Paper Review

MLCF

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

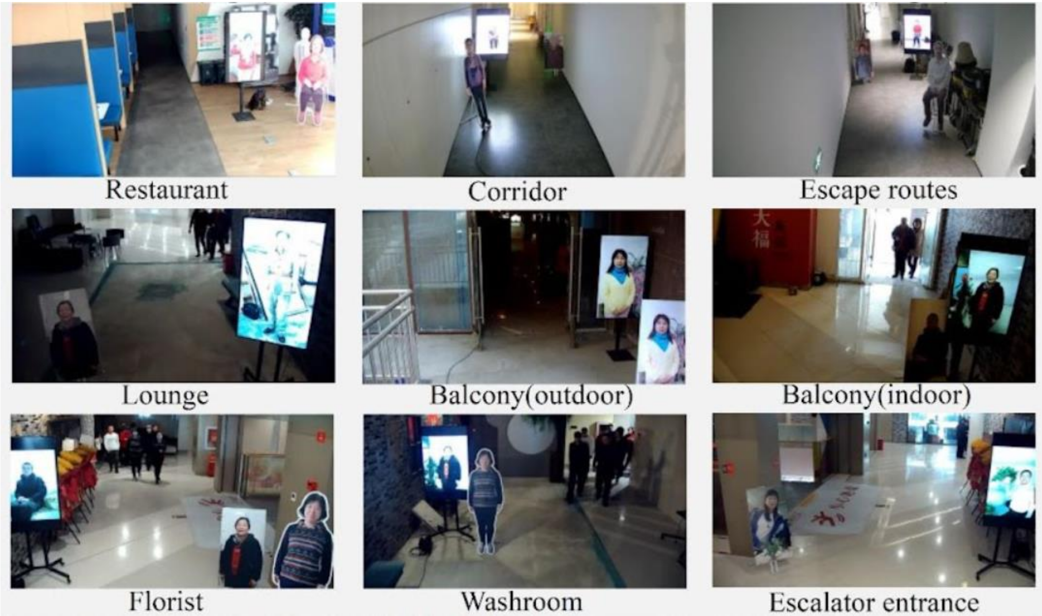
Koushik Srivatsan, et al.

