CFPL-FAS: Class Free Prompt Learning for Generalizable Face Anti-spoofing

Ajian Liu[1], Shuai Xue[2], Jianwen Gan[3], Jun Wan[1,4,5]*, Yanyan Liang[4]*
Jiankang Deng[6], Sergio Escalera[7], Zhen Lei[1,5,8]
[1]MAIS, CASIA, China; [2]BIT, Zhuhai; [3]GC&UKLIS, Wuzhou University; [4]M.U.S.T, Macau
[5]SAI, UCAS, China; [6]ICL, UK; [7]CVC, Spain; [8]CAIR, HKISI, CAS
[1]{ajian.liu,jun.wan}@ia.ac.cn, [4]yyliang@must.edu.mo [1]zlei@nlpr.ia.ac.cn

Abstract

*Domain generalization (DG) based Face Anti-Spoofing (FAS) aims to improve the model's performance on unseen domains. Existing methods either rely on domain labels to align domain-invariant feature spaces, or disentangle generalizable features from the whole sample, which inevitably lead to the distortion of semantic feature structures and achieve limited generalization. In this work, we make use of large-scale VLMs like CLIP and leverage the textual feature to dynamically adjust the classifier's weights for exploring generalizable visual features. Specifically, we propose a novel Class Free Prompt Learning (**CFPL**) paradigm for DG FAS, which utilizes two lightweight transformers, namely Content Q-Former (CQF) and Style Q-Former (SQF), to learn the different semantic prompts conditioned on content and style features by using a set of learnable query vectors, respectively. Thus, the generalizable prompt can be learned by two improvements: (1) A Prompt-Text Matched (**PTM**) supervision is introduced to ensure CQF learns visual representation that is most informative of the content description. (2) A Diversified Style Prompt (**DSP**) technology is proposed to diversify the learning of style prompts by mixing feature statistics between instance-specific styles. Finally, the learned text features modulate visual features to generalization through the designed Prompt Modulation (**PM**). Extensive experiments show that the CFPL is effective and outperforms the state-of-the-art methods on several cross-domain datasets.*

1. Introduction

Face Anti-Spoofing (FAS) is an important step in protecting the security of face recognition systems from print-attack [58], replay-attack [4] and mask-attack [6, 20]. Despite the existing presentation attack detection (PAD) methods [7, 19, 21, 30, 32, 52, 55] obtain remarkable performance in intra-dataset experiments where training and testing data are from the same domain, their performance severely degraded in cross-dataset experiments due to large distribution discrepancies among different domains. Domain Generalization (DG) based FAS aims to mitigate the impact of distribution discrepancy by accessing multiple domains. As shown in Fig. 1 (a) (left), a typical strategy [12, 40, 41, 50] is rely on domain labels to learn a domain-invariant feature space by adversarial training, which is also generalized to unseen domains. While it is difficult to seek a compact and generalized feature space for all domains. Even, there is no guarantee that such a feature

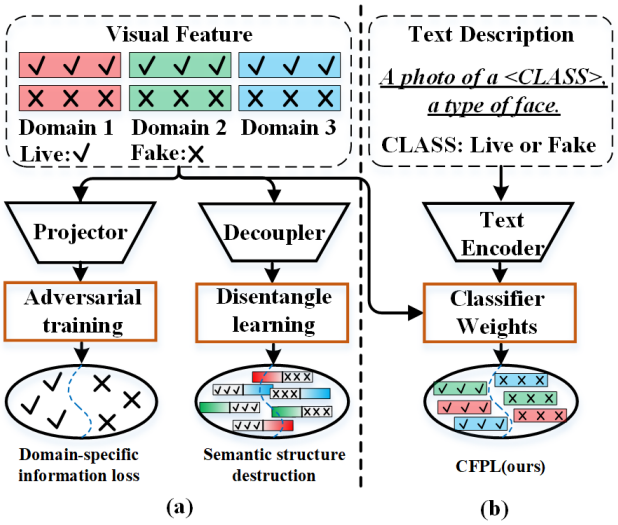


Figure 1. Comparison with existing DG FAS methods. (a) the previous methods either rely on a projector to align domain-invariant feature spaces with adversarial training, or disentangle generalizable features from the whole sample with a decoupler, which inevitably leads to the distortion of semantic structures and achieves limited generalization. (b) Our CFPL framework is built on CLIP to learn generalized visual features by using the text features as weights of the classifier.

*Corresponding author.

space exists among multiple domains due to large distribution discrepancies. Considering the manually labeled domain labels are arbitrary and subjective, which cannot truly reflect the diversity of samples in a domain, as shown in Fig. 1 (a) (right), another strategy [3, 27, 63] is based on the instance to disentangle or generate generalized features from liveness-irrelevant features by disentangled representation learning.Based on the above analysis, these methods either rely on domain labels to align domain-invariant feature spaces, or disentangle generalizable features from instance-specific and liveness-irrelevant features, which inevitably leads to the distortion of semantic feature structures and achieve limited generalization.

Rethinking the reason for the poor generalization of FAS can be attributed to the liveness-irrelevant features interfering with the classifier's recognition of spoofing clues. If the weight of the classifier can be dynamically adjusted based on sample instances, such as, to weaken the interference factors and strengthen the generalized features, it will effectively improve its generalization. Inspired by large vision-language models like *CLIP* [39], which jointly trains an image encoder and a text encoder to predict the correct of pairings (image, text), As shown in Fig. 1 (b), we generate corrective text features by a high-capacity text encoder, which allows open-set visual concepts and broader semantic spaces compared to discrete domain labels. How to learn generalizable prompts for text encoder to adaptively adjust the classifier's weights? Unlike general DG task, which have domain information such as 'This image belongs to Cartoon/Sketch/Art Painting/Photo', a FAS dataset is regarded as a domain, usually named with the publishing agency, such as MSU [51], CASIA [58], Idiap [4], or OULU [1], without providing any valuable domain semantic information to assist generalizable prompt learning.

