FLIP: Cross-domain Face Anti-spoofing with Language Guidance

Koushik Srivatsan Muzammal Naseer Karthik Nandakumar
Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI)
Abu Dhabi, United Arab Emirates
{koushik.srivatsan, muzammal.naseer, karthik.nandakumar}@mbzuai.ac.ae

Abstract

Face anti-spoofing (FAS) or presentation attack detection is an essential component of face recognition systems deployed in security-critical applications. Existing FAS methods have poor generalizability to unseen spoof types, camera sensors, and environmental conditions. Recently, vision transformer (ViT) models have been shown to be effective for the FAS task due to their ability to capture long-range dependencies among image patches. However, adaptive modules or auxiliary loss functions are often required to adapt pre-trained ViT weights learned on large-scale datasets such as ImageNet. In this work, we first show that initializing ViTs with multimodal (e.g., CLIP) pre-trained weights improves generalizability for the FAS task, which is in line with the zero-shot transfer capabilities of vision-language pre-trained (VLP) models. We then propose a novel approach for robust cross-domain FAS by grounding visual representations with the help of natural language. Specifically, we show that aligning the image representation with an ensemble of class descriptions (based on natural language semantics) improves FAS generalizability in low-data regimes. Finally, we propose a multimodal contrastive learning strategy to boost feature generalization further and bridge the gap between source and target domains. Extensive experiments on three standard protocols demonstrate that our method significantly outperforms the state-of-the-art methods, achieving better zero-shot transfer performance than five-shot transfer of "adaptive ViTs". Code: https://github.com/koushiksrivats/FLIP

1. Introduction

From personal devices to airport boarding gates, face recognition systems have become a ubiquitous tool for recognizing people. This may be attributed to recent advances in face recognition technology based on deep learning, as well as its simplicity and non-contact nature. However, these systems are vulnerable to face presentation attacks, where an attacker tries to spoof the identity of a bonafide individual with the help of presentation attack instruments (PAI) such as printed photos, replayed videos, or 3D synthetics masks [52]. Therefore, face anti-spoofing (FAS) or face presentation attack detection (FPAD) is essential to secure face recognition systems against presentation attacks.

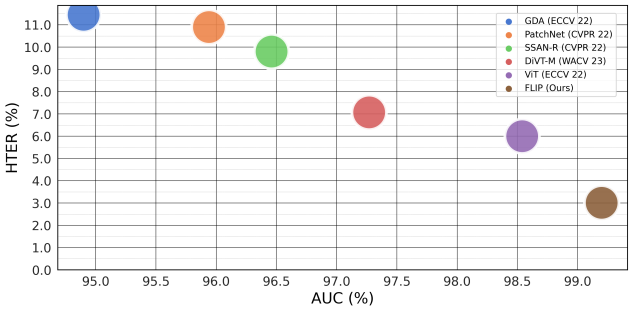

Figure 1. Area Under ROC Curve (AUC %) and Half Total Error Rate (HTER %) comparison between our proposed method and state-of-the-art (SOTA). Our method achieves the highest AUC (↑) performance with the lowest HTER (↓) for cross-domain face anti-spoofing on MCIO datasets, surpassing all the SOTA methods.

Prior works [59, 30, 51, 47, 54, 53, 42] have shown that impressive FAS accuracy can be achieved in intra-domain scenarios, where the training and test distributions are similar. However, existing FAS methods fail to generalize well to the unseen target domains due to two main reasons: (a) variations due to camera sensors, presentation attack instruments, illumination changes, and image resolution cause a large domain gap between the source and target distributions that is inherently hard to bridge; and (b) commonly used FAS benchmark datasets have limited training data, causing the model to overfit to the source domain(s). Consequently, achieving robust cross-domain FAS performance has remained an elusive challenge thus far.

The problem of cross-domain FAS has been formulated in different ways in the literature. Unsupervised domain adaptation (UDA) methods [40, 12, 15, 21, 45, 44, 43, 19, 67, 56] make use of the unlabeled target domain data and labeled source domain data to learn a generalized decision boundary. Few-shot learning methods [29, 32, 31, 16]

use a small subset of labeled target domain data during training to learn features that adapt well to the target domain. However, both these methods assume access to the target domain either in the form of a large set of unlabeled samples or a few labeled samples, which may not always be available. Domain generalization (DG) methods [38, 39, 6, 28, 46, 27, 26, 18, 48, 63, 23] propose to learn domain-agnostic discriminative features from multiple source domains that generalize to an unseen target domain. While zero-shot learning and DG settings are more challenging, they are more applicable in practice. 就是本篇

Recent works [10, 16, 23] have established the effectiveness of vision transformers (ViT) for cross-domain FAS. Since ViTs [9] split the image into fixed-size patches and have the ability to capture long-range dependencies among these patches, they can independently detect the local spoof patterns and aggregate them globally to make an informed decision. However, these methods have two limitations. Firstly, these ViTs are learned using only image data and their learning is guided only by the corresponding image labels, which might not be representative enough. This limits their generalization ability, especially when presented with limited training data. Secondly, they typically require adaptive modules, additional domain labels, or attack-type information to finetune pre-trained weights. This requires explicit network modifications or custom curation of additional information such as attack type or domain labels.

While multimodal vision-language pre-trained (VLP) models have achieved striking zero-shot performance and good generalization in some applications [60, 66, 13, 36, 68, 20, 35], there is still a debate on whether incorporating language supervision yields vision models with more generalizable representations [8, 37]. Therefore, the objective of this work is to examine the following questions: (i) Can initialization of ViTs using multimodal pre-trained weights lead to better cross-domain FAS performance compared to ViTs pre-trained only on images?; (ii) Besides leveraging the image encoder of a VLP model, can the text encoder also be utilized to improve the FAS generalization performance?; and (iii) Can the large domain gap and limited training data availability in FAS be surmounted by exploiting self-supervision techniques during the adaptation of VLP models for the FAS task? The main contributions of this work are as follows:

- We show that direct finetuning of a multimodal pre-trained ViT (e.g., CLIP image encoder) achieves better FAS generalizability without any bells and whistles.
- We propose a new approach for robust cross-domain FAS by grounding the visual representation using natural language semantics. This is realized by aligning the image representation with an ensemble of text prompts (describing the class) during finetuning.
- We propose a multimodal contrastive learning strategy, which enforces the model to learn more generalized features that bridge the FAS domain gap even with limited training data. This strategy leverages view-based image self-supervision and view-based cross-modal image-text similarity as additional constraints during the learning process.

