# Bottom-Up Domain Prompt Tuning for Generalized Face Anti-Spoofing

Si-Qi Liu[1,*,✉], Qirui Wang[1,*], and Pong C. Yuen[2]

[1] Shenzhen Research Institute of Big Data
[2] Department of Computer Science, Hong Kong Baptist University
siqiliu@sribd.cn, qwangcw@connect.ust.hk, pcyuen@comp.hkbu.edu.hk

**Abstract.** Face anti-spoofing (FAS) which plays an important role in securing face recognition systems has been attracting increasing attention. Recently, vision-language model CLIP has been proven to be effective for FAS, where outstanding performance can be achieved by simply transferring the class label into textual prompt. In this work, we aim to improve the generalization ability of CLIP-based FAS from a prompt learning perspective. Specifically, a Bottom-Up Domain Prompt Tuning method (BUDoPT) that covers the different levels of domain variance, including the domain of recording settings and domain of attack types is proposed. To handle domain discrepancies of recording settings, we design a context-aware adversarial domain-generalized prompt learning strategy that can learn domain-invariant prompt. For spoofing domain with different attack types, we construct a fine-grained textual prompt that guides CLIP to look through the subtle details of different attack instruments. Extensive experiments are conducted on five FAS datasets with variations of camera types, resolutions, image qualities, lighting conditions, and recording environments. The effectiveness of our proposed method is evaluated with different amounts of source domains from multiple angles, where we boost the generalizability compared with the state of the arts with multiple or limited numbers of training datasets.

**Keywords:** Face Anti-spoofing · Face presentation attack detection · Face liveness detection

## 1 Introduction

With the wide deployment of face recognition in real-world scenarios, the security concerns of FR systems attract increasing attention. Face images can be easily acquired from social networks, and the costs of conducting a face spoofing attack can be very low. Prints or screens of mobile phones or tablets are the two common low-cost spoofing instruments. 3D mask attack could be more challenging since both the 3D shape and skin texture can be imitated. To counter these presentation attacks, a series of face anti-spoofing (FAS) methods based on different liveness cues are proposed, various from the hand-crafted features [1,19,40] to the deep-learning-based representations [34, 62, 64].

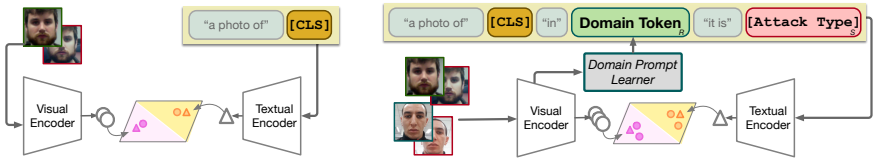
* Equal contribution, ✉ Corresponding author

Early works are mainly developed for intra-domain testing where the recording environment, devices, and attack types are identical. Since the main difference between genuine and attack samples is located on small vestiges such as the texture and color tone, the subtle variations introduced by recording cameras, lighting conditions, and attack mediums form the domain gaps in FAS.

To improve the robustness and generalizability, the generic domain adaptation and domain generalization approaches are applied or customized in the context of FAS [52,53,57]. Few-shot domain adaptation and unsupervised domain adaptation methods tend to have better results since they can learn from the testing data with or without the labels [21,23,54]. The zero-shot domain generalization setting is more practical as Since the testing data may not be accessible in real scenarios. Popular techniques such as adversarial learning, contrastive learning, or meta-learning are adopted to align representations from different domains and keep on boosting the performances [5, 44]. Directly constructing generalized liveness cues can also be effective. The facial heartbeat signal (rPPG) based methods exhibit strong generalizability to different types of attacks but existing methods still require long observation time [29, 30].

Recently, vision-language models like CLIP exhibited astonishing generalizability and zero-shot performance with the joint embedding of image-text pairs. The first trial of CLIP for FAS (FLIP) that transforms binary class labels into text and measures their similarity to image embeddings also achieves high performance [49]. However, since CLIP is trained for generic tasks with image pairs from the internet, the domain discrepancy of lighting conditions, camera qualities, or attack types in FAS can hardly be well addressed by simply finetuning it with the image-to-class-text contrastive loss. Intuitively, we may enrich the class text prompt with domain labels like the camera types, lighting conditions, or recording scene and construct hand-crafted prompts such as `{a photo of [CLS] captured under [Recording Setting]}` to learn a fine-grained representation. However, this approach may not work since 1) the detailed domain information in FAS can hardly be defined with text since it is related to multiple factors; 2) learning fine-grained representation that can tell domain discrepancies may contradict the final classification of genuine and spoofing samples.

In this work, we aim to improve the generalization ability of CLIP-based FAS from a prompt learning perspective. Specifically, we proposed to learn a domain-invariant prompt from the visual embedding that can guide CLIP to avoid interference from the specific domain. Inspired by the CLIP style contrastive learning objective which pulls images to their textual descriptions while pushes away unmatched pairs, we achieve domain-generalized prompt learning by pulling together image embeddings with a domain-invariant prompt and pushing away the image embeddings with their domain-specific prompt. Considering the discrepancies of recording domain are low-level features that scatter in the entire image, we extract domain-specific information from both the facial content and background context regions as the domain-specific negative prompt (push away). The domain-generalized prompt embedding extract from multiple domains is regarded as the positive prompt (pull together). An adversarial




(a) Intuitive solution of CLIP for Face Anti-Spoofing (b) *Bottom-Up* Domain-Generalized Prompt Tuning for FAS (ours)

**Fig. 1:** The comparison between (a) the intuitive solution of CLIP for face anti-spoofing and (b) the proposed bottom-up domain-generalized prompt tuning. The `Domain Token` prompt guides the model to learn domain-generalized FAS features for different recording domains. The `Attack Type` prompt boosts the discriminability by encouraging the model to differentiate fine-grained attack instruments. The Domain Token is learned from the visual embedding through the *Domain Prompt Learner*.

learning strategy is designed to further ensure the prompt learner can learn the domain-generalized information. The various attack instruments for spoofing samples can be regarded as another perspective of domains. Different from the treatment of recording domain prompt learning, we want the visual encoder to distinguish the attack types so that it can learn more discriminative FAS features [20, 24]. In this case, we form a textual prompt that guides the generic visual encoder to adapt to the FAS task and look through the fine-grained difference for various attack types. We refer to this newly proposed method that learns the low-level domain-generalized visual prompt and tunes high-level attack-type textual prompt as the Bottom-Up Domain Prompt Tuning (BUDoPT) for generalized face anti-spoofing.