In this work, without relying on domain semantics, the generalizable prompts can be based on image content and style learned by reducing their correlation [50, 63]. Based on this determination, inspired by BLIP-2 [14] and TGPT [44], we design two lightweight transformers, namely Content Q-Former (CQF) and Style Q-Former (SQF), to learn the expected prompts conditioned on content and style features by using a set of learnable query vectors, respectively. To further ensure CQF can learn to extract the visual representation that is most informative of the content description, we introduce a Prompt-Text Matched (**PTM**) supervision to optimize the learning of content prompts, where each sample's content description is generated by template description. Due to the inability to accurately describe style information in text, instead, we propose a Diversified Style Prompt (**DSP**) technology to diversify style prompts by mixing feature statistics between instance-specific styles. Finally, the generalized visual features are learned through the designed Prompt Modulation (**PM**) function, which uses visual features as modulation factors. To sum up, the main contributions of this paper are summarized as follows:

- Instead of directly manipulating visual features, it is the first work to explore DG FAS via textual prompt learning, namely CFPL, which allows a broader semantic space to adjust the visual features to generalization.
- In order to release the requirement for categories in the text description, our CFPL first learns the prompts conditioned on content and style features with two lightweight transformers, namely Content Q-Former (CQF) and Style Q-Former (SQF). Then, the Prompts are further optimized through two improvements: (1) A Prompt-Text Matched (**PTM**) aims to ensure CQF learns semantic visual representation; (2) A Diversified Style Prompt (**DSP**) technology to diversify the learning of style prompts. Finally, the learned prompts modulate visual features to generalization through the designed Prompt Modulation (**PM**) function.
- Extensive cross-domain experiments show that the proposed CFPL is effective and outperforms the state-of-the-art (SOTA) methods by an undeniable margin.

2. Related Work

2.1. Face Anti-Spoofing

Methods on Intra-datasets. The essence of FAS is a defensive measure for face recognition systems and has been studied for over a decade. Some CNN-based methods [10, 29, 31, 32, 45] design a unified framework of feature extraction and classification in an end-to-end manner. Intuitively, the live faces in any scene have consistent face-like geometry. Inspired by this, some works [30, 41, 49, 53] leverage the physical-based depth information instead of binary classification loss as supervision, which are more faithful attack clues in any domain. With the popularity of high-quality 2d attacks, *i.e.*, OULU-NPU [1], SiW [30], CelebA-Spoof [57] and high-fidelity mask attacks, *i.e.*, MARsV2 [26], WMCA [9, 37], HiFiMask [18, 20] and SuHiFiMask [5, 6] with more realistic in terms of color, texture, and geometry structure, it is very challenging to mine spoofing traces from the visible spectrum alone. Methods based on multimodal fusion [8, 9, 16, 17, 56] have proven to be effective in alleviating the above problems. The motivation for these methods is that indistinguishable fake faces may exhibit quite different properties under the other spectrum. In order to alleviate the limitation of consistency between testing and training modalities, flexible modality based methods [15, 22, 54] aims to improve the performance on any single modality by leveraging available multimodal data. However, above methods are not specially designed to solve the domain generalization.

Domain Generalization Methods. Domain Adaptation

(DA) [33, 36, 46] aims to minimize the distribution discrepancy between the source and target domain by leveraging the unlabeled target data. However, the target data is difficult to collect, or even unknown during training. Domain Generalization (DG) can conquer this by taking the advantage of multiple source domains without seeing any target data. MADDG [40], SSDG [12], DR-MD-Net [47] aim to learn a generalized feature space via adversarial training. SSAN [50] reduces the model's overfitting of style by randomly assembling the content and style of the samples. RFM [41], MT-FAS [38], D^{2}AM [3], and SDA [48] aim to find the generalized feature directions via meta-learning strategies. In addition to aligning a domain-invariant feature space, SA-FAS [43] encourages domain separability while aligning the live-to-spoof transition to be the same for all domains. Considering the domain information lies in the style features, ANRL [27] and SSAN [50] use IN strategy to separate complete representation into content and style features according to image statistics. DRDG [28] iteratively reweights the relative importance between samples to further improve the generalization. The latest emerging strategy is to improve the generalization with the help of domain-specific information. CIFAS [34] adopts causal intervention with backdoor adjustment to mitigate domain bias for learning generalizable features. AMEL [62] exploits the domain-specific feature as a complement to common domain-invariant features to further improve the generalization. IADG [63] learns generalizable visual features by weakening the features' sensitivity to instance-specific styles. In addition to supervised training, Liu et al. [35] propose the first unsupervised DG framework for FAS, which could exploit large amounts unlabeled data to learn generalizable features.

2.2. Vision-Language Models (VLMs).

The vision-language models have undergone a leapfrog development since CLIP [39] was proposed. This approach has stimulated thinking and innovation in many fields, such as object detection, image generation [25], and image forgery detection [23, 24]. In terms of text models, GPT-3 already has strong language processing capabilities. By combining it with visual basic models to build, and adding some necessary link parameters, the trained vision-language model achieves a combined understanding of images and text. For example, BLIP [14] has made significant progress in multimodality by freezing the constructed image encoder and text encoder during training to train an additional small query transformer. Similarly, LLaVa and minigpt-4 reduce the cost of model training by simply linearly mapping image features to the word embedding space. On the basis of deep exploration of VLM by researchers, we want to further extend VLM in FAS.

3. CFPL: Class Free Prompt Learning

3.1. Semanticized Prompts Generation

Visual Content and Style features. Considering the challenge of DG FAS is the interference of liveness-irrelevant signals on spoofing cues, we need to model these two types of information with different prompts and alleviate this interference by reducing their correlation. Based on previous research [50, 63], the liveness-irrelevant signals lie more in the instance-specific styles while spoofing clues is an image attribute hidden in the content. Based on this determination, we design Content Q-Former (CQF) and Style Q-Former (SQF) to generate content and style prompts conditioned corresponding visual features, respectively. Inspired by Adaptive instance regularization (AdaIN) [11], given a sample, we first calculate the mean and standard deviation at l-th layer, i.e., $\mu(\boldsymbol{v}^l)$ and $\sigma(\boldsymbol{v}^l)$; Then, we concatenate them to obtain the style statistics $\boldsymbol{v}_s^l$ of this layer; Finally, the style feature $\boldsymbol{v}_s$ is obtained by averaging style statistics from all layers for this sample. We calculate the content feature $\boldsymbol{v}_c$ of the sample from the output of the image encoder by normalizing. The details are as follows:

$$\begin{aligned} \boldsymbol{v}_s &= \frac{\sum_{l=1}^{L} \boldsymbol{v}_s^l}{L}, \boldsymbol{v}_s^l = \left[\mu(\boldsymbol{v}^l) \,||\, \sigma(\boldsymbol{v}^l)\right], \boldsymbol{v}_s \in \mathbb{R}^{1\times 2d}, \\ \boldsymbol{v}_c &= \frac{\boldsymbol{v}^L - \mu(\boldsymbol{v}^L)}{\sigma(\boldsymbol{v}^L)}, \boldsymbol{v}_c \in \mathbb{R}^{n\times d} \end{aligned} \tag{1}$$

where L is the total layers of the image encoder. $[\cdot\,||\,\cdot]$ represents concatenating features along embedding dimension.