2. Related Work

Domain Adaptation and Few-shot Learning: Several methods have been proposed to leverage unlabeled data from the target domain along with labeled source data. One approach is to align the source and target feature distributions either by reducing the Maximum Mean Discrepancy [21] or by using adversarial domain adaptation [43]. Other methods use semi-supervised learning [19] and progressive transfer learning strategies [33] to exploit the availability of a few labeled samples from the target domain. In [22], a FAS model trained with sufficient labeled training data is distilled to application-specific domains for which training samples are scarce. In [67], cross-domain FAS is treated as a style transfer problem, where target data is transformed to the source domain style via image translation. Vision transformers with ensemble adapter modules and feature-wise transformation layers are employed in [16] for adapting to the target domain. Pseudo-labeled samples containing domain-invariant liveness features from the source domain and content features from the target domain are generated in [56] and both these features are disentangled through domain adversarial training. However, all the above methods assume access to the unlabeled/labeled target domain data, which may not always be available.

Domain Generalization: The idea of learning a shared generalized feature space for FAS was first proposed in [38], where a multi-adversarial discriminative domain generalization framework was presented. A fine-grained meta-learning-based approach was proposed in [39] by simulating the domain shift during training. The concept of separating the features into style and content components to create a stylized feature space was introduced in [48], upon which a contrastive learning strategy is applied emphasizing on liveness-related style information to learn a generalized representation. Recently, vision transformers with two additional losses were used in [23], where one loss enforces the real data from multiple domains to be compact and the other enforces a domain-invariant attack type separation. Though these methods demonstrate promising cross-domain performance, they still require additional information such as attack types and domain labels, or make use of non-trivial auxiliary supervision.

Vision Language Pre-training: Vision-language pre-

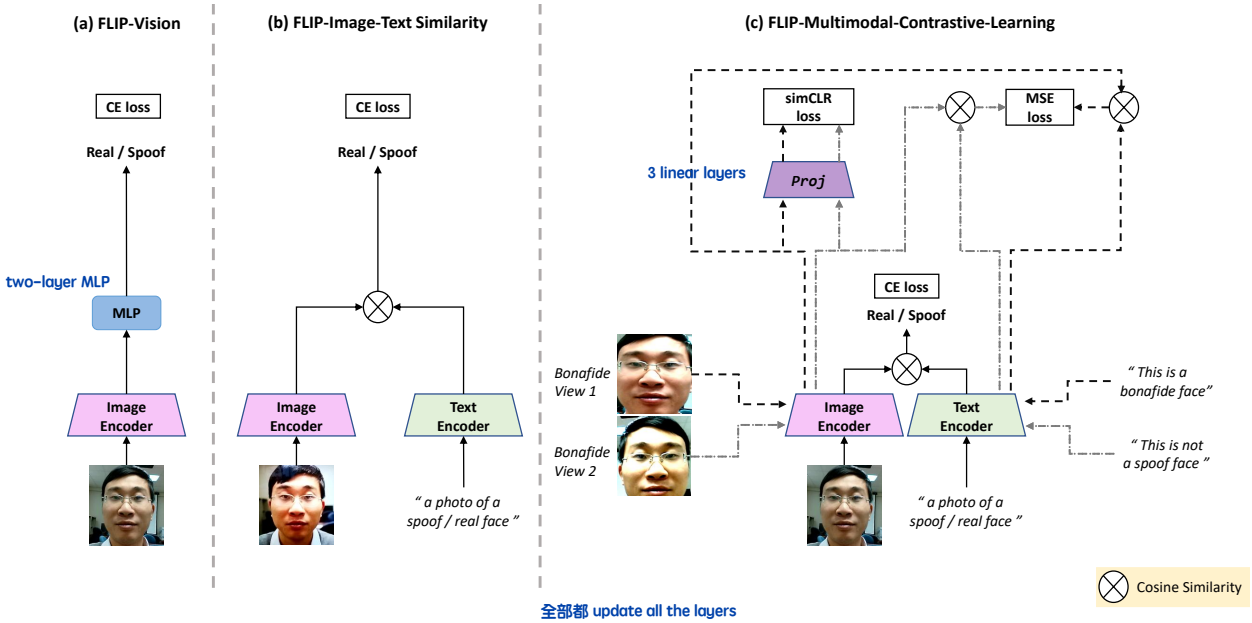

Figure 2. Overview of the proposed FLIP framework for cross-domain face anti-spoofing.

trained (VLP) models encode rich multimodal representations and have demonstrated excellent generalization performance on various downstream applications [60, 66, 13, 36, 68, 20, 35]. Riding on the success of transformer models [41, 9], contrastive representation learning [5, 14], and web-scale training datasets [17, 34], several VLP models have been proposed recently to learn joint image-text representations [34, 17, 58, 50, 55]. However, the issue of whether language supervision enhances the generalizability of vision models is still being debated [8, 37]. In this work, we use contrastive language-image pre-training (CLIP) [34] as the base VLP model.

3. Proposed Method

The goal of cross-domain FAS is to achieve high presentation attack detection accuracy on out-of-distribution face datasets containing bonafide images and presentation attacks. In the many-to-one DG setting, the model is learned from a set of N different source domain datasets $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_N\}$ and evaluated on a single target domain dataset $\mathcal{T}$. In the one-to-one DG setting, the model is trained on images from a single source domain $\mathcal{S}_i$ to generalize to the target domain. Let $I^r_{\mathcal{D}}$ denote a real (bonafide) face image from domain $\mathcal{D} \in (\mathcal{S} \cup \mathcal{T})$. Similarly, let $I^s_{\mathcal{D}}$ represent a spoof (presentation attack) image from $\mathcal{D}$.

We propose a framework called **F**ace Anti-Spoofing with **L**anguage-**I**mage **P**retraining (FLIP) for cross-domain FAS (see Figure 2). The proposed framework uses CLIP [34] as the base model and is finetuned using different strategies to obtain three variants: FLIP-Vision (FLIP-V), FLIP-Image-Text Similarity (FLIP-IT), and FLIP-Multimodal-Contrastive-Learning (FLIP-MCL). We first outline the working of the base model before describing the variants.

3.1. Contrastive Language-Image Pre-Training

CLIP [34] is trained using millions of image-text pairs sourced from the internet. CLIP encodes the input image $I \in \mathbb{R}^{H \times W \times 3}$ and the corresponding text description t into a shared embedding space as detailed below.