Extensive experiments are conducted on five FAS datasets with a large number of variations (camera types, resolutions, image qualities, lighting conditions, and recording environments). The effectiveness of our proposed method is evaluated with different amounts of source domains from multiple angles, where we boost the generalizability compared with the state of the arts with multiple training datasets or with only one dataset.

In sum, the main contributions of this work are:

- We improve the domain generalizability of FAS from a prompt tuning perspective within the context of vision-language model CLIP. A Bottom-Up Domain Prompt Tuning method (BUDoPT) that covers the different levels of domain variance, including the domain of recording settings and domain of attack types is proposed.
- We design 1) a context-aware adversarial domain-generalized prompt learning strategy that can learn domain-invariant prompt for different recording settings, 2) an effective textual prompt that guides CLIP to adapt to FAS task and look through the subtle details of different attack types.
- We improve the generalizability compared with the state of the arts with different scales of training domains, and give in-depth analysis with extensive experiments on five FAS datasets that cover typical variations in FAS.

## 2 Related Work

**Face Anti-Spoofing**. Face anti-spoofing has attracted increasing attention in the last decades, developed from the early handcrafted liveness feature construction [37, 60] to the latest deep learning-based design [64]. Early works focus on searching the discriminative liveness cues such as the appearance artifacts [31, 42, 58], and the motion difference between genuine faces and static attack instruments. Accordingly, hand-crafted features, such as SIFT [41], HoG [60], and LBP [3] are used to characterize spoof patterns. More discriminative features can be learned with Deep learning-based [38, 59]. Auxiliary tasks, such as the regression of facial depth map [32], 3D structure [12], and material [61], encourage the network to learn fine-grained appearance details. There are also classification-based methods [34, 51, 59] and Generative models [22, 33, 57]. Spatiotemporal information can also be learned through CNN and LSTM framework [9, 39].

**Generalized FAS**. Generalized FAS that can perform robustly on different domains are necessary for real-world application scenarios. Recent works proposed to learn the FAS model on multiple source domains and evaluated a target domain. Under the domain adaptation settings where the target domain is accessible, aligning the source and target feature distributions is the typical solution [21,52,53,65,71]. Adversarial learning strategies or the triplet loss encourage the model to learn shared feature for generalization [20, 47, 57]. Meta-learning and continue learning approaches are also proposed by [5,6,27,44,48] to simulate the domain shift during training.

**Prompt Learning**. Prompt learning originated from the field of Natural Language Processing (NLP) motivates to shape the output of pre-trained language models by providing contextual information or constraints, thereby enhancing their ability to adapt to downstream tasks [43]. Prompt engineering was extended from NLP to Visual-Language Models [45], such as CLIP which classifies images by matching them with class prompts. CoOp [69] is designed to turn the handcrafted textual prompt into a learnable visual prompt. CoCoOp [68] incorporates the image as the condition for prompts and improves the model's generalization capabilities. Within the visual-textual contrastive learning framework, generalizability can be improved with a smaller data scale. CFPL[1] [25] and this work aims to improve the generalizability for FAS with prompt learning. Different from CFPL which regards style (in AdaIN [18]) as the domain-specific cue, in this work we learn the domain-invariant prompt from an intrinsic CLIP-style contrastive learning perspective.

## 3 Preliminary on CLIP for FAS

CLIP [45] comprises an image encoder denoted $f_v$ and a text encoder denoted $f_t$. In the context of the FAS, the two class text prompt $\boldsymbol{p_k}$ encompasses the

[1] CFPL was published after this manuscript was submitted for review

descriptions of real and spoof faces, representing two classes. CLIP determines the similarity between the input image $\boldsymbol{x}$ and $\boldsymbol{p}$ instead of directly training the classification head. The model predicts the image's category $\hat{y}$ by selecting the class prompt with the highest response:

$$\hat{y} = \underset{k \in \{\mathrm{r,s}\}}{\operatorname{argmax}} \langle f_v(\boldsymbol{x}), f_t(\boldsymbol{p}_k) \rangle \tag{1}$$

where $\langle , \rangle$ denotes the cosine simiarity. $p_r$ and $p_s$ represent the prompt with real and spoof face descriptions.

With such learning framework, pretrained CLIP exhibits great performance in FAS with the standard finetuning process [49].However, since CLIP is trained for generic tasks with image pairs from the internet, when handling the low-level feature in cross-domain settings in FAS, the domain discrepancy of lighting conditions, camera qualities, or attack types can hardly be well addressed.

## 4 Proposed method

**Overview**. The Bottom-Up Domain Prompt Tuning method (BUDoPT) is proposed to improve the generalization ability of FAS in the context of the vision language model CLIP. On top of the intuitive CLIP-based face anti-spoofing which turns the class label into textual prompt, we incorporate the domain token for the domain discrepancies from different recording settings and the attack type token that describes the fine-grained types of attacks. First, the context-aware adversarial domain-generalized prompt learning is to train the domain prompt learner given the input of facial visual feature. A CLIP score style domain prompt alignment mechanism and an adversarial learning strategy are designed to learn a domain-generalized prompt. Second, the textual prompt of fine-grained attack types aims to boost discriminability by looking through the subtle details of different attack instruments.

### 4.1 Context-aware Domain Generalized Prompt Learning

**Recording Domain Generalized Prompt Learning**. Although text prompt can effectively assist in classifying concrete instances in upper-stream tasks, it encounters difficulties in describing the recording domain with various complex factors, including camera qualities, lighting conditions, and blurriness in FAS. Therefore we proposed the to learn domain-generalized prompt from the visual embedding. Inspired by [68], we employ vision domain features as conditional inputs to generate learnable prompts. The primary objective of the component is learning domain-invariant prompts to mitigate the impact of domain gaps.