CQF and SQF. As shown in Fig. 2, CQF and SQF share a similar backbone, which consists of alternating layers of multiheaded self-attention (MSA), multiheaded cross-attention (MCA) and MLP blocks. Firstly, we create N learnable query embeddings $\boldsymbol{Q} = \{\boldsymbol{q}^1, \boldsymbol{q}^2, \cdots, \boldsymbol{q}^N\} \in \mathbb{R}^{N\times d}$ as input to the backbone, where each query has a dimension of $d = 512$ (same dimension with multi-modal embedding space); Then, the queries interact with each other through MSA block, and interact with image features $\boldsymbol{v} \in \mathbb{R}^d$ through MCA block; Finally, we obtain the prompt $\boldsymbol{P} = \{\boldsymbol{p}^1, \boldsymbol{p}^2, \ldots, \boldsymbol{p}^N\} \in \mathbb{R}^{N\times d}$ after the queries pass through MLP block. This process can be expressed as:

$$\begin{aligned} \boldsymbol{Q'} &= \boldsymbol{Q} + \texttt{MSA}(\texttt{LN}(\boldsymbol{Q})), \boldsymbol{Q'} \in \mathbb{R}^{N\times d} \\ \boldsymbol{Q''} &= \boldsymbol{Q'} + \texttt{MCA}(\texttt{LN}(\boldsymbol{Q'}), \texttt{LN}(\boldsymbol{v})), \boldsymbol{Q''} \in \mathbb{R}^{N\times d} \\ \boldsymbol{P} &= \boldsymbol{Q''} + \texttt{MLP}(\texttt{LN}(\boldsymbol{Q''})), \boldsymbol{P} \in \mathbb{R}^{N\times d} \end{aligned} \tag{2}$$

where Layernorm (LN) is applied before every block, and residual connections after every block. The MLP contains two layers with a GELU non-linearity. Based on this training mechanism, CQF and SQF can bridge the gap between visual and language modalities by interacting prompt queries with corresponding image features, and output the content prompt $\boldsymbol{P}_c$ and style prompt $\boldsymbol{P}_s$, respectively.

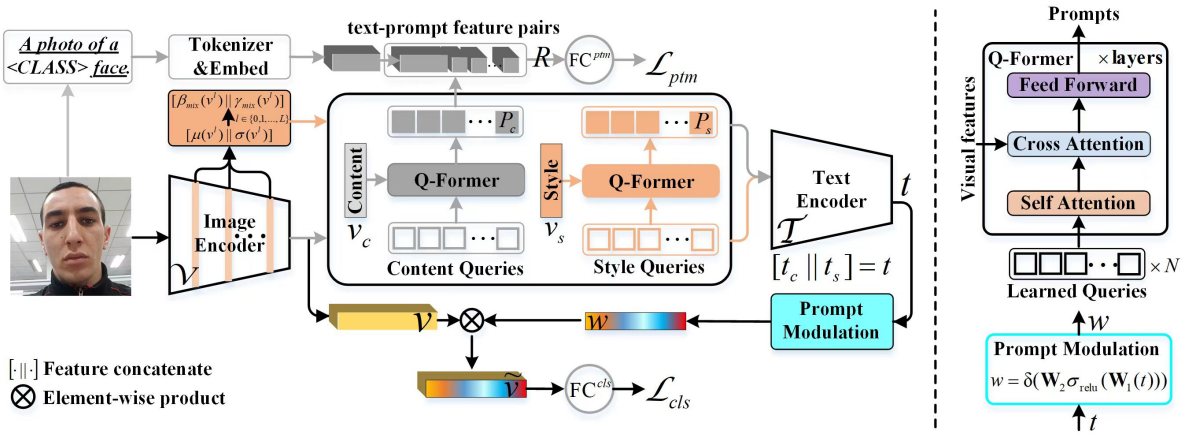


Figure 2. Our CFPL is built on CLIP [39] consists of image encoder $\mathcal{V}$ and text encoder $\mathcal{T}$, and adaptes to FAS tasks via prompt learning with four contributions: (1) CQF and SQF. CFPL introduces two lightweight transformers, namely Content Q-Former (CQF) and Style Q-Former (SQF) to learn the different semantic prompts conditioned on content and style features from the image encoder by using a set of learnable query vectors, respectively; (2) Prompt-Text Matched (**PTM**) supervision. The fixed template description of each sample is used as a supervise to ensure CQF learns semantic visual representation; (3) A Diversified Style Prompt (**DSP**). The style from each layer of the image encoder is diversified through mixing feature statistics; (4) Prompt Modulation (**PM**). The generalized visual feature is adjusted by the modulation factor, which is generated by the text feature through the designed modulation function.

3.2. Generalized Prompt Optimization

Text Supervision in Content Prompt.

Due to the lack of semantics for CLIP in the FAS categories, it is not suitable to align queries and text representations with the concept of maximizing their mutual information. Instead, we guide the CQF to understand the FAS's categories at a higher level with a binary classification task, where the model is asked to predict whether a prompt-text pair is matched (PTM).

Given a min-batch with B samples, we first generate a text description containing category attribute for each sample, such as "a photo of a ⟨CLASS⟩ face.", where "CLASS" is "live" for real face and "fake" for spoof face, respectively; After that, the text descriptions $\boldsymbol{T}$ are transformed into text supervisions $\boldsymbol{S} \in \mathbb{R}^{B\times 77\times d}$ through `Tokenizer` and `Embed` layers, sequentially; Then, we construct positive and negative feature pairs of prompt-text for prediction by CQF. Specifically, we concatenate content prompt $\boldsymbol{P_c} \in \mathbb{R}^{B\times N\times d}$ and text supervision $\boldsymbol{S}$ according to the embedding dimension, and obtain the positive feature pairs $\boldsymbol{R}_p \in \mathbb{R}^{B\times N\times 2d}$. We adopt the hard negative mining strategy from ALBEF [13] to create informative negative pairs. Such as for each prompt, one negative text is selected with the contrastive similarity distribution, where texts that are more similar to the prompt have a higher chance of being sampled. A similar strategy for one hard negative prompt for each text. Therefore, we can obtain negative feature pairs $\boldsymbol{R}_n^{prompt} \in \mathbb{R}^{B\times N\times 2d}$ and $\boldsymbol{R}_n^{text} \in \mathbb{R}^{B\times N\times 2d}$ by mining prompt and text, respectively; After that, we concatenate all positive and negative feature pairs $\boldsymbol{R}_p$, $\boldsymbol{R}_n^{prompt}$, and $\boldsymbol{R}_n^{text}$ to obtain the joint features $\boldsymbol{R} \in \mathbb{R}^{3B\times N\times 2d}$ according to the batch dimension. This process can be expressed as:

$$\begin{aligned}
\boldsymbol{S} &= \texttt{Embed}(\texttt{Tokenizer}(\boldsymbol{T})), \boldsymbol{S} \in \mathbb{R}^{B\times 77\times d}, \\
\boldsymbol{S} &= \texttt{Mean\&Expand}(\boldsymbol{S}), \boldsymbol{S} \in \mathbb{R}^{B\times N\times d}, \\
\boldsymbol{R}_p &= \left[\boldsymbol{P} \,||\, \boldsymbol{S}\right]_2, \boldsymbol{R}_p \in \mathbb{R}^{B\times N\times 2d}, \\
\boldsymbol{R} &= \left[\boldsymbol{R}_p \,||\, \boldsymbol{R}_n^{prompt} \,||\, \boldsymbol{R}_n^{text}\right]_0, \boldsymbol{R} \in \mathbb{R}^{3B\times N\times 2d}
\end{aligned} \tag{3}$$

where the number of word tokens in text supervisions is aligned with the number of queries in the content prompt by averaging (`Mean`) and expanding (`Expand`) N times. $[\cdot\,||\,\cdot]_{\texttt{dim}}$ represents concatenating features along `dim` dimension. Finally, the optimization of text supervision is achieved by predicting the matched and unmatched probabilities for the joint features $\boldsymbol{R}$:

$$\mathcal{L}_{ptm} = \sum_{i=1}^{3B} \mathcal{H}(\boldsymbol{y}_i^{ptm}, \texttt{Mean}(\texttt{FC}^{ptm}(\boldsymbol{R}_i))) \tag{4}$$

where we feed each query embedding into a two-class linear classifier to obtain a logit, and average (`Mean`) the logits across all queries as the output matching score. $\mathcal{H}(.,.)$ is the cross-entropy loss, $\texttt{FC}^{ptm}$ is a fully-connected layer followed by softmax, and $\boldsymbol{y}^{ptm} \in \{0,1\}$ is a 2-dimensional one-hot vector representing the ground-truth label.

Diversified Style Prompt. Due to the indescribability of the sample style, we are unable to complete this task using text supervision. Implicitly, we borrow a strategy from MixStyle [59] that mixes style feature statistics between instances to achieve diversification of style prompts.

Specifically, given the visual style stastics $[\mu(\boldsymbol{v}), \sigma(\boldsymbol{v})]$ of a min-batch, we first obtain reference statistics $\mu(\hat{\boldsymbol{v}})$ and

Method	OCI→M			OMI→C			OCM→I			ICM→O			avg.
	HTER↓	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER
MADDG [40]	17.69	88.06	-	24.50	84.51	-	22.19	84.99	-	27.98	80.02	-	23.09
DR-MD-Net [47]	17.02	90.10	-	19.68	87.43	-	20.87	86.72	-	25.02	81.47	-	20.64
RFMeta [41]	13.89	93.98	-	20.27	88.16	-	17.30	90.48	-	16.45	91.16	-	16.97
NAS-FAS [52]	19.53	88.63	-	16.54	90.18	-	14.51	93.84	-	13.80	93.43	-	16.09
D^2AM [3]	12.70	95.66	-	20.98	85.58	-	15.43	91.22	-	15.27	90.87	-	16.09
SDA [48]	15.40	91.80	-	24.50	84.40	-	15.60	90.10	-	23.10	84.30	-	19.65
DRDG [28]	12.43	95.81	-	19.05	88.79	-	15.56	91.79	-	15.63	91.75	-	15.66
ANRL [27]	10.83	96.75	-	17.83	89.26	-	16.03	91.04	-	15.67	91.90	-	15.09
SSDG-R [12]	7.38	97.17	-	10.44	95.94	-	11.71	96.59	-	15.61	91.54	-	11.28
SSAN-R [50]	6.67	98.75	-	10.00	96.67	-	8.88	96.79	-	13.72	93.63	-	9.81
PatchNet [45]	7.10	98.46	-	11.33	94.58	-	13.40	95.67	-	11.82	95.07	-	10.91
SA-FAS [43]	5.95	96.55	-	8.78	95.37	-	6.58	97.54	-	10.00	96.23	-	7.82
IADG [63]	5.41	98.19	-	8.70	96.44	-	10.62	94.50	-	8.86	97.14	-	8.39
CFPL(Ours)	**3.09**	**99.45**	94.28	**2.56**	**99.10**	66.33	**5.43**	**98.41**	85.29	**3.33**	**99.05**	90.06	**3.60**
ViTAF*-5-shot [10]	2.92	99.62	91.66	1.40	99.92	98.57	1.64	99.64	91.53	5.39	98.67	76.05	2.83
FLIP-MCL* [42]	4.95	98.11	74.67	0.54	99.98	100.00	4.25	99.07	84.62	2.31	99.63	92.28	3.01
CFPL*(Ours)	1.43	99.28	98.57	2.56	99.10	66.33	5.43	98.41	85.29	2.50	99.42	94.72	2.98

Table 1. The results (%) of Protocol 1 on MSU-MFSD (M), CASIA-FASD (C), ReplayAttack (I), and OULU-NPU (O) datasets. Note that the $*$ indicates the corresponding method using CelebA-Spoof [57] as the supplementary source dataset and '5-shot' represents 5 images from the target datasets participating in the training phase.

$\sigma(\hat{\boldsymbol{v}})$ by shuffling the order of batch dimension for $\mu(\boldsymbol{v})$ and $\sigma(\boldsymbol{v})$, respectively; Then, we generate mixture of feature statistics γ_{mix} and β_{mix} through a weighted approach:

$$\begin{aligned} \gamma_{mix} &= \lambda\sigma(\boldsymbol{v}) + (1-\lambda)\sigma(\hat{\boldsymbol{v}}), \\ \beta_{mix} &= \lambda\mu(\boldsymbol{v}) + (1-\lambda)\mu(\hat{\boldsymbol{v}}) \end{aligned} \tag{5}$$

where λ is an instance-specific, random weight sampled from the beta distribution, $\lambda \sim Beta(\alpha, \alpha)$. α is set to 0.1 according to the suggestion in [59]. Finally, the mixture of style statistics $[\beta_{mix}, \gamma_{mix}]$ are used to calculate style features according to Eq. 1.

3.3. Prompt Modulation on Visual Features

Class Free Prompt Modulation. Due to the content and style prompts are generated based on sample instances, they are more suitable as a set of fine-tuning factors (class free) for adaptively recalibrating channel-wise visual feature responses, compared to using them as classifier's weights (with class) to predict visual feature. In this work, we use prompts to distinguish between generalized features and liveness-irrelevant signals by explicitly modeling interdependencies between channels.