Image Encoder: The image encoder is a vision transformer $\mathcal{V}$ consisting of K transformer blocks $\{\mathcal{V}_k\}_{k=1}^{K}$. To encode the input image I, it is first split into M fixed-size patches and these patches are projected linearly into patch embeddings $\boldsymbol{e}_0 \in \mathbb{R}^{M \times d_v}$. Patch embeddings $\boldsymbol{e}_{k-1}$ are then input to the k^{th} transformer block $(\mathcal{V}_k)$ after appending a learnable class token c_{k-1}, and processed through the K transformer blocks sequentially.

$$[\mathrm{c}_k, \boldsymbol{e}_k] = \mathcal{V}_k([\mathrm{c}_{k-1}, \boldsymbol{e}_{k-1}]) \qquad k = 1, 2, \cdots, K.$$

The final image representation $\boldsymbol{x}$ is obtained by linearly projecting the class token c_K from the last transformer block $(\mathcal{V}_K)$ into a shared vision-language space via `ImageProj`:

$$\boldsymbol{x} = \texttt{ImageProj}(\mathrm{c}_K) \qquad \boldsymbol{x} \in \mathbb{R}^{d_{vl}}.$$

Text Encoder: The text encoder $\mathcal{L}$ generates feature representations for the description t by first tokenizing the words and then projecting them into word embeddings $\boldsymbol{w}_0 = [w_0^1, w_0^2, \cdots, w_0^Q] \in \mathbb{R}^{Q \times d_l}$. At each stage, $\boldsymbol{w}_{k-1}$ is input to the k^{th} transformer block $(\mathcal{L}_k)$ to obtain

$$\boldsymbol{w}_k = \mathcal{L}_k(\boldsymbol{w}_{k-1}) \qquad k = 1, 2, \cdots, K.$$

The final text representation $\boldsymbol{z}$ is obtained by projecting the text embeddings corresponding to the last token of the last transformer block ($\mathcal{L}_K$) into a shared vision-language latent space via $\texttt{TextProj}$.

$$\boldsymbol{z} = \texttt{TextProj}(\boldsymbol{w}_K^Q) \qquad \boldsymbol{z} \in \mathbb{R}^{d_{vl}}.$$

The CLIP model has been pre-trained using a contrastive loss that maximizes the cosine similarity of the image ($\boldsymbol{x}$) and text ($\boldsymbol{z}$) embeddings of n corresponding (image, text) pairs in a batch while minimizing the cosine similarity of the embeddings of the $(n^2 - n)$ incorrect pairings.

3.2. FLIP-Vision

Representations produced by CLIP have shown impressive out-of-the-box performance for many downstream vision applications based on natural images such as classification [60, 66], object detection [13, 36, 68], and segmentation [20, 35]. However, these features cannot be directly used for the FAS task, which requires identifying subtle variations among similar face images. Hence, we first fine-tune only the vision backbone for FAS and refer to this approach as FLIP-Vision (FLIP-V). In this method, we take a pre-trained CLIP model and use only its image encoder $\mathcal{V}$ and discard the text encoder $\mathcal{L}$. This gives us a simple ViT initialized with language-image pre-trained weights. Given a batch of balanced images from N source domains, we use the image encoder to extract the class token ($\texttt{c}_K$) from the last transformer block ($\mathcal{V}_K$) prior to $\texttt{ImageProj}$. This class token is then passed to a multi-layer perceptron (MLP) classification head, to decide if the input image is spoof or real. The image encoder and the MLP head are updated using the standard cross entropy loss L_{ce}.

Prompt No.	Real Prompts	Spoof Prompts
P1	This is an example of a real face	This is an example of a spoof face
P2	This is a bonafide face	This is an example of an attack face
P3	This is a real face	This is not a real face
P4	This is how a real face looks like	This is how a spoof face looks like
P5	A photo of a real face	A photo of a spoof face
P6	This is not a spoof face	A printout shown to be a spoof face

Table 1. Natural language descriptions (context prompts) of the real and spoof classes used to guide the FLIP-IT model.

3.3. FLIP-Image-Text Similarity

In **FLIP**-**I**mage-**T**ext similarity, we obtain the prediction with the help of language supervision instead of using the MLP head. Specifically, we leverage textual prompts/descriptions corresponding to the real and spoof classes (denoted as t_r and t_s, respectively), whose feature representations are computed using the text encoder $\mathcal{L}$. The cosine similarity between the image representation ($\boldsymbol{x}$) and text representations corresponding to the two classes ($\boldsymbol{z}_r$ and $\boldsymbol{z}_s$) is computed, resulting in two values for every image in the batch. These similarity values are considered as class logits and passed to the cross entropy loss computation.

During inference, the predicted class $\hat{y}$ is determined by the class description having the highest cosine similarity score ($sim(\cdot,\cdot)$) with the given image I. Hence,

$$p(\hat{y}|x) = \frac{\exp(sim(\boldsymbol{x}, \boldsymbol{z}_{\hat{y}})/\tau)}{\exp(sim(\boldsymbol{x}, \boldsymbol{z}_r)/\tau) + \exp(sim(\boldsymbol{x}, \boldsymbol{z}_s)/\tau)},$$

where τ is the temperature parameter and $\hat{y} \in \{r, s\}$ is the predicted class label. To account for the limited availability of training data, we align each image to an ensemble of class descriptions/ prompts called *context prompts*. We consider P descriptions per class and compute the text representation $\boldsymbol{z}$ for each description. An average of these representations ($\bar{\boldsymbol{z}}$) gives an ensemble of the context in the embedding space. Aligning the image with a multitude of natural language class descriptions enables the model to learn class-specific clues. The specific language descriptions used to describe the real and spoof classes are provided in Table 1.

3.4. FLIP-Multimodal-Contrastive-Learning

In **FLIP**-**M**ultimodal-**C**ontrastive-**L**earning (FLIP-MCL), we propose an additional multimodal contrastive learning objective to further enhance the generalizability of the extracted features and surmount the domain-gap and limited-data problems. This approach is motivated by the tremendous promise of contrastive view-based self-supervised learning methods [5, 57, 2]. In addition to the cross-entropy loss applied on the cosine similarity logits as described in Section 3.3, we also apply self-supervised $\texttt{simCLR}$ loss and mean squared error ($\texttt{MSE}$) loss. While the $\texttt{simCLR}$ loss is applied on a pair of image views, the $\texttt{MSE}$ loss enforces consistency between pairs of image-text views.

For the $\texttt{simCLR}$ loss, we follow the approach in [5] to create two views (denoted as I^{v_1} and I^{v_2}) of the given image I by applying different transformations. The features corresponding to the two transformed images are extracted using the image encoder $\mathcal{V}$ and further projected using a non-linear projection network $\mathcal{H}$. Finally, a contrastive loss is applied on the projected features.