Specifically, we propose a Prompt learner $PL$ to obtain the domain prompt from the image $\boldsymbol{x}$. $PL$ utilizes the features from $i$-th layer of $f_v$ as the input. The domain feature $D$ is defined as:

$$D(\boldsymbol{x}) = PL(f_{v,i}(\boldsymbol{x})), \tag{2}$$

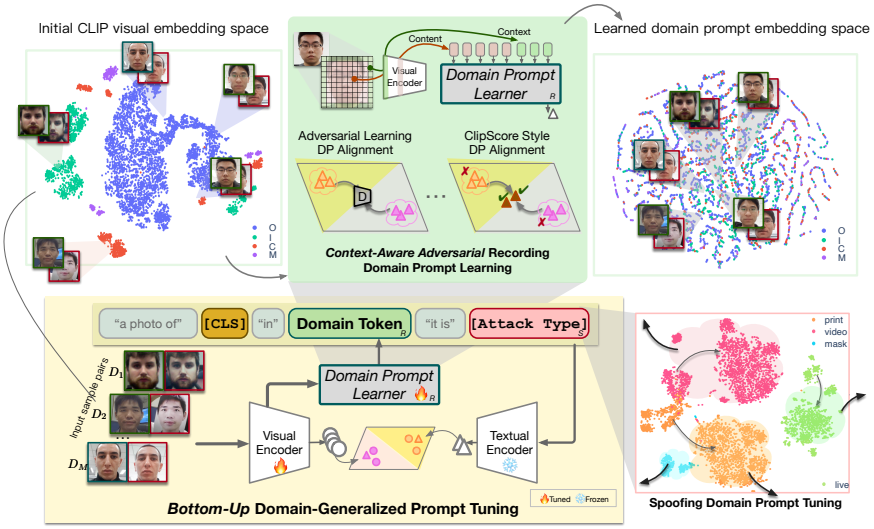



**Fig. 2:** The proposed Bottom-Up Domain Prompt Tuning method (BUDoPT) covers the different levels of domain variance for generalized FAS, including the domain of recording settings and domain of attack types. The context-aware adversarial domain-generalized prompt learning can learn domain-invariant prompt for different recording settings. The textual prompt of fine-grained attack types can boost the discriminability by looking through the subtle details of different attack instruments.

where $f_{v,i}$ denotes $i$-layer features of $f_v$.

To accurately represent the domain feature while minimizing influences from the content feature of a single image, we utilize all images' domain features in a mini-batch to generate the domain prompt. Given the training mini-batch $\{\boldsymbol{x}_n\}_{n=1}^{N}$ for domain $m$, the domain prompt $\boldsymbol{p}_{D_m}$ is defined as:

$$\boldsymbol{p}_{D_m} = f_t(\frac{1}{N}\sum_{n=1}^{N} D(\boldsymbol{x}_n^m)), \tag{3}$$

where $N$ is the mini-batch size and $x_n^m$ denotes the $n$-th image from $m$.

Inspired by the CLIP style contrastive learning objective, we design a CLIP score-based domain generalization learning strategy that pulls together image embeddings with a domain-invariant prompt and pushes away the image embeddings with their domain-specific prompt. Given a batch of training data with $M$ domains, the generalized domain prompt $\boldsymbol{p}_{GD}$ can be defined as:

$$\boldsymbol{p}_{GD} = f_t(\frac{1}{N*M}\sum_{m=1}^{M}\sum_{n=1}^{N} D(\boldsymbol{x}_n^m)), \tag{4}$$

As the Figure 3 shows, only the labels belonging to the generalized domain can be marked as positive while other domain-specific prompts are always labeled

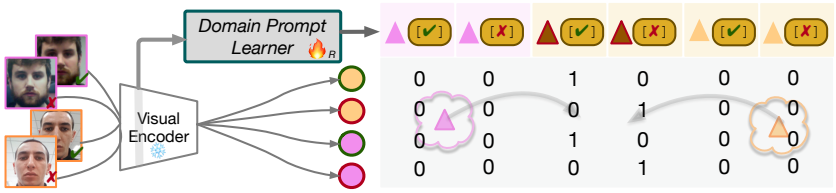



**Fig. 3:** Illustration of the CLIP score-based domain generalization learning strategy, where we pull together image embeddings with a domain-generalized prompt (▲) and push away the image embeddings with their domain-specific prompt

as negative. We denote $\boldsymbol{p}$ as the set of prompts and then calculate the image cross-entropy loss $\mathcal{L}_{ce}$ by:

$$\mathcal{L}_{ce} = -\sum_{z=1}^{2\times(M+1)} y \log\left(\frac{\exp(\langle f_v(\boldsymbol{x}), f_t(\boldsymbol{p}_z)\rangle)}{\sum_{k=1}^{2\times(M+1)} \exp(\langle f_v(\boldsymbol{x}), f_t(\boldsymbol{p}_k)\rangle)}\right), \tag{5}$$

where $y$ is the ground truth label, $k$ and $z$ denote the indices in $\boldsymbol{p}$. Note that we have $2 \times (M+1)$ classes in total, i.e., $M$ domains plus $\boldsymbol{p}_{GD}$ for each class since the prompt set $\boldsymbol{p} = \{\boldsymbol{p}^{real}_{D_1}, \cdots, \boldsymbol{p}^{real}_{D_M}, \boldsymbol{p}^{real}_{GD}; \boldsymbol{p}^{fake}_{D_1}, \cdots, \boldsymbol{p}^{fake}_{D_M}, \boldsymbol{p}^{fake}_{GD}\}$,

During inference, we assume that all samples in the batch are sourced from the same domain and possess the identical prompt. The model selects the class prompt with the highest similarity score as the final prediction.

**Context-aware Domain Clue**. To better learn the domain-generalized prompt, we try to extract more domain-specific information from the visual embedding as it is treated as the negative label to be pushed away. Considering the discrepancies of recording domain are low-level features that scatter in the entire image, we extract domain-specific information from both the facial content and background context regions. Specifically, we want to highlight the context information from the surrounding background regions with purer domain information without disturbance of facial content. Consequently, we classify the output patches of $f_v$ into two types: context patches $c^{ctx}$ and content patches $c^{ctn}$. Then the Equation (2) can be rewritten as follows:

$$D(x) = PL([c_i^{ctx}, c_i^{ctn}]), \tag{6}$$

where $c_i^{ctx}$ denotes the context patch obtained from $f_{v,i}$. To enhance the representation of domain features while reducing the influence of facial content, we extract context patches $c_s^{ctx}$ from a shallower layer ($s < i$) of $f_v$ and replace the original $c_i^{ctx}$, as shown in Figure 2(green block). For the VIT visual encoder in CLIP, the context patches in shallower layers preserve low-level recording domain information as they have less interaction with content patches. Hence, the domain feature is formed as:

$$D(x) = PL([c_s^{ctx}, c_i^{ctn}]), \tag{7}$$

to emphasize the domain-specific information in the prompts.