Concretely, we first input content prompt $\boldsymbol{P_c}$ and style prompt $\boldsymbol{P_s}$ into the text encoder to produce text features, $\boldsymbol{t_c} \in \mathbb{R}^{B\times d}$ and $\boldsymbol{t_s} \in \mathbb{R}^{B\times d}$, respectively; After that, we concatenate these two types of text features along the embedding dimension to obtain modulation features $\boldsymbol{t} \in \mathbb{R}^{B\times 2d}$ with rich visual concepts; Then, we employ a gating mechanism $\mathtt{g}_e$ with a sigmoid activation δ to map modulation features to the weighting factors $\boldsymbol{w} \in \mathbb{R}^{B\times d}$. This process can be expressed as:

$$\begin{aligned} \boldsymbol{w} &= \delta(\mathtt{g}_e(\boldsymbol{t}, \mathbf{W})) = \delta(\mathbf{W}_2\sigma_{\text{relu}}(\mathbf{W}_1\boldsymbol{t})), \\ \tilde{\boldsymbol{v}} &= \left[\tilde{\boldsymbol{v}}^1, \tilde{\boldsymbol{v}}^2, \ldots, \tilde{\boldsymbol{v}}^d\right], \tilde{\boldsymbol{v}}^c = \boldsymbol{w}^c \cdot \boldsymbol{v}^c, \\ \mathcal{L}_{cls} &= \sum_{i=1}^{B} \mathcal{H}(\boldsymbol{y}_i^{cls}, \mathtt{FC}^{cls}(\tilde{\boldsymbol{v}}_i)), \tilde{\boldsymbol{v}} \in \mathbb{R}^{B\times d} \end{aligned} \tag{6}$$

where σ_{relu} is the ReLU function, $\mathbf{W}_1 \in \mathbb{R}^{\frac{d}{r}\times 2d}$ and $\mathbf{W}_2 \in \mathbb{R}^{d\times\frac{d}{r}}$ are trainable parameters for two fully connected (FC) layers in $\mathtt{g}_e$ function. r is a reduction ratio, with a value of 16 in this work. Finally, the adapted visual features $\tilde{\boldsymbol{v}} \in \mathbb{R}^{B\times d}$ are obtained by weighting the channel-wise feature $\boldsymbol{v}^c$ with the scalar $\boldsymbol{w}^c$, and append a fully-connected ($\mathtt{FC}^{cls}$) layer followed by softmax to predict a two-class (i.e., live or fake) probability. $y^{cls} \in \{0, 1\}$ is the label for live or spoof face.

Model Training and Inference. In the training stage, parameters from two designed Q-Formers, i.e., CQF and SQF, two fully connected layers for classifiers, i.e., $\mathtt{FC}^{ptm}$ and $\mathtt{FC}^{cls}$, one gate function, i.e., $\mathtt{g}_e$, and image encoder $\mathcal{V}$ are updated and the text encoder $\mathcal{T}$ is fixed. The full training objective of CFPL is:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{ptm} \tag{7}$$

In the inference stage, our CQF and SQF will adaptively generate the semanticized prompt as input to the text encoder based on each sample instance. Finally, the text encoder generates continuous and widely adjustable modulation factors for weighting visual features to generalization.

Method	CS→W			SW→C			CW→S			avg.
	HTER↓	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER
ViT* [10]	7.98	97.97	73.61	11.13	95.46	47.59	13.35	94.13	49.97	10.82
ViTAF*-5-shot [10]	2.91	99.71	92.65	6.00	98.55	78.56	11.60	95.03	60.12	6.83
FLIP-MCL* [42]	4.46	99.16	83.86	9.66	96.69	59.00	11.71	95.21	57.98	8.61
CFPL*(Ours)	4.40	99.11	85.23	8.13	96.70	62.41	8.50	97.00	55.66	7.01
ViT [10]	21.04	89.12	**30.09**	17.12	89.05	22.71	17.16	90.25	30.23	18.44
CLIP-V [39]	20.00	87.72	16.44	17.67	89.67	20.70	**8.32**	**97.23**	57.28	15.33
CLIP [39]	17.05	89.37	8.17	15.22	**91.99**	17.08	9.34	96.62	**60.75**	13.87
CoOp [61]	9.52	90.49	10.68	18.30	87.47	11.50	11.37	95.46	40.40	13.06
CFPL (Ours)	**9.04**	**96.48**	25.84	**14.83**	90.36	**8.33**	8.77	96.83	53.34	**10.88**

Table 2. The results (%) of Protocol 2 on CASIA-SURF (S), CASIA-SURF CeFA (C), and WMCA (W) datasets. Note that the $*$ indicates the corresponding method using CelebA-Spoof [57] as the supplementary source dataset and '5-shot' represents 5 images from the target datasets participating in the training phase.

4. Experimental Setup

Datasets, Protocols and Evaluation Metrics. Following the prior work [10], two Protocols are used to evaluate the generalization in this work. For Protocol 1, we use four widely-used benchmark datasets, MSU-MFSD (M) [51], CASIA-FASD (C) [58], Idiap Replay-Attack (I) [4], and OULU-NPU (O) [1]. For Protocol 2, we use RGB samples in CASIA-SURF (S) [56], CASIA-SURF CeFA (C) [16], and WMCA (W) [9] datasets, which contain more subjects, diverse attack types, and rich collection environments. In each Protocol, we treat each dataset as a domain and apply the leave-one-out testing for generalization evaluation. We adopt three metrics to evaluate the performance of a model: (1) HTER. It computes the average of the FRR and the FAR. (2) AUC. It evaluates the theoretical performance of the model. (3) TPR at a fixed False Positive Rate (FPR=1%). It can be used to select a suitable trade-off threshold according to a given real application.

Implementation Details. We set the length of style and content queries to 16, where each query has a dimension of 512; The depth of the CQF and SQF is set to 1. Style prompt diversification is activated in the training phase with a probability of 0.5 and does not participate in the test phase; All models are trained with a batch size of 12, an Adam optimizer with a weight decay of 0.05. The minimum learning rate at the second stage is $1e-6$. We resize images to 224×224, augmented with random resized cropping and horizontal flipping, and train all models with 500 epochs.

4.1. Cross-domain Results

For Protocol 1, we report the results of recent SOTA methods on Tab. 1, such as SA-FAS [43] encourages domain separability while aligning the live-to-spoof transition; IADG [63] learns the generalizable feature by weakening the features' sensitivity to instance-specific styles; FLIP-MCL* [42] improves the visual representations with the help of natural language, and ViTAF*-5-shot [10] tackles a real-world anti-spoofing when 5 images are available from target datasets. From Tab. 1, without using CelebA-Spoof [57], it can be seen that our method achieves the best performance for all metrics on four test datasets. Specifically, on the HTER metric (similar conclusions on AUC), CFPL outperforms IADG [63] for all target domains with lower values: **M** (3.09% *vs.*, 5.41%), **C** (2.56% *vs.*, 8.70%), **I** (5.43% *vs.*, 10.62%) and **O** (3.33% *vs.*, 8.86%). Finally, an average HTER of 3.60% is achieved, significantly better than the previous best result of 7.82%. After introducing the CelebA-Spoof [57] dataset as the supplementary training data, CFPL* outperforms FLIP-MCL* [42] in terms of the average metric of HTER, such as 2.98% *vs.*, 3.01%. It indicates that our algorithm aligns images with text representation by class-free prompt learning is superior to an ensemble of class descriptions. In addition, CFPL* is slightly inferior to the ViTAF*-5-shot [10] on the average of HTER, such as 2.98% *vs.*, 2.83%, which uses 5 additional samples from target domain in training data, greatly alleviating the overfitting on the domain-specific distribution.