Simple framework for Contrastive Learning of visual Representations"的縮寫

$$\boldsymbol{x}^{v_1} = \mathcal{V}(I^{v_1}), \quad \boldsymbol{x}^{v_2} = \mathcal{V}(I^{v_2})$$

$$\boldsymbol{h}_1 = \mathcal{H}(\boldsymbol{x}^{v_1}) \ , \ h_2 = \mathcal{H}(\boldsymbol{x}^{v_2}) \qquad \boldsymbol{h}_1, \boldsymbol{h}_2 \in \mathbb{R}^{d_h}.$$

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

For the $\texttt{MSE}$ loss, we first randomly sample two different prompts from the ground-truth class and get their text representations $\boldsymbol{z}^{v_1}$ and $\boldsymbol{z}^{v_2}$. We now have two image views and two text views. For each pair of image-text views, we compute the cosine similarity score between the image and text representations and enforce the consistency between the two similarity scores.

$$L_{mse} = (sim(\boldsymbol{x}^{v_1}, \boldsymbol{z}^{v_1}) - sim(\boldsymbol{x}^{v_2}, \boldsymbol{z}^{v_2}))^2$$

Table 2. Evaluation of cross-domain performance in Protocol 1, between MSU-MFSD (**M**), CASIA-MFSD (**C**), Replay Attack (**I**) and OULU-NPU (O). We run each experiment 5 times under different seeds and report the mean HTER, AUC, and TPR@FPR=1%.

	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
0-shot	MADDG (CVPR' 19) [38]	17.69	88.06	–	24.50	84.51	–	22.19	84.99	–	27.98	80.02	–	23.09
	MDDR (CVPR' 20) [44]	17.02	90.10	–	19.68	87.43	–	20.87	86.72	–	25.02	81.47	–	20.64
	NAS-FAS (TPAMI' 20) [53]	16.85	90.42	–	15.21	92.64	–	11.63	96.98	–	13.16	94.18	–	14.21
	RFMeta (AAAI' 20) [39]	13.89	93.98	–	20.27	88.16	–	17.30	90.48	–	16.45	91.16	–	16.97
	D^2AM (AAAI' 21) [6]	12.70	95.66	–	20.98	85.58	–	15.43	91.22	–	15.27	90.87	–	16.09
	DRDG (IJCAI' 21) [28]	12.43	95.81	–	19.05	88.79	–	15.56	91.79	–	15.63	91.75	–	15.66
	Self-DA (AAAI' 21) [46]	15.40	91.80	–	24.50	84.40	–	15.60	90.10	–	23.10	84.30	–	19.65
	ANRL (ACM MM' 21) [27]	10.83	96.75	–	17.85	89.26	–	16.03	91.04	–	15.67	91.90	–	15.09
	FGHV (AAAI' 21) [26]	9.17	96.92	–	12.47	93.47	–	16.29	90.11	–	13.58	93.55	–	12.87
	SSDG-R (CVPR' 20) [18]	7.38	97.17	–	10.44	95.94	–	11.71	96.59	–	15.61	91.54	–	11.28
	SSAN-R (CVPR' 22) [48]	6.67	98.75	–	10.00	96.67	–	8.88	96.79	–	13.72	93.63	–	9.80
	PatchNet (CVPR' 22) [42]	7.10	98.46	–	11.33	94.58	–	13.40	95.67	–	11.82	95.07	–	10.90
	GDA (ECCV' 22) [67]	9.20	98.00	–	12.20	93.00	–	10.00	96.00	–	14.40	92.60	–	11.45
0-shot	DiVT-M (WACV' 23) [23]	2.86	99.14	–	8.67	96.62	–	3.71	99.29	–	13.06	94.04	–	7.07
	ViT (ECCV' 22) [16]	**1.58**	**99.68**	**96.67**	5.70	98.91	88.57	9.25	97.15	51.54	7.47	98.42	69.30	6.00
5-shot	ViT (ECCV' 22) [16]	3.42	98.60	95.00	1.98	99.75	94.00	2.31	99.75	87.69	7.34	97.77	66.90	3.76
	ViTAF* (ECCV' 22) [16]	2.92	99.62	91.66	1.40	99.92	98.57	1.64	99.64	91.53	5.39	98.67	76.05	3.31
0-shot	FLIP-V	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	5.27	98.41	79.33	**0.44**	**99.98**	99.86	**2.94**	**99.42**	**84.62**	3.61	99.15	84.76	3.06
	FLIP-MCL	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**

We define the joint training objective as:

$$L_{mcl} = L_{ce} + L_{simCLR} + L_{mse}$$

We follow the same cosine similarity method described in Section 3.3 for inference.

4. Experiments

4.1. Experimental Setup

Datasets and DG Protocols: We evaluate our method on three different protocols. Following [16], we set up the first two protocols as a leave-one-domain-out testing protocol, where each dataset is considered as a domain and we evaluate the cross-domain performance on the left-out domain. In **Protocol 1**, we evaluate on the widely used cross-domain FAS benchmark datasets, MSU-MFSD **(M)** [49], CASIA-MFSD **(C)** [65], Idiap Replay Attack **(I)** [7], and OULU-NPU **(O)** [3]. For example, **OCI → M** represents the scenario where **O**, **C**, and **I** datasets are considered as source domains and **M** is the target domain. In **Protocol 2**, we evaluate our method on the large-scale FAS datasets, WMCA **(W)** [11], CASIA-CeFA **(C)** [25, 24], and CASIA-SURF **(S)** [61, 62]. To further evaluate the performance in the low-data regime, we follow [56] and set up **Protocol 3** as a single-source-to-single-target protocol. We use the **M**, **C**, **I**, and **O** datasets, where each source domain will have 3 combinations, one each with the other domains, giving us a total of 12 different scenarios. In each of the three protocols, similar to [16], we include CelebA-Spoof [64] as the supplementary training data to increase the diversity of training samples.

Implementation Details: We crop and resize the face images to $224 \times 224 \times 3$ and split them into a patch size of 16×16. For the image encoder, we use the ViT variant of the CLIP model. For the text input, we have curated a set of custom text prompts for each of the real and spoof classes as shown in Table 1. We use the Adam optimizer and set the initial learning rate to 10^{-6} and weight decay to 10^{-6}. For each domain, we set a batch size of 3 in **Protocol 1** and **Protocol 3** and a batch size of 8 in **Protocol 2**. For FLIP-V we use a two-layer MLP head containing fully-connected layers of dimensions 512 and 2 respectively. The dimensionality of the image representation is $d_v = 768$ and the dimension of the shared vision-language embedding space is $d_{vl} = 512$. For all the 3 variants of our approach, we train for 4000 iterations. In FLIP-V we update all the layers of the image encoder and MLP, for FLIP-IT we update all the layers of the image and text encoders, and for FLIP-MCL we update all the layers of the image encoder, text encoder, and the non-linear projection network $\mathcal{H}$. In FLIP-MCL, $\mathcal{H}$ consists of 3 linear layers of dimensions 512, 4096, and 256, and the first two layers are followed by BatchNorm and ReLU.

