Conference on Machine Learning (ICML). pp. 1779–1788 (2020)
14. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG). pp. 1–7 (2012)
15. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG). pp. 1–7 (2012)
16. Cui, Q., Zhou, B., Guo, Y., Yin, W., Wu, H., Yoshie, O., Chen, Y.: Contrastive vision-language pre-training with limited resources. In: European Conference on Computer Vision (ECCV). pp. 236–253 (2022)
17. Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., et al.: Auggpt: Leveraging chatgpt for text data augmentation. arXiv preprint arXiv:2302.13007 (2023)
18. Du, Z., Li, J., Zuo, L., Zhu, L., Lu, K.: Energy-based domain generalization for face anti-spoofing. In: ACM International Conference on Multimedia (ACM MM). pp. 1749–1757 (2022)
19. Erdogmus, N., Marcel, S.: Spoofing 2d face recognition systems with 3d masks. In: Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG). pp. 1–8 (2013)
20. Fang, H., Liu, A., Jiang, N., Lu, Q., Zhao, G., Wan, J.: Vl-fas: Domain generalization via vision-language model for face anti-spoofing. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4770–4774 (2024)
21. Fei, J., Wang, T., Zhang, J., He, Z., Wang, C., Zheng, F.: Transferable decoding with visual entities for zero-shot image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3136–3146 (2023)
22. de Freitas Pereira, T., Anjos, A., De Martino, J.M., Marcel, S.: Lbp- top based countermeasure against face spoofing attacks. In: Computer Vision-ACCV 2012 Workshops: ACCV 2012 International Workshops(ACCV). pp. 121–132 (2013)
23. George, A., Marcel, S.: On the effectiveness of vision transformers for zero-shot face anti-spoofing. In: International Joint Conference on Biometrics (IJCB). pp. 1–8 (2021)
24. Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W.: Deep reconstruction-classification networks for unsupervised domain adaptation. In: European Conference on Computer Vision (ECCV). pp. 597–613 (2016)
25. Goyal, S., Kumar, A., Garg, S., Kolter, Z., Raghunathan, A.: Finetune like you pretrain: Improved finetuning of zero-shot vision models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19338–19347 (2023)
26. Hu, L., Kan, M., Shan, S., Chen, X.: Duplex generative adversarial network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1498–1507 (2018)

27. Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., Wang, L.: Scaling up vision-language pre-training for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17980–17989 (2022)
28. Huang, H.P., Sun, D., Liu, Y., Chu, W.S., Xiao, T., Yuan, J., Adam, H., Yang, M.H.: Adaptive transformers for robust few-shot cross-domain face anti-spoofing. In: European Conference on Computer Vision (ECCV). pp. 37–54 (2022)
29. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning (ICML). pp. 4904–4916 (2021)
30. Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8484–8493 (2020)
31. Jia, Y., Zhang, J., Shan, S., Chen, X.: Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing. Pattern Recognition (PR) **115**, 107888 (2021)
32. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems (NeurIPS) pp. 18661–18673 (2020)
33. Kim, G., Eum, S., Suhr, J.K., Kim, D.I., Park, K.R., Kim, J.: Face liveness detection based on texture and frequency analyses. In: International Conference on Biometrics (ICB). pp. 67–72 (2012)
34. Kim, S., Yu, S., Kim, K., Ban, Y., Lee, S.: Face liveness detection using variable focusing. In: International Conference on Biometrics (ICB). pp. 1–6 (2013)
35. Kim, Y.E., Lee, S.W.: Domain generalization with pseudo-domain label for face anti-spoofing. In: Asian Conference on Pattern Recognition (ACPR). pp. 431–442 (2021)
36. Komulainen, J., Hadid, A., Pietikäinen, M.: Context based face anti-spoofing. In: International Conference on Biometrics: Theory, Applications and Systems (BTAS). pp. 1–8 (2013)
37. Le, B.M., Woo, S.S.: Gradient alignment for cross-domain face anti-spoofing. arXiv preprint arXiv:2402.18817 (2024)