For Protocol 2, we list the results of different methods on Tab. 2. Without using CelebA-Spoof [57], our CFPL significantly surpasses several baselines in terms of the HTER metric when tested on the **W** and **C** domains. However, when tested on the **S** dataset, CFPL is slightly inferior to the CLIP-V [39] on the HTER (8.77% *vs.*8.32%), AUC (96.83% *vs.*97.23%) and TPR@FPR=1% (53.34% *vs.*57.28%) metrics. Due to the obvious spoofing traces on the CASIA-SURF dataset [56], relying solely on visual features is relatively generalizable. Similar to the conclusion of Protocol 1, CFPL* outperforms baseline ViT* [10] and FLIP-MCL* [42] when introducing the CelebA-Spoof [57] dataset into training data. For example, our CFPL* achieves the optimal mean on the HTER (7.01%) metric.

Baseline	PTM	DSP	PM	HTER(%)↓	AUC(%)	TPR(%) @FPR=1%
CoOp [61]	-	-	-	8.78	94.77	43.71
✓	-	-	-	8.11	96.09	51.59
✓	✓	-	-	7.50	96.39	54.78
✓	✓	✓	-	7.08	96.79	57.61
✓	✓	✓	✓	6.72	97.09	60.35

Table 3. Ablation of each component, where each result is the average on all sub-protocols.

Method	HTER(%)↓	AUC(%)	TPR(%)@FPR=1%
CoCoOp [60]	6.80	97.27	60.41
CQF	5.12	98.65	73.67
SQF	4.84	98.75	87.08
CFPL	3.33	99.05	90.06

Table 4. Ablation of the structures for CQF and SQF on ICM→O.

HTER(%)↓	Length			
Depth	×8	×16	×32	×64
×1	3.47	*3.33*	*3.33*	***3.30***
×4	3.45	**3.42**	3.45	3.45
×8	3.56	3.56	**3.47**	3.47
×12	*3.41*	**3.33**	3.33	3.33

Table 5. Ablation of the length for Queries and the depth for Q-former on ICM→O. The optimal value for each row/column is represented in bold/italics.

4.2. Ablation Study

Effectiveness of each component. In order to evaluate the contribution of each component in our framework, such as Text Supervision (abbreviated as 'PTM'), Diversification of Style Prompt (abbreviated as 'DSP'), and Prompt Modulation (abbreviated as 'PM'), we conduct ablation studies on Protocol 1 and 2 by gradually introducing one of them into the Baseline (abbreviated as 'B'), where the Baseline is obtained by removing all contributions from the CFPL. And report the average results on all sub-protocols in Tab. 3.

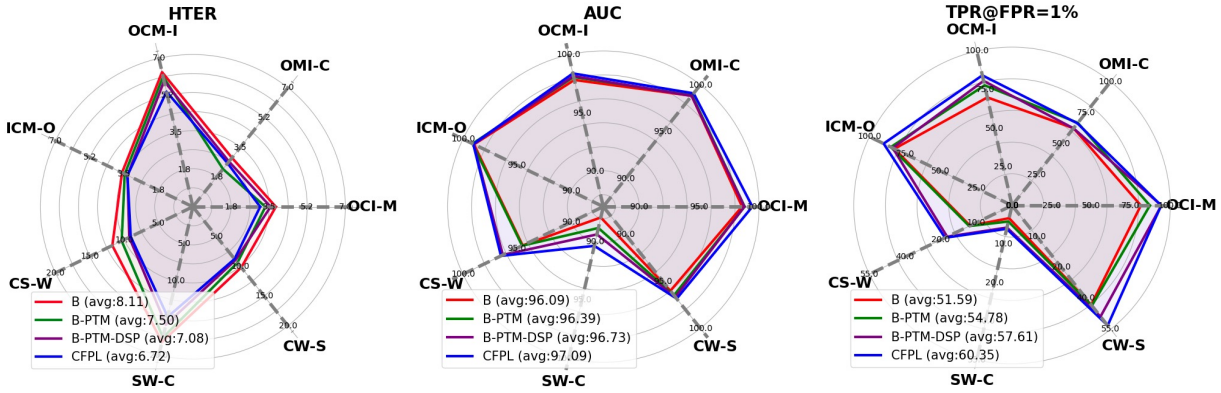
Specifically, instead of modeling a prompt's context words with free learnable vectors in CoOp [61], our 'Baseline' introduces two lightweight transformers CQF and SQF to learn the expected prompts conditioned on content and style features from the image encoder, achieving significant generalization benefits of -0.67% (HTER), +1.32% (AUC) and +7.88% (TPR@FPR=1%), respectively. After introducing the text supervision in the content prompt, the results of the three metrics are optimized to 7.50% (HTER), 96.39% (AUC), and 54.78% (TPR@FPR=1%), respectively. It indicates that promoting the content prompt carries sufficient category attributes that can be converted into generalization benefits. Further diversification of style statistics in style prompts can further expand performance benefits, i.e., -0.42% (HTER), +0.4% (AUC), and +2.83% (TPR@FPR=1%). Instead of using the designed gate function $\mathcal{G}_e$, we calculate the modulation factor $\boldsymbol{w}$ by calculating the mean of content and style features, such as $\boldsymbol{w} = (\boldsymbol{t_c} + \boldsymbol{t_s})/2, \boldsymbol{w} \in \mathbb{R}^{B\times d}$. From the Tab. 3, it can be seen that if removing the designed gate function, the generalization significantly decreased from 6.72% (HTER), 97.09% (AUC) and 60.35% (TPR@FPR=1%) to 7.08% (HTER), 96.79% (AUC) and 57.61% (TPR@FPR=1%).

Furthermore, in Fig. 3, we detailed the result of each method on three metrics across all sub-protocols, where the red line represents the Baseline, and the blue line represents our CFPL. From Fig. 3, it can be clearly seen that the blue line shrinks with the smallest area in the polar coordinate system of the HTER, while is distributed with the largest area for the AUC and TPR@FPR=1%. The opposite conclusion applies to the Baseline of red lines. Almost all methods have an undeniable advantage over Baseline.