Evaluation Metrics: Following [16], we evaluate the model performance using the Half Total Error Rate (HTER), Area Under the Receiver Operating Characteristic Curve (AUC), and True Positive Rate (TPR) at a fixed False Positive Rate (FPR). Unlike most prior works that simply report the best result over a single trial, we run each

因為選擇是 true 或 false 是根據兩者的相似度大小，而標準可以調整，不一定是誰高就選誰

Table 3. Evaluation of cross-domain performance in Protocol 2, between CASIA-SURF (**S**), CASIA-CeFA (**C**), and WMCA (**W**). We run each experiment 5 times under different seeds and report the mean HTER, AUC, and TPR@FPR=1%

	Method	CS → W			SW → C			CW → S			Avg.
		HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER
0-shot	ViT (ECCV' 22) [16]	7.98	97.97	73.61	11.13	95.46	47.59	13.35	94.13	49.97	10.82
5-shot	ViT (ECCV' 22) [16]	4.30	99.16	83.55	7.69	97.66	68.33	12.26	94.40	42.59	6.06
	ViTAF* (ECCV' 22) [16]	2.91	99.71	92.65	6.00	98.55	78.56	11.60	95.03	60.12	5.12
0-shot	FLIP-V	6.13	97.84	50.26	10.89	95.82	53.93	12.48	94.43	53.00	9.83
	FLIP-IT	4.89	98.65	59.14	10.04	96.48	59.4	15.68	91.83	43.27	10.2
	FLIP-MCL	**4.46**	**99.16**	**83.86**	**9.66**	**96.69**	**59.00**	**11.71**	**95.21**	**57.98**	**8.61**

Table 4. Evaluation of cross-domain performance in Protocol 3, for all the 12 different combinations between MSU-MFSD (**M**), CASIA-MFSD (**C**), Replay Attack (**I**) and OULU-NPU (O). We run each experiment 5 times under different seeds and report the mean HTER.

	Method	C → I	C → M	C → O	I → C	I → M	I → O	M → C	M → I	M → O	O → C	O → I	O → M	Avg.
0-shot	ADDA (CVPR' 17) [40]	41.8	36.6	-	49.8	35.1	-	39.0	35.2	-	-	-	-	39.6
	DRCN (ECCV' 16) [12]	44.4	27.6	-	48.9	42.0	-	28.9	36.8	-	-	-	-	38.1
	DupGAN (CVPR' 18) [15]	42.4	33.4	-	46.5	36.2	-	27.1	35.4	-	-	-	-	36.8
	KSA (TIFS' 18) [21]	39.3	15.1	-	12.3	33.3	-	9.1	34.9	-	-	-	-	24.0
	DR-UDA (TIFS' 20) [45]	15.6	9.0	28.7	34.2	29.0	38.5	16.8	3.0	30.2	19.5	25.4	27.4	23.1
	MDDR (CVPR' 20) [44]	26.1	20.2	24.7	39.2	23.2	33.6	34.3	8.7	31.7	21.8	27.6	22.0	26.1
	ADA (ICB' 19) [43]	17.5	9.3	29.1	41.5	30.5	39.6	17.7	5.1	31.2	19.8	26.8	31.5	25.0
	USDAN-Un (PR' 21) [19]	16.0	9.2	-	30.2	25.8	-	13.3	3.4	-	-	-	-	16.3
	GDA (ECCV' 22) [67]	15.10	**5.8**	-	29.7	20.8	-	12.2	2.5	-	-	-	-	14.4
	CDFTN-L (AAAI' 23) [56]	**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
0-shot	FLIP-V	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	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	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**

設定 FPR 為 1% 的情況下，模型的 TPR 值

of our experiments 5 times with different random seeds and report the mean HTER, AUC, and TPR@FPR=1% in all the results. The standard deviation of the performance metrics is reported in the supplementary material along with the statistical hypothesis testing results.

Baseline Methods: The closest and state-of-the-art (SOTA) baseline methods for the proposed FLIP framework are ViT-based FAS methods reported in [16] and [23]. While [16] reports both zero-shot and five-shot performance, it uses only vanilla ViT for the zero-shot case, but both vanilla and adaptive ViTs (ViTAF) for the five-shot case. Only zero-shot performance is considered in [23]. Note that zero-shot refers to the setting where no sample from the target domain is used during training, while five-shot refers to the setting where 5 labeled samples from the target domain are used during training.

4.2. Cross-domain FAS Performance

Table 2, Table 3, and Table 4 report the zero-shot cross-domain performance for **Protocol 1**, **Protocol 2**, and **Protocol 3**, respectively. We can further extend the proposed FLIP framework for the five-shot setting following techniques similar to [16], and the corresponding five-shot results are provided in the supplementary material.

Comparison of proposed training strategies: Firstly, we analyze the performance of the FLIP-V variant, which is obtained by simple finetuning of a multimodal pre-trained ViT. The results in Tables 2, 3, and 4 show that even this simple strategy can achieve SOTA performance (in terms of average HTER) on all three protocols, demonstrating the zero-shot transfer capabilities of VLP models. Note that this result belies claims in [16] and [10] that full finetuning of a pre-trained ViT image encoder inhibits its generalizability. In two of the three protocols considered (Protocols 1 and 3), the FLIP-IT variant outperforms the FLIP-V variant. This illustrates the power of natural language supervision in generating more generalizable representations, especially when the training data is limited. Even in the case of Protocol 2, the FLIP-IT variant generalizes better than FLIP-V in two of the three scenarios (see Table 3), with poor performance only in the **CW → S** case. Finally, the proposed FLIP-MCL variant significantly outperforms all the SOTA methods for all three protocols in the zero-shot setting. In the case of Protocol 1, the zero-shot performance of FLIP-MCL is better than even the five-shot performance of the SOTA ViTAF. This clearly demonstrates the effectiveness of the proposed multimodal contrastive learning strategy.

Cross-domain performance in Protocol 1: The FLIP framework outperforms SOTA zero-shot methods in three out of four target domains (C=+5.2, I=+0.76, O=+5.16) and five-shot methods in two out of four target domains (C=+0.86, O=+3.08) by large margins. We observe that the performance drop in M (-3.37) is primarily due to the real samples being categorized as presentation attacks, thereby increasing the false negative error rate. Compared to zero-shot methods, we can also observe huge gains in TPR@FPR=1% in three out of the four domains (C=+11.43, I=+33.08, O=+22.98).

Cross-domain performance in Protocol 2: The proposed FLIP framework performs better than zero-shot ViT in all three domains (W=+3.52, C=+1.47, and S=+1.64) in terms of HTER. In terms of TPR@FPR=1%, we are able to see high gains of +10.25, +11.41, and +8.01 for the target domains W, C, and S respectively. Compared to Protocol 1, Protocol 2 has much more subjects ($>$ 1000 in CASIA-CeFA/SURF, compared to $\approx$50 in MCIO) and richer environmental variations, which once again proves the effectiveness of our approach in learning generalized features across different data regimes.