38. Li, H., Li, W., Cao, H., Wang, S., Huang, F., Kot, A.C.: Unsupervised domain adaptation for face anti-spoofing. IEEE Transactions on Information Forensics and Security (TIFS) pp. 1794–1809 (2018)
39. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (ICML). pp. 12888–12900 (2022)
40. Li, P., Liu, G., He, J., Zhao, Z., Zhong, S.: Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 374–383 (2023)
41. Li, P., Liu, G., Tan, L., Liao, J., Zhong, S.: Self-supervised vision-language pre-training for medial visual question answering. In: International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (2023)
42. Liao, C.H., Chen, W.C., Liu, H.T., Yeh, Y.R., Hu, M.C., Chen, C.S.: Domain invariant vision transformer learning for face anti-spoofing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6098–6107 (2023)

43. Liu, A., Tan, Z., Yu, Z., Zhao, C., Wan, J., Lei, Y.L.Z., Zhang, D., Li, S.Z., Guo, G.: Fm-vit: Flexible modal vision transformers for face anti-spoofing. IEEE Transactions on Information Forensics and Security (TIFS) (2023)
44. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in Neural Information Processing Systems (NeurIPS) (2024)
45. Liu, M., Mu, J., Yu, Z., Ruan, K., Shu, B., Yang, J.: Adversarial learning and decomposition-based domain generalization for face anti-spoofing. Pattern Recognition Letters (PRL) pp. 171–177 (2022)
46. Liu, S., Lu, S., Xu, H., Yang, J., Ding, S., Ma, L.: Feature generation and hypothesis verification for reliable face anti-spoofing. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 1782–1791 (2022)
47. Liu, S., Zhang, K.Y., Yao, T., Bi, M., Ding, S., Li, J., Huang, F., Ma, L.: Adaptive normalized representation learning for generalizable face anti-spoofing. In: ACM International Conference on Multimedia (ACM MM). pp. 1469–1477 (2021)
48. Liu, S., Zhang, K.Y., Yao, T., Sheng, K., Ding, S., Tai, Y., Li, J., Xie, Y., Ma, L.: Dual reweighting domain generalization for face presentation attack detection. arXiv preprint arXiv:2106.16128 (2021)
49. Liu, Y., Chen, Y., Dai, W., Gou, M., Huang, C.T., Xiong, H.: Source-free domain adaptation with contrastive domain alignment and self-supervised exploration for face anti-spoofing. In: European Conference on Computer Vision (ECCV) (2022)
50. Liu, Y., Chen, Y., Dai, W., Gou, M., Huang, C.T., Xiong, H.: Source-free domain adaptation with domain generalized pretraining for face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2024)
51. Liu, Y., Chen, Y., Gou, M., Huang, C.T., Wang, Y., Dai, W., Xiong, H.: Towards unsupervised domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
52. Mu, L., Bai, J., He, X., Ye, J., Liang, X., Yang, Y., Zhuang, J., Hu, H.: Tegdg: Textually guided domain generalization for face anti-spoofing. arXiv preprint arXiv:2311.18420 (2023)
53. Nag, S., Zhu, X., Song, Y.Z., Xiang, T.: Zero-shot temporal action detection via vision-language prompting. In: European Conference on Computer Vision (ECCV). pp. 681–697 (2022)
54. Niu, X., Yu, Z., Han, H., Li, X., Shan, S., Zhao, G.: Video-based remote physiological measurement via cross-verified feature disentangling. In: European Conference on Computer Vision (ECCV). pp. 295–310 (2020)
55. Panwar, A., Singh, P., Saha, S., Paudel, D.P., Van Gool, L.: Unsupervised compound domain adaptation for face anti-spoofing. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG). pp. 1–8 (2021)
56. Parelli, M., Delitzas, A., Hars, N., Vlassis, G., Anagnostidis, S., Bachmann, G., Hofmann, T.: Clip-guided vision-language pre-training for question answering in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5606–5611 (2023)
57. Patel, K., Han, H., Jain, A.K., Ott, G.: Live face video vs. spoof face video: Use of moiré patterns to detect replay video attacks. In: International Conference on Biometrics (ICB). pp. 98–105 (2015)
58. Pramanick, S., Song, Y., Nag, S., Lin, K.Q., Shah, H., Shou, M.Z., Chellappa, R., Zhang, P.: Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5285–5297 (2023)

59. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
60. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
61. Sarhan, M.H., Navab, N., Eslami, A., Albarqouni, S.: Fairness by learning orthogonal disentangled representations. In: European Conference on Computer Vision (ECCV). pp. 746–761 (2020)
62. Schwartz, W.R., Rocha, A., Pedrini, H.: Face spoofing detection through partial least squares and low-level descriptors. In: International Joint Conference on Biometrics (IJCB). pp. 1–8 (2011)
63. Shao, R., Lan, X., Li, J., Yuen, P.C.: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
64. Shao, R., Lan, X., Li, J., Yuen, P.C.: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10023–10031 (2019)
65. Shao, R., Lan, X., Yuen, P.C.: Regularized fine-grained meta face anti-spoofing. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 11974–11981 (2020)
66. Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. Advances in Neural Information Processing Systems (NeurIPS) pp. 14274–14289 (2022)
67. Singh, A.K., Joshi, P., Nandi, G.C.: Face liveness detection through face structure analysis. International Journal of Applied Pattern Recognition (IJAPR) pp. 338–360 (2014)
68. Smith, D.F., Wiliem, A., Lovell, B.C.: Face recognition on consumer devices: Reflections on replay attacks. IEEE Transactions on Information Forensics and Security (TIFS) pp. 736–745 (2015)
69. Srivatsan, K., Naseer, M., Nandakumar, K.: Flip: Cross-domain face anti-spoofing with language guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19685–19696 (2023)
70. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7167–7176 (2017)
71. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS) (2017)
72. Wang, C.Y., Lu, Y.D., Yang, S.T., Lai, S.H.: Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20281–20290 (2022)
73. Wang, G., Han, H., Shan, S., Chen, X.: Improving cross-database face presentation attack detection via adversarial domain adaptation. In: International Conference on Biometrics (ICB). pp. 1–8 (2019)

74. Wang, G., Han, H., Shan, S., Chen, X.: Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6678–6687 (2020)
75. Wang, G., Han, H., Shan, S., Chen, X.: Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. IEEE Transactions on Information Forensics and Security (TIFS) pp. 56–69 (2020)
76. Wang, J., Zhang, J., Bian, Y., Cai, Y., Wang, C., Pu, S.: Self-domain adaptation for face anti-spoofing. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 2746–2754 (2021)
77. Wang, J., Ge, Y., Yan, R., Ge, Y., Lin, K.Q., Tsutsui, S., Lin, X., Cai, G., Wu, J., Shan, Y., et al.: All in one: Exploring unified video-language pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6598–6608 (2023)
78. Wang, M., Xing, J., Liu, Y.: Actionclip: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472 (2021)
79. Wang, W., Liu, P., Zheng, H., Ying, R., Wen, F.: Domain generalization for face anti-spoofing via negative data augmentation. IEEE Transactions on Information Forensics and Security (TIFS) pp. 2333–2344 (2023)
80. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for all vision and vision-language tasks. arXiv preprint arXiv:2208.10442 (2022)
81. Wang, Z., Xu, Y., Wu, L., Han, H., Ma, Y., Li, Z.: Improving face anti-spoofing via advanced multi-perspective feature learning. ACM Transactions on Multimedia Computing, Communications and Applications (TOMM) pp. 1–18 (2023)
82. Wang, Z., Wang, Z., Yu, Z., Deng, W., Li, J., Gao, T., Wang, Z.: Domain generalization via shuffled style assembly for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4123–4133 (2022)
83. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Transactions on Information Forensics and Security (TIFS) pp. 746–761 (2015)
84. Wu, W., Sun, Z., Ouyang, W.: Revisiting classifier: Transferring vision-language models for video recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)). pp. 2847–2855 (2023)
85. Xie, J., Zheng, S.: Zero-shot object detection through vision-language embedding alignment. In: IEEE International Conference on Data Mining Workshops (ICDMW). pp. 1–15 (2022)