Effect of the Structures of CQF and SQF. In Tab. 4, we list the results of CoCoOp [60], CFPL removing SQF (abbreviated as 'CQF'), CFPL removing CQF (abbreviated as 'SQF'), and CFPL on the ICM→O experiment. From Tab. 4, we can see that a simple two-layer bottleneck structure cannot effectively alleviate the FAS task with significant domain differences, such as achieving results of 6.80%, 97.27%, and 60.41% on metrics HTER, AUC, and TPR@FPR=1%. When replacing MetaNet with CQF and SQF, we obtain performance benefits of -1.68% (HTER), +1.38% (AUC), and +13.26% (TPR@FPR=1%) and -1.96% (HTER), +1.48% (AUC) and +26.67% (TPR@FPR=1%), respectively. Finally, by collaborating with CQF and SQF, our CFPL achieves the best results, such as 3.33% (HTER), 99.05% (AUC), and 90.06% (TPR@FPR=1%). Furthermore, our CFPL further improves their performance through collaborative CQF and SQF, indicating that the CQF and SQF not only bring significant benefits but also have a positive collaborative effect.

The Length of Queries and the Depth of Q-Former. The number of learnable queries and the depth of CQF and SQF can also affect the performance. We search for the optimal value on the ICM→O experiment for the length of queries and the depth of Q-Former from two sets of changing values, such as $[8, 16, 32, 64]$ for the former, and $[1, 4, 8, 12]$ for the latter, respectively.

From Tab. 5, the following two conclusions can be drawn: (1) The length of the queries is set to around 16, achieving optimal performance. By observing at different depth settings, the number of learnable queries increases ex-



(a) (b) (c)

Figure 3. The results of each method on three metrics across all sub-protocols, where the red line represents the Baseline, and the blue line represents our CFPL. For the HTER metric, the smaller area enclosed by lines, the better performance of the corresponding methods. The opposite conclusion applies to metrics AUC and TPR@FPR=1%.

ponentially, the performance benefits of the model are subtle, and there is even a trend toward degradation. Specifically, at a length of 16, models at different depths achieve decent performance, with values of 3.33%, 3.42%, 3.56%, and 3.33% for HTER. (2) The depth of Q-Former has a negligible impact on the performance. In detail, under each length setting, the performance of models with different depths fluctuates around a certain value of HTER. For example, when the length is set to 8, the HTER is 3.45%, while when the length is 16, 32, and 64, the HTER is 3.33%. Based on the experimental results, we suggest setting the number of queries to 16 and the depth of the Q-Former to 1.

4.3. Visualization and Analysis

With attention-model explainability tool [2], we visually validate the superiority of the proposed CFPL from the visual attention maps, compared to the Baseline, which is obtained by removing all contributions from the CFPL. The results on all protocols are shown in Fig. 4, where the maps of the Baseline correspond to misclassified samples, while our CFPL correctly classifies these samples.

Specifically, for the OCM→I experiment on Protocol 1, the Baseline classifies live face errors due to focusing more on the background. Our CFPL correctly classifies it by correcting the focus area to the boundary between face and background. For the playback attack, the Baseline does not pay attention to spoofing clues, such as the reflection spot on the electronic screen, and the feature map area illuminated by our method. Similar conclusions can be drawn on other sub-protocols.

5. Conclusion

In this work, we target DG FAS via textual prompt learning for the first time, and present a cross-domain FAS framework CFPL, which utilizes two lightweight transformers,

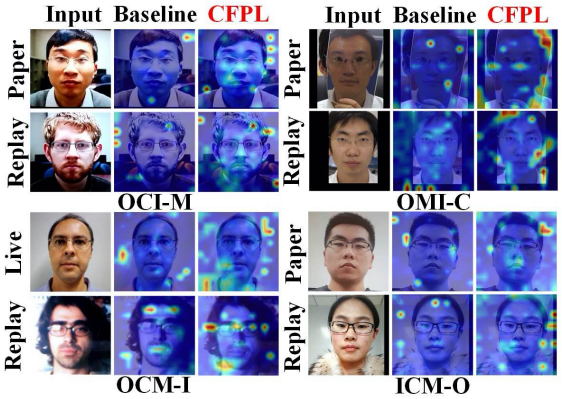


Figure 4. Using visualization tool [2], the attention maps on all sub-protocols from Protocol 1, where the Baseline caused classification errors due to its failure to detect spoofing regions, and our CFPL correctly classifies these samples by correcting the region of interest.

CQF and SQF, to learn the different semantic prompts conditioned on content and style features. Finally, we introduce text supervision, diverse style prompt, and prompt modulation to promote the generalization.

6. Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant 2021YFE0205700, Beijing Natural Science Foundation JQ23016, the Science and Technology Development Fund of Macau Project 0123/2022/A3, 0096/2023/RIA2, and 0070/2020/AMJ, Guangdong Provincial Key R&D Programme: 2019B010148001, the Chinese National Natural Science Foundation Project 62276254, U23B2054, and the InnoHK program.

References

[1] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *FGR*, pages 612–618, 2017. 2, 6

[2] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, 2021. 8

[3] Zhihong Chen, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Feiyue Huang, and Xinyu Jin. Generalizable representation learning for mixture domain face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1132–1139, 2021. 2, 3, 5

[4] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, 2012. 1, 2, 6

[5] Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, and Zhen Lei. Surveillance face presentation attack detection challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6360–6370, 2023. 2

[6] Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Chenxu Zhao, Xu Zhang, Stan Z Li, and Zhen Lei. Surveillance face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 2023. 1, 2

[7] Anjith George and Sébastien Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *ICB*, 2019. 1

[8] Anjith George and Sébastien Marcel. Cross modal focal loss for rgbd face anti-spoofing. In *CVPR*, pages 7882–7891, 2021. 2

[9] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *TIFS*, 2019. 2, 6

[10] Hsin-Ping Huang, Deqing Sun, Yaojie Liu, Wen-Sheng Chu, Taihong Xiao, Jinwei Yuan, Hartwig Adam, and Ming-Hsuan Yang. Adaptive transformers for robust few-shot cross-domain face anti-spoofing. *arXiv preprint arXiv:2203.12175*, 2022. 2, 5, 6

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 3

[12] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8484–8493, 2020. 1, 3, 5

[13] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 4

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3

[15] Ajian Liu and Yanyan Liang. Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1180–1186. International Joint Conferences on Artificial Intelligence Organization, 2022. 2

[16] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1179–1187, 2021. 2, 6

[17] Ajian Liu, Zichang Tan, Jun Wan, Yanyan Liang, Zhen Lei, Guodong Guo, and Stan Z Li. Face anti-spoofing via adversarial cross-modality translation. *IEEE Transactions on Information Forensics and Security*, 16:2759–2772, 2021. 2

[18] Ajian Liu, Chenxu Zhao, Zitong Yu, Anyang Su, Xing Liu, Zijian Kong, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zhen Lei, et al. 3d high-fidelity mask face presentation attack detection challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 814–823, 2021. 2