**Adversarial Domain Prompt Learning**. We propose the Adversarial Domain Prompt Learning strategy to further ensure that the prompt learner can learn the domain-generalized information. A Discriminator $DC$ is constructed to differentiate the domain features and determine their belonging domain. $PL$, as the generator, is required to minimize the domain-specific trace in the output to fool $DC$:

$$\mathcal{L}_{adv} = -\frac{1}{N*M}\sum_{m=1}^{M}\sum_{n=1}^{N} y_m \log(DC(D(\boldsymbol{x}_n^m))), \tag{8}$$

$y_m$ is the domain label. $\mathcal{L}_{adv}$ encourages discriminator $DC$ to try its best to find the domain-specific vestige so that domain prompt learner can learn the domain-invariant information. Empirically, the adversarial learning strategy aggravates the unstable loss convergence. To ensure a more consistent learning direction, we employ a distribution constraint between domain features themselves. The Jensen–Shannon divergence is adopted to pull every two features from $M$ domains together in a training batch:

$$\mathcal{L}_{dis} = \sum_{j,t=1,\cdots,N*M,j<t} \mathrm{JS}(D(\boldsymbol{x}_j), D(\boldsymbol{x}_t)). \tag{9}$$

Here $N*M$ is the total number of samples in a batch.

### 4.2 Spoofing Domain Discriminative Prompt Tuning.

In contrast to the visual recording domain generalized prompt that aims to avoid bias, the attack types of spoofing domain can be utilized to boost the discriminability. Specific artifacts are highly correlated with the attack instruments, which can be regarded as spoofing clues, such as the saturation change in printing attacks, the moire pattern in video replay attacks, and the edge defects of 3D masks. Different from the low-level recording domain prompt, the attack types can be clearly described with the (high-level) textual prompt. Regularize the visual embedding with finer grained attack types text helps the visual encoder learn more discriminative FAS features. This will also set clearer boundaries between genuine faces and spoofing attacks thereby enhancing model generalizability to unseen attack types. Therefore, we propose the Spoof Domain Prompt tuning strategy to encourage the model to differentiate finer-grained attack types.

Specifically, we design a textual prompt template, a structured sentence that describes spoofing face photos with a placeholder of the attack types, such as "a {fake} face photo {and it is the printing attack}". Given $N*M$ samples in a training batch, we generate the same amount of text prompts $\boldsymbol{p}a$. Spoofing Domain Discriminative Prompt Tuning updates the parameters of $f_v$ with the text cross-entropy loss $\mathcal{L}_t$, which is calculated by:

$$\mathcal{L}_t = -\sum_{o=1}^{N*M} o \log(\frac{\exp(\langle f_v(\boldsymbol{x}), f_t(\boldsymbol{p}a_o)\rangle)}{\sum_{w=1}^{N*M}\exp(\langle f_v(\boldsymbol{x}), f_t(\boldsymbol{p}a_w)\rangle)}), \tag{10}$$

where $o$ and $w$ are the indices in $\boldsymbol{pa}$.

The final loss function of our method is formed as:

$$\mathcal{L} = \alpha\mathcal{L}_{ce} + \beta\mathcal{L}_t - \gamma\mathcal{L}_{adv} + \delta\mathcal{L}_{dis}. \tag{11}$$

where $\alpha, \beta, \gamma, \delta$ are hyperparameters that balance each loss.

## 5 Experiments

### 5.1 Experimental Setups

**Databases and Protocols.** We evaluate our method with six widely used FAS databases: MSU-MFSD(**M**) [58], CASIA-FASD(**C**) [67], Idiap Replay attack (**I**) [8], OULU-NPU(**O**) [2], WMCA(**W**) [10] and CelebA-Spoof [66]. Five protocols that cover various domain generalization settings with different amounts of source domains are involved to comprehensively evaluate our method's performance. They are leave-one-out(**P1**) and one-to-one(**P2**) cross-domain generalization evaluation, different domain partition evaluation(**P3**), unseen attack type evaluation(**P4**) and various devices domain evaluation(**P5**).

**Metrics.** Following [49], we evaluate the model performance using three metrics: Half Total Error Rate (HTER), Area Under the Receiver Operating Characteristic Curve (AUC) and True Positive Rate (TPR) at a False Positive Rate (FPR) 1% (TPR@FPR=1%). HTER and AUC reflect the theoretical performance. TPR@FPR=1% evaluates the model performance in practice.

**Implementation Details.** We utilize FAN [4] to obtain facial landmarks, aligning the images based on nose and eye landmarks before cropping them to a size of 224x224. We employ the pre-trained CLIP-ViT-B-16 [45] as the foundation model. An eleven-layer VIT following a two-layer MLP is adopted as the prompt learner $PL$. We extract content patches from the 5th layer of $f_v$ and context patches from the 1st layer. For adversarial learning an eight-layer 1D ResNet is adopted as the discriminator $DC$.

**Training strategy.** We divide the training process into two stages for the Bottom-Up domain prompt tuning. In the first training stage, we conduct the Spoofing Domain Discriminative Prompt Tuning to boost the discriminability of $f_v$. Adam optimizer is adopted with a learning rate of $10^{-6}$. In the second training stage, we freeze $f_v$ and train the $PL$ for the Recording Domain Generalization Prompt Learning. The optimizer for both $PL$ and $DC$ are Adam with learning rate $10^{-3}$ and $3 * 10^{-4}$. Before the adversarial learning, we first train at least one epoch for each domain source to warm up(exclude supplementary training data). Then, we take turns freezing the $PL$ and $DC$ for 50 iterations. For each stage we train 4000 iterations to ensure the largest dataset is traversed. The mini-batch size is 4 ($N = 4$) for each domain. $M$ varies according to the number of datasets used for training under different evaluation protocols. The hyperparameter $\alpha$, $\beta$, $\gamma$, and $\delta$ are set as 1, 1, 1, and 0.5, respectively.