86. Yan, Z., Zhang, Y., Fan, Y., Wu, B.: Ucf: Uncovering common features for generalizable deepfake detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22412–22423 (2023)
87. Yang, J., Lei, Z., Liao, S., Li, S.Z.: Face liveness detection with component dependent descriptor. In: International Conference on Biometrics (ICB). pp. 1–6 (2013)
88. Yin, W., Ming, Y., Tian, L.: A face anti-spoofing method based on optical flow field. In: International Conference on Signal Processing (ICSP). pp. 1333–1337 (2016)
89. Yu, Z., Li, X., Niu, X., Shi, J., Zhao, G.: Face anti-spoofing with human material perception. In: European Conference on Computer Vision (ECCV). pp. 557–575 (2020)

90. Yu, Z., Li, X., Shi, J., Xia, Z., Zhao, G.: Revisiting pixel-wise supervision for face anti-spoofing. IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM) pp. 285–295 (2021)
91. Yu, Z., Wan, J., Qin, Y., Li, X., Li, S.Z., Zhao, G.: Nas-fas: Static-dynamic central difference network search for face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) pp. 3005–3023 (2020)
92. Yue, H., Wang, K., Zhang, G., Feng, H., Han, J., Ding, E., Wang, J.: Cyclically disentangled feature translation for face anti-spoofing. arXiv preprint arXiv:2212.03651 (2022)
93. Zeng, Y., Zhang, X., Li, H.: Multi-grained vision language pre-training: Aligning texts with visual concepts. arXiv preprint arXiv:2111.08276 (2021)
94. Zhang, K.Y., Yao, T., Zhang, J., Tai, Y., Ding, S., Li, J., Huang, F., Song, H., Ma, L.: Face anti-spoofing via disentangled representation learning. In: European Conference on Computer Vision (ECCV). pp. 641–657 (2020)
95. Zhang, Y., Yin, Z., Li, Y., Yin, G., Yan, J., Shao, J., Liu, Z.: Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In: European Conference on Computer Vision (ECCV). pp. 70–85 (2020)
96. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: International Conference on Biometrics (ICB). pp. 26–31 (2012)
97. Zhang, Z., Yi, D., Lei, Z., Li, S.Z.: Face liveness detection by learning multispectral reflectance distributions. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG). pp. 436–441 (2011)
98. Zheng, G., Liu, Y., Dai, W., Li, C., Zou, J., Xiong, H.: Towards unified representation of invariant-specific features in missing modality face anti-spoofing. In: European Conference on Computer Vision (ECCV) (2024)
99. Zheng, T., Li, B., Wu, S., Wan, B., Mu, G., Liu, S., Ding, S., Wang, J.: Mfae: Masked frequency autoencoders for domain generalization face anti-spoofing. IEEE Transactions on Information Forensics and Security (TIFS) (2024)
100. Zheng, T., Yu, Q., Chen, Z., Wang, J.: Famim: A novel frequency-domain augmentation masked image model framework for domain generalizable face anti-spoofing. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4470–4474 (2024)
101. Zheng, Z., Ma, M., Wang, K., Qin, Z., Yue, X., You, Y.: Preventing zero-shot transfer degradation in continual learning of vision-language models. arXiv preprint arXiv:2303.06628 (2023)
102. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16793–16803 (2022)
103. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision (IJCV) pp. 2337–2348 (2022)
104. Zhou, Q., Zhang, K.Y., Yao, T., Lu, X., Ding, S., Ma, L.: Test-time domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 175–187 (2024)
105. Zhou, Q., Zhang, K.Y., Yao, T., Lu, X., Yi, R., Ding, S., Ma, L.: Instance-aware domain generalization for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20453–20463 (2023)

106. Zhou, Q., Zhang, K.Y., Yao, T., Yi, R., Ding, S., Ma, L.: Adaptive mixture of experts learning for generalizable face anti-spoofing. In: ACM International Conference on Multimedia (ACM MM). pp. 6009–6018 (2022)
107. Zhou, Q., Zhang, K.Y., Yao, T., Yi, R., Sheng, K., Ding, S., Ma, L.: Generative domain adaptation for face anti-spoofing. In: European Conference on Computer Vision (ECCV). pp. 335–356 (2022)