Cross-domain performance in Protocol 3: In the challenging single-source to single-target setting, our framework outperforms (in terms of average HTER) SOTA methods by a large margin of +8.36. Specifically, for the target domain O, we observe huge HTER improvements of +26.0, +25.7, and +21.65, when taking C, I, and O as the source domains respectively. Also, for the target domain C, we observe huge improvements of +11.22, +8.61, and +18.91, when taking I, M, and O as the source domains. For the target domain M, we observe improvements of +0.95, and +2.38, for source domains C and I, except for O (-2.1). For the target domain I, we observe that [56] does better for the source domains C and M, but for source domain O, our framework is able to perform on par. These results demonstrate that the FLIP-MCL method can learn strong generalizable features that could handle adverse limited-data and domain-gap problems.

4.3. Ablation Studies

Comparing various ViT initialization methods for FAS: To extend our observation regarding the effect of initialization on FAS generalizability, we take ViT pre-trained with different methods and show the comparative performance in Table 5. Specifically, we adopt the ViT training strategy proposed in [16] and a) train from scratch without any pre-trained weights, b) initialize with self-supervised BeIT [1] pre-training weights, c) initialize with ImageNet pre-trained weights [16] and d) initialize with multimodal CLIP [34] pre-trained weights. It can be seen that multimodal pre-trained initialization achieves better FAS generalizability compared to other initialization methods due to their ability to encode rich multimodal representations, serving as a base for all the experiments aligning image and text representations.

Table 5. Comparing different ViT initialization methods for FAS. We use each initialization method with their default parameters and show the results for **Protocol 1**.

Method	OCI → M		OMI → C		OCM → I		ICM → O		Avg.
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER
Scratch	18.32	87.36	40.05	61.13	19.22	88.15	29.72	73.66	25.86
BeIT [1]	4.73	98.46	7.86	96.62	13.51	92.42	15.19	91.95	8.70
ImageNet [16]	**1.58**	**99.68**	5.70	98.91	9.25	97.15	7.47	98.42	6.00
CLIP (FLIP-V)	3.79	99.31	**1.27**	**99.75**	**4.71**	**98.80**	**4.15**	**98.76**	**3.48**

Table 6. Impact of guidance with different text prompts (described in Table 1). We use FLIP-IT and show the results for **Protocol 1**.

Prompt	OCI → M		OMI → C		OCM → I		ICM → O		Avg.
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER
P1	6.00	98.17	0.54	99.97	3.60	99.19	3.47	99.24	3.40
P2	8.32	96.38	1.05	99.90	2.98	99.48	5.74	98.39	4.52
P3	**4.68**	**98.43**	**0.21**	**99.99**	4.30	99.06	4.07	99.02	3.31
P4	5.78	97.91	0.65	99.93	3.72	99.21	3.54	99.28	3.42
P5	6.48	98.37	0.46	99.96	**2.52**	**99.55**	3.24	99.30	3.17
P6	5.58	98.00	0.3	99.99	2.85	99.28	**3.03**	**99.46**	**2.94**
Ensemble	5.27	98.41	0.44	99.98	2.94	99.42	3.61	99.15	3.06

Table 7. Average HTER performance under different loss weights for Protocol 1. $L_{mcl} = \alpha L_{ce} + \beta L_{simCLR} + \gamma L_{mse}$

(α, β, γ)	**(1,1,1)**	**(1,1,0)**	**(1,0,1)**	**(1,2,2)**	**(1,5,5)**
HTER	3.01	3.15	3.47	3.20	3.67

Impact of different text prompts: In Table 6, we compare the effect of different text prompts in guiding the classification decision. It can be seen that different text prompts perform well for different cross-domain scenarios and it is difficult to choose a single prompt that works well across all the cases. Creating a list of different prompts for real and spoof classes is relatively easier and the performance of ensemble prompts shows that it is able to capture the best representation from each prompt while eliminating any inherent noise. This validates our idea of aligning the image representation to an ensemble of class prompts to learn generalized representations.

Contribution of different loss terms: We weight the different components of the joint training loss of FLIP-MCL as follows: $L_{mcl} = \alpha L_{ce} + \beta L_{simCLR} + \gamma L_{mse}$. A sensitivity analysis based on the tuple (α, β, γ) is provided in Table 7. Note that self-supervised losses L_{simCLR} and L_{mse} provide regularization in combination with the supervised cross-entropy loss L_{ce}. As we increase the

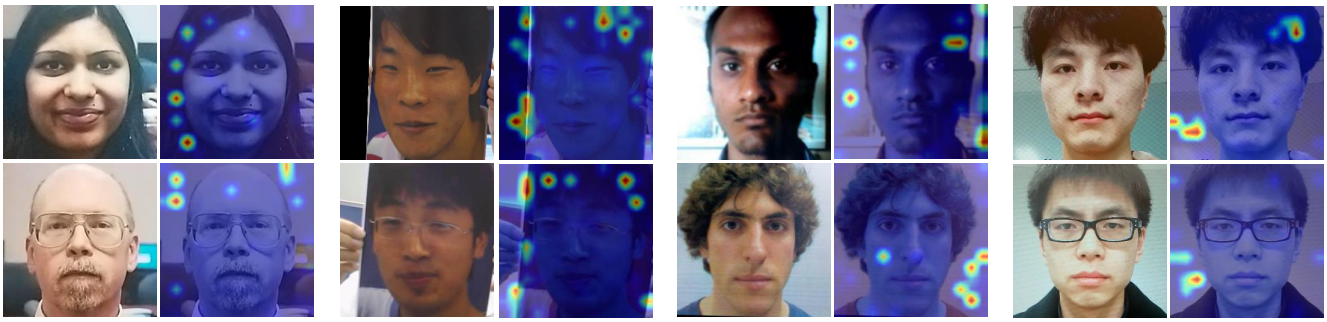

Figure 3. **Attention maps on spoof images from different scenarios in Protocol 1:** We observe that the attention highlights are on the spoof-specific clues such as paper texture (M), edges of the paper (C), and moire patterns (I and O).