**Table 1:** Evaluation of cross-domain performance under leave-one-out protocol (**P1**)

| Methods | 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% | |
| MADDG [47] | 17.69 | 88.06 | - | 24.50 | 84.51 | - | 22.19 | 84.99 | - | 27.98 | 80.02 | - | 23.09 |
| MDDR [53] | 17.02 | 90.10 | - | 19.68 | 87.43 | - | 20.87 | 86.72 | - | 25.025 | 81.47 | - | 20.64 |
| NAS-FAS [64] | 16.85 | 90.42 | - | 15.21 | 92.64 | - | 11.63 | 96.98 | - | 13.16 | 94.18 | - | 14.21 |
| RFMeta [48] | 13.89 | 93.98 | - | 20.27 | 88.16 | - | 17.30 | 90.48 | - | 16.45 | 91.16 | - | 16.97 |
| $D^2AM$ [7] | 12.70 | 95.6 | - | 20.98 | 85.58 | - | 15.43 | 91.22 | - | 15.27 | 90.87 | - | 16.09 |
| DRDG [28] | 12.43 | 95.81 | - | 19.05 | 88.79 | - | 15.56 | 91.79 | - | 15.63 | 91.75 | - | 15.66 |
| Self-DA [55] | 15.40 | 91.80 | - | 24.50 | 84.40 | - | 15.60 | 90.10 | - | 23.10 | 84.30 | - | 19.65 |
| ANRL [27] | 10.83 | 96.75 | - | 17.85 | 89.26 | - | 16.03 | 91.04 | - | 15.67 | 91.90 | - | 15.09 |
| FGHV [26] | 9.17 | 96.92 | - | 12.47 | 93.47 | - | 16.29 | 90.11 | - | 13.58 | 93.55 | - | 12.87 |
| SSDG-R [20] | 7.38 | 97.17 | - | 10.44 | 95.94 | - | 11.71 | 96.59 | - | 15.61 | 91.54 | - | 11.28 |
| SSAB-R [57] | 6.67 | 98.75 | - | 10.00 | 96.67 | - | 8.88 | 96.79 | - | 13.72 | 93.65 | - | 9.80 |
| PatchNet [51] | 7.10 | 98.46 | - | 11.33 | 94.58 | - | 13.40 | 95.67 | - | 11.82 | 95.07 | - | 10.90 |
| GDA [71] | 9.20 | 98.00 | - | 12.20 | 93.00 | - | 10.00 | 96.00 | - | 14.40 | 92.60 | - | 11.45 |
| LDCformer [16] | 6.43 | 98.39 | - | 8.11 | 96.67 | - | 8.57 | 97.09 | - | 11.17 | 95.58 | - | 8.57 |
| TTDG-V [70] | 4.16 | 98.48 | - | 7.59 | 98.18 | - | 9.62 | 98.18 | - | 10 | 96.15 | - | 7.84 |
| BUDoPT(ours) | **0.95** | **99.70** | **87.12** | **2.85** | **98.03** | **55.00** | **4.4** | **98.54** | **86.46** | **2.26** | **98.78** | **55.25** | **2.62** |
| ViT† [15] | 1.58 | 99.68 | 96.67 | 5.70 | 98.97 | 88.57 | 9.25 | 97.15 | 51.54 | 7.47 | 98.42 | 69.30 | 6.00 |
| FLIP-V† [49] | 3.79 | 99.31 | 87.99 | 1.27 | 99.75 | 95.85 | 4.71 | 98.80 | 75.84 | 4.15 | 98.76 | 66.47 | 3.48 |
| FLIP-IT† [49] | 5.27 | 98.42 | 79.33 | 0.44 | 99.98 | 99.86 | 2.94 | 99.42 | 84.62 | 3.61 | 99.15 | 84.76 | 3.06 |
| FLIP-MCL† [49] | 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 |
| BUDoPT-CoOp† | 1.98 | 99.54 | 80.63 | 1.93 | 99.58 | 88.06 | 3.18 | 99.3 | 90.4 | 3.33 | 98.91 | 89.29 | 2.61 |
| BUDoPT-CoCoOp† | 1.51 | 99.43 | 75.55 | 1.55 | 99.56 | 84.74 | 3.53 | 99.2 | 92.53 | 3.51 | 98.98 | 87.98 | 2.53 |
| BUDoPT(ours)† | **0.40** | **99.99** | **99.67** | **0.26** | 99.96 | 99.92 | **1.38** | **99.69** | **98.1** | **1.60** | 99.51 | **97.18** | **0.91** |

† indicates CelebA-Spoof [66] is employed as supplementary training dataset

### 5.2 Evaluation

**Cross-domain generalization evaluation**. Following [49], the leave-one-out (**P1**) and one-to-one (**P2**) on **M**, **C**, **I** and **O** are conducted to evaluate cross-dataset performance. OCM→I refers to the model trains on **O**, **C**, **M** and tests on **I**. For a fair comparison with [49] and other models, we evaluate the performance of our method with and without the supplementary training data CelebA-Spoof [66]. For simplicity, we regard dataset as the domain partition in our proposed method. As shown in Table 1, without supplementary training data, the proposed BUDoPT outperforms SOTA in all settings. With supplementary training data, BUDoPT also surpasses the FLIP family model in HTER for all settings and achieves more than 69% improvement on average. In terms of TPR@FPR=1%, our method performs better in three out of four settings. Although the performance slightly drops in TPR@FPR=1%(C) and AUC(C,O), the margin is relatively small as they already exceed 99%.

In **P2** of low-data regime [49] the model is trained separately on **M**, **C**, **I** and **O** on the remaining datasets, resulting in a total of 12 different scenario combinations. Since the model is trained on one dataset only, domain partition is based on capturing devices(**C**,**M**), environments(**I**), and sessions(**O**) for different datasets. We also test our method with and without supplementary CelebA-Spoof. CelebA-Spoof is regarded as a domain if it is involved. As illustrated in Table 2, without help of CelebA-Spoof, BUDoPT performs better than SOTA in nine out of twelve settings. We also outperform domain adap-

**Table 2:** Evaluation of cross-domain performance under one-to-one protocol (**P2**)

| Methods | Unseen | 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 [50] | No | 41.8 | 36.6 | - | 49.8 | 35.1 | - | 39.0 | 35.2 | - | - | - | - | 39.6 |
| DRCN [11] | No | 44.4 | 27.4 | - | 48.9 | 42.0 | - | 28.9 | 36.8 | - | - | - | - | 38.1 |
| DupGAN [14] | No | 42.1 | 33.4 | - | 46.5 | 36.2 | - | 27.1 | 35.4 | - | - | - | - | 36.8 |
| KSA [23] | No | 39.3 | 15.1 | - | 12.3 | 33.3 | - | 9.1 | 34.9 | - | - | - | - | 24.0 |
| DR-UDA [54] | No | 15.6 | 9.0 | 28.7 | 34.2 | 29.0 | **3.5** | 16.8 | 3.0 | 30.2 | 19.5 | 25.4 | 27.4 | 23.1 |
| ADA [52] | No | 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 [21] | No | 16.0 | 9.2 | - | 30.2 | 25.8 | - | 13.3 | 3.4 | - | - | - | - | 16.3 |
| GDA [71] | No | 15.10 | 5.8 | - | 29.7 | 20.8 | - | 12.2 | 2.5 | - | - | - | - | 14.4 |
| CDFTN-L [65] | No | **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 |
| BUDoPT(ours) | Yes | 9.65 | **4.37** | **6.76** | **8.26** | **5.64** | 7.73 | **6.07** | 2.9 | **6.16** | **4.81** | **2.97** | **2.94** | **5.69** |
| FLIP-V† [49] | Yes | 15.08 | 13.73 | 12.34 | 4.30 | 9.68 | 7.87 | 0.56 | 3.96 | 4.79 | 2.09 | 5.01 | 6.00 | 7.12 |
| FLIP-IT† [49] | Yes | 12.33 | 15.18 | 7.98 | 1.12 | 8.37 | 6.98 | 0.19 | 5.21 | 4.96 | **0.16** | 4.27 | 5.63 | 6.03 |
| FLIP-MCL† [49] | Yes | 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 |
| BUDoPT(ours)† | Yes | **4.33** | **2.62** | 4.98 | **0.48** | **1.83** | **4.14** | **0** | **2.45** | **1.87** | 0.44 | **2.53** | **1.43** | **2.46** |

† indicates CelebA-Spoof is employed as supplementary training data

tive models (Unseen=No) even without any information of target domain. With CelebA-Spoof, BUDoPT surpasses FLIP family models in ten settings.