[19] Ajian Liu, Jun Wan, Ning Jiang, Hongbin Wang, and Yanyan Liang. Disentangling facial pose and appearance information for face anti-spoofing. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4537–4543. IEEE, 2022. 1

[20] Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, et al. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 17:2497–2507, 2022. 1, 2

[21] Ajian Liu, Zichang Tan, Yanyan Liang, and Jun Wan. Attack-agnostic deep face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6335–6344, 2023. 1

[22] Ajian Liu, Zichang Tan, Zitong Yu, Chenxu Zhao, Jun Wan, Yanyan Liang, Zhen Lei, Du Zhang, S. Li, and Guodong Guo. Fm-vit: Flexible modal vision transformers for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 18:4775–4786, 2023. 2

[23] Huan Liu, Zichang Tan, Qiang Chen, Yunchao Wei, Yao Zhao, and Jingdong Wang. Unified frequency-assisted transformer framework for detecting and grounding multi-modal manipulation. *arXiv preprint arXiv:2309.09667*, 2023. 3

[24] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Yao Zhao, and Jingdong Wang. Forgery-aware adaptive transformer for generalizable synthetic image detection. *arXiv preprint arXiv:2312.16649*, 2023. 3

[25] Huan Liu, Xiaolong Liu, Zichang Tan, Xiaolong Li, and Yao Zhao. Padvg: A simple baseline of active protection for audio-driven video generation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20 (6), 2024. 3

[26] Siqi Liu, Baoyao Yang, Pong C Yuen, and Guoying Zhao. A 3d mask face anti-spoofing database with real world variations. In *CVPRW*, 2016. 2

[27] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Mingwei Bi, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Adaptive normalized representation learning for generalizable face anti-spoofing. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1469–1477, 2021. 2, 3, 5

[28] Shubao Liu, Ke-Yue Zhang, Taiping Yao, Kekai Sheng, Shouhong Ding, Ying Tai, Jilin Li, Yuan Xie, and Lizhuang Ma. Dual reweighting domain generalization for face presentation attack detection. *arXiv preprint arXiv:2106.16128*, 2021. 3, 5

[29] Yaojie Liu and Xiaoming Liu. Spoof trace disentanglement for generic face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3813–3830, 2022. 2

[30] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, 2018. 1, 2

[31] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *CVPR*, 2019. 2

[32] Yaojie Liu, Joel Stehouwer, and Xiaoming Liu. On disentangling spoof trace for generic face anti-spoofing. In *ECCV*, pages 406–422. Springer, 2020. 1, 2

[33] Yuchen Liu, Yabo Chen, Wenrui Dai, Mengran Gou, Chun-Ting Huang, and Hongkai Xiong. Source-free domain adaptation with contrastive domain alignment and self-supervised exploration for face anti-spoofing. In *ECCV*, 2022. 3

[34] Yuchen Liu, Yabo Chen, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Causal intervention for generalizable face anti-spoofing. In *ICME*, 2022. 3

[35] Yuchen Liu, Yabo Chen, Mengran Gou, Chun-Ting Huang, Yaoming Wang, Wenrui Dai, and Hongkai Xiong. Towards unsupervised domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[36] Yuchen Liu, Yabo Chen, Wenrui Dai, Mengran Gou, Chun-Ting Huang, and Hongkai Xiong. Source-free domain adaptation with domain generalized pretraining for face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[37] Zohreh Mostaani, Anjith George, Guillaume Heusch, David Geissbuhler, and Sebastien Marcel. The high-quality wide multi-channel attack (hq-wmca) database, 2020. 2

[38] Yunxiao Qin, Zitong Yu, Longbin Yan, Zezheng Wang, Chenxu Zhao, and Zhen Lei. Meta-teacher for face anti-spoofing. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 6

[40] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, 2019. 1, 3, 5

[41] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Regularized fine-grained meta face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11974–11981, 2020. 1, 2, 3, 5

[42] Koushik Srivatsan, Muzammal Naseer, and Karthik Nandakumar. Flip: Cross-domain face anti-spoofing with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19685–19696, 2023. 5, 6

[43] Yiyou Sun, Yaojie Liu, Xiaoming Liu, Yixuan Li, and Wen-Sheng Chu. Rethinking domain generalization for face anti-spoofing: Separability and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24563–24574, 2023. 3, 5, 6

[44] Hao Tan, Jun Li, Yizhuang Zhou, Jun Wan, Zhen Lei, and Xiangyu Zhang. Compound text-guided prompt tuning via image-adaptive cues. *arXiv preprint arXiv:2312.06401*, 2023. 2

[45] Chien-Yi Wang, Yu-Ding Lu, Shang-Ta Yang, and Shang-Hong Lai. Patchnet: A simple face anti-spoofing framework via fine-grained patch recognition. pages 20281–20290, 2022. 2, 5

[46] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Improving cross-database face presentation attack detection via adversarial domain adaptation. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019. 3

[47] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. *TIFS*, 2020. 3, 5

[48] Jingjing Wang, Jingyi Zhang, Ying Bian, Youyi Cai, Chunmao Wang, and Shiliang Pu. Self-domain adaptation for face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2746–2754, 2021. 3, 5

[49] Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In *CVPR*, 2020. 2

[50] Zhuo Wang, Zezheng Wang, Zitong Yu, Weihong Deng, Jiahong Li, Tingting Gao, and Zhongyuan Wang. Domain generalization via shuffled style assembly for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4123–4133, 2022. 1, 2, 3, 5

[51] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE TIFS*, 2015. 2, 6

[52] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z. Li, and Guoying Zhao. Nas-fas: Static-dynamic central difference network search for face anti-spoofing. In *TPAMI*, 2020. 1, 5

[53] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *CVPR*, 2020. 2

[54] Zitong Yu, Ajian Liu, Chenxu Zhao, Kevin H. M. Cheng, Xu Cheng, and Guoying Zhao. Flexible-modal face anti-spoofing: A benchmark, 2023. 2

[55] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Ying Tai, Shouhong Ding, Jilin Li, Feiyue Huang, Haichuan Song, and Lizhuang Ma. Face anti-spoofing via disentangled representation learning. In *ECCV*, 2020. 1

[56] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *TBMIO*, 2(2):182–193, 2020. 2, 6

[57] Yuanhan Zhang, Zhenfei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *ECCV*, 2020. 2, 5, 6

[58] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *ICB*, 2012. 1, 2, 6

[59] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 4, 5

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7

[61] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 6, 7

[62] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Shouhong Ding, and Lizhuang Ma. Adaptive mixture of experts learning for generalizable face anti-spoofing. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6009–6018, 2022. 3

[63] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Ran Yi, Shouhong Ding, and Lizhuang Ma. Instance-aware domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20453–20463, 2023. 2, 3, 5, 6