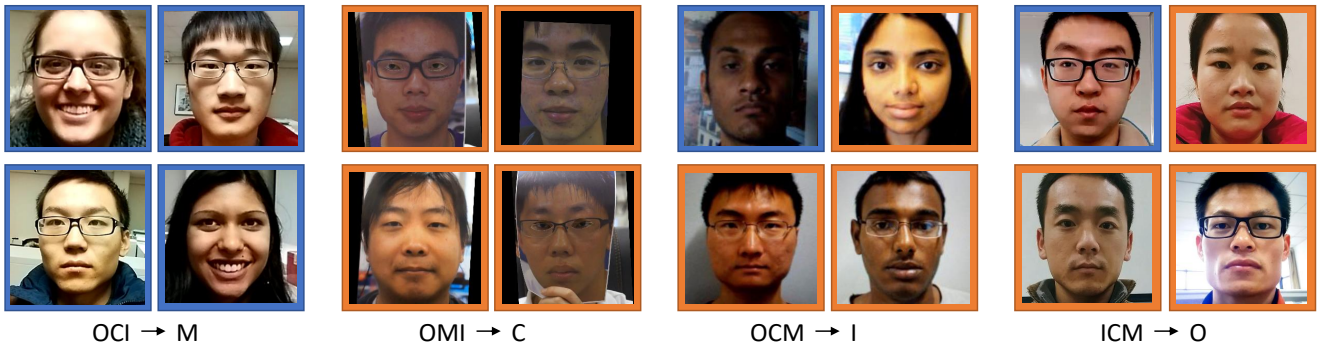

Figure 4. **Mis-Classified Examples in Protocol 1:** Blue boxes indicate real faces mis-classified as spoof. Orange boxes indicate spoof faces mis-classified as real.

都是因為 image resolution and lighting

importance of L_{simCLR} and L_{mse} losses (e.g., $(1, 2, 2)$ and $(1, 5, 5)$), it reduces the overall performance. This is expected because these settings decrease the contribution of L_{ce} during training. Similarly, the performance degrades when $\beta = 0$ or $\gamma = 0$, verifying that the self-supervised losses indeed facilitate better generalization.

4.4. Visualization

Attention maps: In Figure 3 and Figure 5, we use [4] to show the visual attention maps of the FLIP-MCL model on the spoof samples in **Protocol 1** and **Protocol 2** respectively. We can observe that our model is able to effectively localize the spoof patterns in each of the spoof domains to make the classification decision. In **Protocol 1** the datasets contain only print and replay attacks. We observe from the figure that the attention highlights are on the spoof-specific clues such as paper texture (M), edges of the paper (C), and moire patterns (I and O). In **Protocol 2**, for the CS $\rightarrow$ W scenario, we observe that the model focuses on spoof clues such as the edges of the paper/screen or the reflection on the screen. For the SW $\rightarrow$ C scenario, we observe that the model focuses on the region with cloth wrinkles. For the CW $\rightarrow$ S scenario, we observe that the model focuses on the cut region of the nose, or eyes.

Mis-Classified examples: In Figure 4, we show examples of images being mis-classified in **Protocol 1**. It is interesting to observe that for the OCI $\rightarrow$ M scenario, there are no false positive cases. i.e., none of the spoof samples have been predicted as real. However, as shown in Figure 4, some of the bonafide samples are mis-classified as spoof due to low image resolution and lighting variations, causing the performance to drop as shown in Table 2. In contrast, for the OMI $\rightarrow$ C scenario, we observe that none of the real samples are mis-classified as spoof, but a few high-resolution spoof samples are mis-classified as real. This could be due to the presence of high-resolution images from OULU (O) in training. For the OCM $\rightarrow$ I scenario, we observe that only 0.62% of the real samples are incorrectly classified. For the spoof samples, the mis-classification could be attributed to the adverse change in lighting conditions. For the ICM $\rightarrow$ O scenario, we again observe that a very low percentage (0.2%) of the real samples are mis-classified as spoof. Samples in O have higher resolution compared to the other datasets as shown, and this could be attributed to mis-classifying spoof as real.

In Figure 6, we show the examples of images being mis-

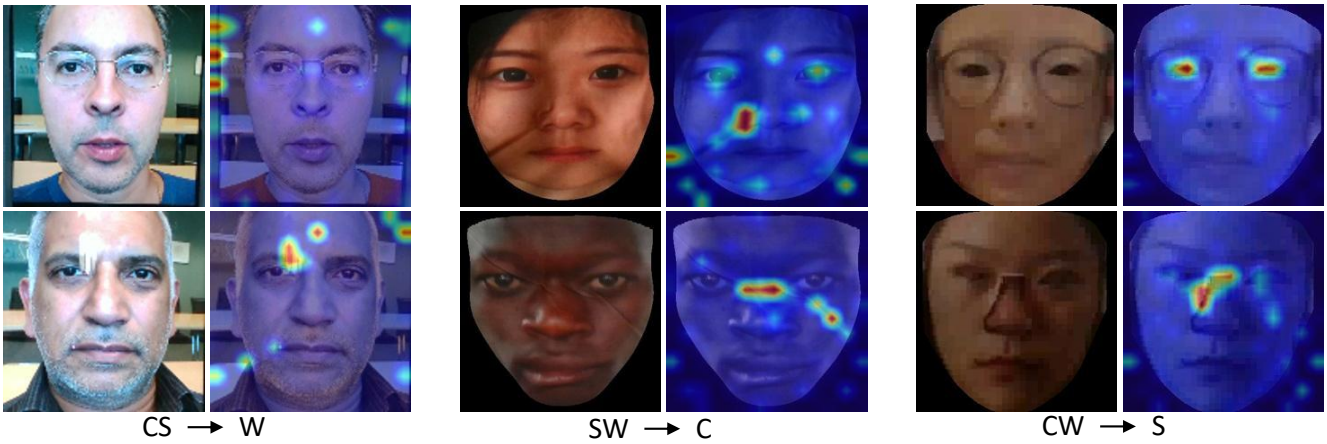

Figure 5. **Attention maps on spoof images from different scenarios in Protocol 2:** We observe that the attention highlights are on the spoof-specific clues such as screen edges/ screen reflection (W), wrinkles in printed cloth (C), and cut-out eyes/nose (S).

Figure 6. **Mis-Classified Examples in Protocol 2**: Blue boxes indicate real faces mis-classified as spoof. Orange boxes indicate spoof faces mis-classified as real.

classified in **Protocol 2**. For the CS → W scenario, we observe that some real samples are mis-classified as spoof due to the texture in the background region, which is identified as a moire spoof pattern visible in replay attacks. For the spoof samples being mis-classified as real, we observe that there are no clear visible spoof clues on these print and replay mediums. For the SW → C scenario, we observe that real samples in darker lighting conditions or a few faces with darker skin tones are mis-classified as spoof. The spoof sample mis-classification can be attributed to a realistic cloth print or print attack with no visible spoof clues, making it challenging for the model. For the CW → S scenario, we observe that most of the samples are of poor image resolution with a lot of pixelization. The real samples being mis-classified as spoof is either due to a) Pixelization, b) extreme pose changes, or c) darker lighting conditions. Some of the spoof samples that have higher resolution compared to the other samples get mis-classified as real.