**Effect of Domain Partition**. We set MI→O and MI→C to evaluate our model with different amounts of training data sources. Here we treat dataset as domain. Results in Table 3 show that with less training data, BUDoPT outperforms competitors by a larger margin. However, we also note that comparing with the M→O results in **P2** the performance increase is limited. We believe this is due to the difference in domain partition strategies. For M→O the domain is defined based on the device while for MI→O it is based on dataset. The results indicate that a fine-grained partition strategy may help achieve better performance for our method even with a limited data scale. More analysis of partition strategy is in supplementary Sec 2.1.

**Generalize to Unseen attack type**. We employ **W** with high-quality 3D silicon mask attack samples to evaluate the model generalizability to unseen attack types (**P4**). Different from the baselines trained on **IMCO**, we only utilize **C** as the source dataset. Same domain partition strategy in **P2** is adopted. As shown in Table 4, with less training data BUDoPT still outperforms the others. This result not only demonstrates our generalizability to the unseen attack types but also shows the learning efficiency of our method.

**Device Variance Only**. In **P5**, we utilize the **O**'s intra protocol 3 and protocol 4 [2] to evaluate the performance with device variance only. The domain gaps in **P5** are located on the unseen device settings and parameters and other underlying factors are similar. Same domain partition strategy in **P2** is adopted. As shown in Table 5, under constraint setting BUDoPT can handle the domain gap of devices well. This indicates the usability in real applications where some local samples can be available for fine-tuning.

**Table 3:** Evaluation of different domain partition (**P3**)

| Methods | MI → C | | MI → O | |
|---|---|---|---|---|
| | HTER | AUC | HTER | AUC |
| SSAN-R [57] | 25.56 | 83.89 | 24.44 | 82.86 |
| HFN+MP [6] | 30.89 | 72.48 | 20.94 | 86.71 |
| CIFAS [35] | 22.67 | 83.89 | 24.63 | 81.48 |
| DiVT-M [24] | 20.11 | 86.71 | 23.61 | 85.73 |
| SSDG-R [20] | 19.86 | 86.46 | 27.92 | 78.72 |
| DGUA-FAS [13] | 19.22 | 86.81 | 20.05 | 88.75 |
| BUDoPT(ours) | **5.33** | **98.92** | **5.94** | **98.37** |

**Table 4:** Evaluation of unseen attack type performance (**P4**)

| Methods | IMCO → WMCA | |
|---|---|---|
| | HTER | AUC |
| DiVT-M [24] | 22.36 | 86.82 |
| DGUA-FAS [13] | 20.62 | 88.07 |
| | C → WMCA | |
| BUDoPT(ours) | **9.20** | **95.93** |

**Table 5:** Evaluation of Device Variance Only (**P5**) on **O**. P refers protocol

| Methods | P | ACPER | BCPER | ACER |
|---|---|---|---|---|
| DC-CDN [63] | 3 | 2.2±2.8 | 1.6±2.1 | 1.9±1.1 |
| RAEDFL [17] | 3 | 1.38±1.78 | 0.28±0.68 | 0.83±0.86 |
| TranFAS [56] | 3 | 0.6±0.7 | 1.1±2.5 | 0.9±1.1 |
| LDCFormer [16] | 3 | 2.35±2.05 | 0.28±0.68 | 1.31±1.03 |
| BUDoPT(ours) | 3 | **0±0** | **0±0** | **0±0** |
| DC-CDN [63] | 4 | 5.4±3.3 | 2.5±4.2 | 4.0±3.1 |
| RAEDFL [17] | 4 | 5.41±6.40 | 2.50±2.74 | 3.96±3.90 |
| TranFAS [56] | 4 | 2.1±2.2 | 3.8±3.5 | 2.9±2.4 |
| LDCFormer [16] | 4 | 1.08±1.28 | 1.17±1.94 | 1.13±1.02 |
| BUDoPT(ours) | 4 | **0±0** | **0±0** | **0±0** |

**Table 6:** Ablation study of major components of the proposed BUDoPT

| SD | SDD | RDG | AD | CDC | ICM → O |
|---|---|---|---|---|---|
| ✓ | - | - | - | - | 3.51 |
| - | ✓ | - | - | - | 5.78 |
| ✓ | - | ✓ | - | - | 3.01 |
| ✓ | - | ✓ | ✓ | - | 2.50 |
| ✓ | - | ✓ | ✓ | ✓ | 2.26 |

### 5.3 Ablation Studies

We analyze the effectiveness of major components(Table 6), test the intuitive textual recording domain prompt tuning(Table 6), and compared with generic visual prompt tuining CoOp [69], CoCoOp [68](Table 1).

Table 6 reports the results of the ablation studies of our major components. We set Spoofing Domain Discriminative Prompt Tuning(SD) as the baseline and evaluate performance with and without Recording Domain Generalization Prompt Learning (RDG), Adversarial Domain Learning(AD) and Context-aware Domain Clue(CDC). With Recording Domain Generalization Prompt Learning, the HTER is improved by 14%. With Adversarial Domain Learning, the HTER decreases by around 17%. The Context-aware Domain Clue contributes around 10% improvement. The ablation studies indicate that each component within our method is crucial for improving the domain generalization ability.

**Intuitive textual recording domain prompt tuning.** Considering the success of employing context information as a prompt in upper-stream tasks, we evaluate an intuitive domain prompt tuning strategy, i.e., using the textual domain description as the prompt. We mark the textual domain prompts with the

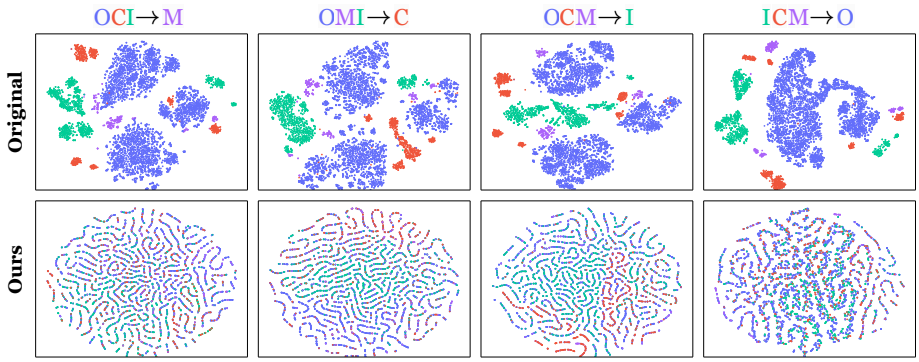



**Fig. 4:** The t-SNE visualizations of original visual embedding (first row) and the learned domain-invariant prompt embedding (second row) under **P1**.