5. Conclusion

In this work, we have shown that vision transformer models learned using vision-language pre-training (e.g., CLIP) have excellent generalization ability for the face anti-spoofing task, compared to their counterparts trained only on images. The rich multimodal representations learned by these models enable them to work well, even if only the image encoder is finetuned and used for presentation attack detection. On top of this baseline, we have shown that aligning the image representations to text representations produced by the text encoder further boosts generalizability. Using multimodal contrastive learning also enhances the generalizability across data regimes and domain gaps. The limitation of the later approaches is the additional computational overhead involved in invoking the text encoder during training. In the future, we plan to explore if these conclusions hold for other VLP foundation models. Prompt learning is also a potential way to further improve performance.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 7

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. 4

[3] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 612–618, 2017. 5

[4] 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, October 2021. 8

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3, 4

[6] 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*, volume 35, pages 1132–1139, 2021. 2, 5

[7] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, pages 1–7, 2012. 5

[8] Benjamin Devillers, Bhavin Choksi, Romain Bielawski, and Rufin VanRullen. Does language help generalization in vision models? In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 171–182, Online, Nov. 2021. Association for Computational Linguistics. 2, 3

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 3

[10] Anjith George and Sébastien Marcel. On the effectiveness of vision transformers for zero-shot face anti-spoofing. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021. 2, 6

[11] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 15:42–55, 2020. 5

[12] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. 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*, pages 597–613. Springer, 2016. 1, 6

[13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. 2, 3, 4

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[15] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Duplex generative adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1498–1507, 2018. 1, 6

[16] 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. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 37–54. Springer, 2022. 1, 2, 5, 6, 7

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[18] 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. 2, 5

[19] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing. *Pattern Recognition*, 115:107888, 2021. 1, 2, 6

[20] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2, 3, 4

[21] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7):1794–1809, 2018. 1, 2, 6

[22] Haoliang Li, Shiqi Wang, Peisong He, and Anderson Rocha. Face anti-spoofing with deep neural network distillation. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):933–946, 2020. 2

[23] Chen-Hao Liao, Wen-Cheng Chen, Hsuan-Tung Liu, Yi-Ren Yeh, Min-Chun Hu, and Chu-Song Chen. Domain invariant vision transformer learning for face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6098–6107, 2023. 2, 5, 6

[24] Ajian Liu, Xuan Li, Jun Wan, Yanyan Liang, Sergio Escalera, Hugo Jair Escalante, Meysam Madadi, Yi Jin, Zhuoyuan Wu, Xiaogang Yu, et al. Cross-ethnicity face anti-spoofing recognition challenge: A review. *IET Biometrics*, 10(1):24–43, 2021. 5

[25] 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 *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1178–1186, 2021. 5

[26] Shice Liu, Shitao Lu, Hongyi Xu, Jing Yang, Shouhong Ding, and Lizhuang Ma. Feature generation and hypothesis verification for reliable face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1782–1791, 2022. 2, 5

[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, 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. 2, 5

[29] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4680–4689, 2019. 1

[30] Yaojie Liu, Joel Stehouwer, and Xiaoming Liu. On disentangling spoof trace for generic face anti-spoofing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 406–422. Springer, 2020. 1

[31] Daniel Pérez-Cabo, David Jiménez-Cabello, Artur Costa-Pazo, and Roberto J López-Sastre. Learning to learn face-pad: a lifelong learning approach. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2020. 1

[32] Yunxiao Qin, Chenxu Zhao, Xiangyu Zhu, Zezheng Wang, Zitong Yu, Tianyu Fu, Feng Zhou, Jingping Shi, and Zhen Lei. Learning meta model for zero-and few-shot face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11916–11923, 2020. 1

[33] Ruijie Quan, Yu Wu, Xin Yu, and Yi Yang. Progressive transfer learning for face anti-spoofing. *IEEE Transactions on Image Processing*, 30:3946–3955, 2021. 2

[34] 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. 3, 7

[35] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 2, 3, 4

[36] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *Adv. Neural Inform. Process. Syst.*, 2022. 2, 3, 4

[37] Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. Is a caption worth a thousand images? a study on representation learning. In *International Conference on Learning Representations*. 2, 3

[38] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10031, 2019. 2, 5

[39] Rui Shao, Xiangyuan Lan, and Pong C Yuen. Regularized fine-grained meta face anti-spoofing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11974–11981, 2020. 2, 5

[40] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 1, 6

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[42] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20281–20290, 2022. 1, 5

[43] 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. 1, 2, 6

[44] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. 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*, pages 6678–6687, 2020. 1, 5, 6

[45] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 16:56–69, 2020. 1, 6

[46] 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*, volume 35, pages 2746–2754, 2021. 2, 5

[47] Yu-Chun Wang, Chien-Yi Wang, and Shang-Hong Lai. Disentangled representation with dual-stage feature learning for face anti-spoofing. In *Proceedings of the IEEE/CVF Win-*

ter Conference on Applications of Computer Vision, pages 1955–1964, 2022. 1

[48] 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. 2, 5

[49] Di Wen, Hu Han, and Anil K. Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015. 5

[50] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 3

[51] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Face anti-spoofing with human material perception. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 557–575. Springer, 2020. 1

[52] Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Deep learning for face anti-spoofing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[53] 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. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3005–3023, 2020. 1, 5

[54] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5295–5305, 2020. 1

[55] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3

[56] Haixiao Yue, Keyao Wang, Guosheng Zhang, Haocheng Feng, Junyu Han, Errui Ding, and Jingdong Wang. Cyclically disentangled feature translation for face anti-spoofing. *arXiv preprint arXiv:2212.03651*, 2022. 1, 2, 5, 6, 7

[57] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 4

[58] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 3

[59] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 641–657. Springer, 2020. 1

[60] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, page 493–510, Berlin, Heidelberg, 2022. Springer-Verlag. 2, 3, 4

[61] 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. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020. 5

[62] Shifeng Zhang, Xiaobo Wang, Ajian Liu, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 919–928, 2019. 5

[63] Wentian Zhang, Haozhe Liu, Feng Liu, Raghavendra Ramachandra, and Christoph Busch. Effective presentation attack detection driven by face related task. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 408–423, Cham, 2022. Springer Nature Switzerland. 2

[64] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 70–85. Springer, 2020. 5

[65] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z. Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 26–31, 2012. 5

[66] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2, 3, 4

[67] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Kekai Sheng, Shouhong Ding, and Lizhuang Ma. Generative domain adaptation for face anti-spoofing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 335–356. Springer, 2022. 1, 2, 5, 6

[68] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 350–368. Springer, 2022. 2, 3, 4