Spoofing Domain Discriminative Prompt Tuning as SDD [2]. As shown in Table 6, adding the intuitive textual recording domain prompt (SDD) harms the model's performance. This aligns with our analysis that the textual domain prompt of context encourages the model to extract more domain-specific information and contradicts the goal of domain-generalized FAS.

**Generic visual prompt tuning.** Since the proposed method follows the idea of visual prompt tuning, the classical visual prompt learning methods: CoOp and CoCoOp are compared. For a fair comparison, the CLIP visual encoder of the two models is finetuned with the Spoofing Domain Discriminative Prompt Learning module in our framework so we regard them as BUDoPT-CoOp and BUDoPT-CoCoOp. As the results shown in Table 1, by comparing with the FLIP-family models, BUDoPT-CoOp, and BUDoPT-CoCoOp, we comprehensively verify the effectiveness of our proposed method that adapts CLIP to FAS with better generalization ability.

We also conduct hyperparameter analysis of loss function. Results are summarized in Eq. (11) in the supplementary Section 2.2.

### 5.4 Visualizations

**Domain Prompt Visualization.** Figure 4 shows the t-SNE [36] visualization of original visual embedding (first row) and the learned domain-invariant prompt embedding (second row) under **P1**. The original visual embedding naturally forms multiple clusters of domain-specific information. Each cluster shares some low-level underlying factors, such as the camera qualities (see supplementary material for details). Our BUDoPT learns a domain-generalized prompt embedding where samples from different domains are grouped together in a single cluster, thereby guiding the model to minimize the influences of domain gaps.

**Attention Visualization.** We utilize the GradCam [46] to visualize attention

[2] Please refer to supplementary material for details

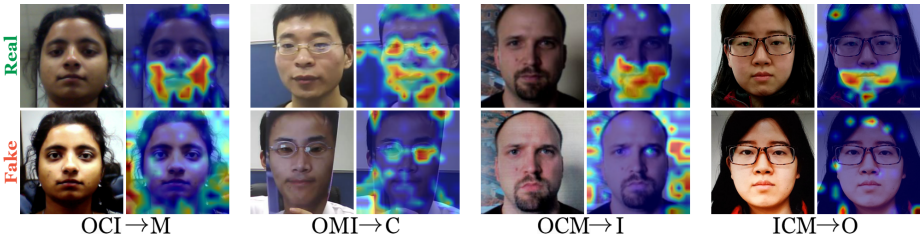



**Fig. 5:** Typical attention maps of real and fake samples under **P1**.

maps for genuine and spoofing samples in **P1**. Typical results are visualized in Figure 5. We can observe that BUDoPT takes different attention for live and spoof faces. Specifically, the model focuses on face regions for real samples while highlights spoof clues of fake faces. With the help of Spoofing Domain Discriminative Prompt Tuning, the model is able to detect the intrinsic spoofing clues for different types of attack, including moire patterns(**M**, **O**), paper mask edges(**C**) and paper texture(**I**).

## 6 Conclusion

In this work, we aim to improve the generalization ability of CLIP-based FAS from a prompt learning perspective. A Bottom-Up Domain Prompt Tuning method (BUDoPT) that covers the different levels of domain variance, including the domain of recording settings and domain of attack types is proposed. It contains a context-aware adversarial domain-generalized prompt learning strategy that can learn domain-invariant prompt for different recording settings and a fine-grained textual prompt that guides CLIP to look through the subtle details of different attack instruments. Extensive experiments are conducted on five FAS datasets with a large number of variations (camera types, resolutions, image qualities, lighting conditions, and recording environments). The effectiveness of our proposed method is evaluated with different amounts of source domains from multiple angles, where we boost the generalizability compared with the state of the arts with multiple training datasets or with only one dataset. We also observe the limitations of our methods with different domain partitioning strategies. Designing an adaptive domain partition approach is a potential way to further improve the generalizability of our method.

## 7 Acknowledgement

This work is supported by GuangDong Basic and Applied Basic Research Foundation 2023A1515110706, RGC-NSFC joint project N_HKBU212/23, and NSFC Young Scientist Fund Application No. 6240075068.

## References


1. Bao, W., Li, H., Li, N., Jiang, W.: A liveness detection method for face recognition based on optical flow field. In: IASP (2009)
2. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: OULU-NPU: A mobile face presentation attack database with real-world variations. In: FG (2017)
3. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face spoofing detection using colour texture analysis. IEEE Trans. on Information Forensics and Security **11**(8), 1818–1830 (2016)
4. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: International Conference on Computer Vision (2017)
5. 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. pp. 8037–8048 (2023)
6. Cai, R., Li, Z., Wan, R., Li, H., Hu, Y., Kot, A.C.: Learning meta pattern for face anti-spoofing. IEEE Transactions on Information Forensics and Security **17**, 1201–1213 (2022)
7. 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. vol. 35, pp. 1132–1139 (2021)
8. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: BIOSIG. pp. 1–7 (2012)
9. Ge, H., Tu, X., Ai, W., Luo, Y., Ma, Z., Xie, M.: Face anti-spoofing by the enhancement of temporal motion. In: CTISC. pp. 106–111. IEEE (2020)
10. George, A., Mostaani, Z., Geissenbuhler, D., Nikisins, O., Anjos, A., Marcel, S.: Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE Trans. on Information Forensics and Security (2019)
11. Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W.: Deep reconstruction-classification networks for unsupervised domain adaptation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 597–613. Springer (2016)
12. Guo, J., Zhu, X., Xiao, J., Lei, Z., Wan, G., Li, S.Z.: Improving face anti-spoofing by 3d virtual synthesis. arXiv (2019)
13. Hong, Z.W., Lin, Y.C., Liu, H.T., Yeh, Y.R., Chen, C.S.: Domain-generalized face anti-spoofing with unknown attacks. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 820–824. IEEE (2023)
14. 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. pp. 1498–1507 (2018)
15. 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. pp. 37–54. Springer (2022)
16. Huang, P.K., Chiang, C.H., Chong, J.X., Chen, T.H., Ni, H.Y., Hsu, C.T.: Ldcformer: Incorporating learnable descriptive convolution to vision transformer for face anti-spoofing. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 121–125. IEEE (2023)
17. Huang, P.K., Chin, M.C., Hsu, C.T.: Face anti-spoofing via robust auxiliary estimation and discriminative feature learning. In: Asian Conference on Pattern Recognition. pp. 443–458. Springer (2021)

18. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
19. Jee, H.K., Jung, S.U., Yoo, J.H.: Liveness detection for embedded face recognition system. International Journal of Biological and Medical Sciences **1**(4), 235–238 (2006)
20. 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. pp. 8484–8493 (2020)
21. Jia, Y., Zhang, J., Shan, S., Chen, X.: Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing. Pattern Recognition **115**, 107888 (2021)
22. Jourabloo, A., Liu, Y., Liu, X.: Face de-spoofing: Anti-spoofing via noise modeling. In: ECCV. pp. 290–306 (2018)
23. 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 **13**(7), 1794–1809 (2018)
24. 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. pp. 6098–6107 (2023)
25. Liu, A., Xue, S., Gan, J., Wan, J., Liang, Y., Deng, J., Escalera, S., Lei, Z.: Cfpl-fas: Class free prompt learning for generalizable face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 222–232 (2024)
26. 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. vol. 36, pp. 1782–1791 (2022)
27. 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: Proceedings of the 29th ACM international conference on multimedia. pp. 1469–1477 (2021)
28. 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)
29. Liu, S.Q., Lan, X., Yuen, P.C.: Multi-channel remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. IEEE Trans. on Information Forensics and Security **16**, 2683–2696 (2021)
30. Liu, S.Q., Lan, X., Yuen, P.C.: Learning temporal similarity of remote photoplethysmography for fast 3d mask face presentation attack detection. IEEE Transactions on Information Forensics and Security **17**, 3195–3210 (2022)
31. Liu, Y., Tai, Y., Li, J., Ding, S., Wang, C., Huang, F., Li, D., Qi, W., Ji, R.: Aurora guard: Real-time face anti-spoofing via light reflection. arXiv (2019)
32. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: CVPR. pp. 389–398 (2018)
33. Liu, Y., Liu, X.: Spoof trace disentanglement for generic face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3813–3830 (2022)
34. Liu, Y., Stehouwer, J., Jourabloo, A., Liu, X.: Deep tree learning for zero-shot face anti-spoofing. In: CVPR (2019)

35. Liu, Y., Chen, Y., Dai, W., Li, C., Zou, J., Xiong, H.: Causal intervention for generalizable face anti-spoofing. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 01–06. IEEE (2022)
36. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
37. Määttä, J., Hadid, A., Pietikäinen, M.: Face spoofing detection from single images using micro-texture analysis. In: IJCB. pp. 1–7 (2011)
38. Menotti, D., Chiachia, G., Pinto, A., Schwartz, W.R., Pedrini, H., Falcão, A.X., Rocha, A.: Deep representations for iris, face, and fingerprint spoofing detection. IEEE Trans. on Information Forensics and Security **10**(4), 864–879 (2015)
39. Muhammad, U., Holmberg, T., de Melo, W.C., Hadid, A.: Face anti-spoofing via sample learning based recurrent neural network (rnn). In: BMVC. p. 113 (2019)
40. Pan, G., Sun, L., Wu, Z., Lao, S.: Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In: ICCV. pp. 1–8 (2007)
41. Patel, K., Han, H., Jain, A.K.: Secure face unlock: Spoof detection on smartphones. IEEE transactions on information forensics and security **11**(10), 2268–2283 (2016)
42. 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: ICB. pp. 98–105 (2015)
43. Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.H., Riedel, S.: Language models as knowledge bases? arXiv preprint arXiv:1909.01066 (2019)
44. Qin, Y., Yu, Z., Yan, L., Wang, Z., Zhao, C., Lei, Z.: Meta-teacher for face anti-spoofing. IEEE transactions on pattern analysis and machine intelligence **44**(10), 6311–6326 (2021)
45. 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. pp. 8748–8763. PMLR (2021)
46. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
47. 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. pp. 10023–10031 (2019)
48. Shao, R., Lan, X., Yuen, P.C.: Regularized fine-grained meta face anti-spoofing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11974–11981 (2020)
49. 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. pp. 19685–19696 (2023)
50. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7167–7176 (2017)
51. 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. pp. 20281–20290 (2022)
52. Wang, G., Han, H., Shan, S., Chen, X.: Improving cross-database face presentation attack detection via adversarial domain adaptation. In: 2019 International Conference on Biometrics (ICB). pp. 1–8. IEEE (2019)

53. 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. pp. 6678–6687 (2020)
54. 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 **16**, 56–69 (2020)
55. 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. vol. 35, pp. 2746–2754 (2021)
56. Wang, Z., Wang, Q., Deng, W., Guo, G.: Face anti-spoofing using transformers with relation-aware mechanism. IEEE Transactions on Biometrics, Behavior, and Identity Science **4**(3), 439–450 (2022)
57. 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. pp. 4123–4133 (2022)
58. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Trans. on Information Forensics and Security **10**(4), 746–761 (2015)
59. Yang, J., Lei, Z., Li, S.Z.: Learn convolutional neural network for face anti-spoofing. arXiv preprint arXiv:1408.5601 (2014)
60. Yang, J., Lei, Z., Liao, S., Li, S.Z.: Face liveness detection with component dependent descriptor. In: 2013 International Conference on Biometrics (ICB). pp. 1–6. IEEE (2013)
61. Yu, Z., Li, X., Niu, X., Shi, J., Zhao, G.: Face anti-spoofing with human material perception. In: ECCV. pp. 557–575. Springer (2020)
62. Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z., Zhao, G.: Deep learning for face anti-spoofing: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(5), 5609–5631 (2022)
63. Yu, Z., Qin, Y., Zhao, H., Li, X., Zhao, G.: Dual-cross central difference network for face anti-spoofing. arXiv preprint arXiv:2105.01290 (2021)
64. 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 Trans. on Pattern Analysis and Machine Intelligence (2020)
65. Yue, H., Wang, K., Zhang, G., Feng, H., Han, J., Ding, E., Wang, J.: Cyclically disentangled feature translation for face anti-spoofing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3358–3366 (2023)
66. 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) (2020)
67. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: ICB (2012)
68. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
69. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
70. 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. pp. 175–187 (2024)

71. 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. pp. 335–356. Springer (2022)