ICCV 2023

Reviewed by Susang Kim

Contents

1.Introduction - FAS Application

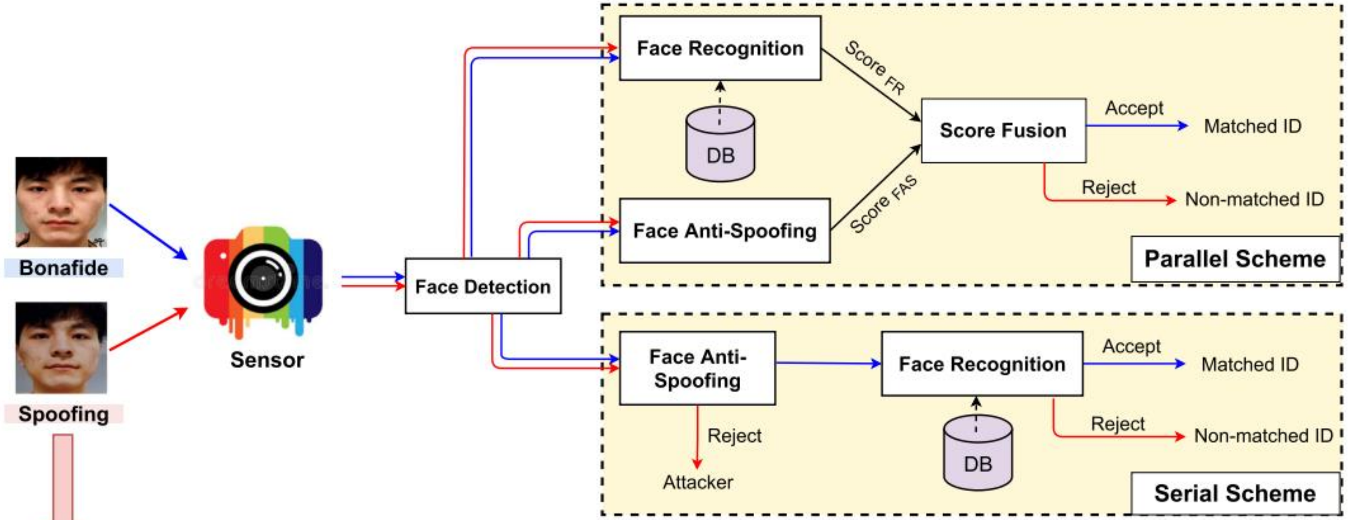


Surveillance scenes

Face Anti-spoofing in the wild

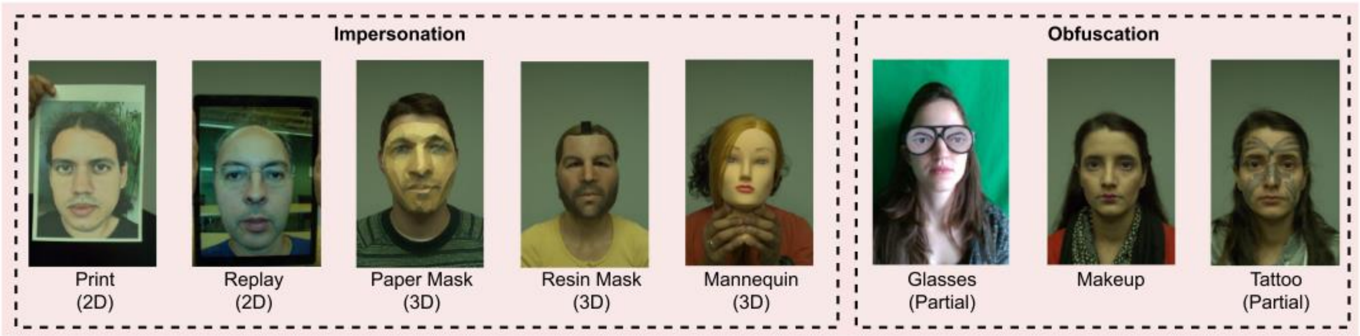
Face anti-spoofing (FAS) plays a vital role in securing face recognition systems from presentation attacks.

1.Introduction - FAS pipeline



(a) FAS pipeline in face recognition systems

(a) FAS could be integrated with face recognition systems with paralled or serial scheme for reliable face ID matching.

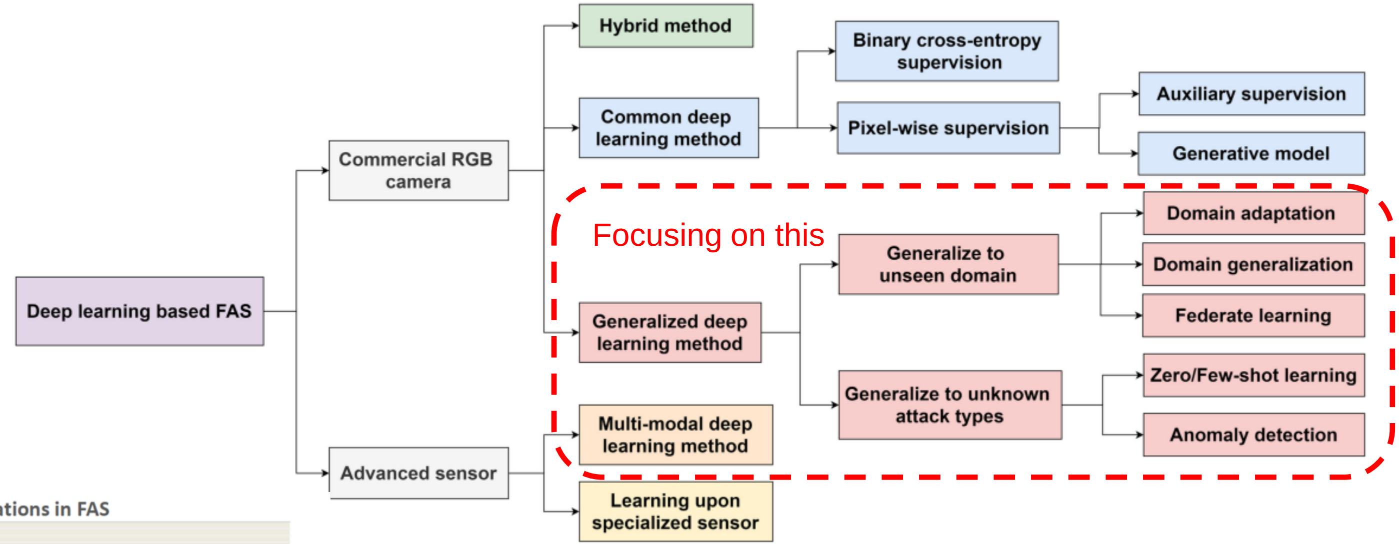


(b) Face spoofing attacks

(b) Visualization of several classical face spoofing attack types in terms of impersonation/obfuscation, 2D/3D, and whole/partial evidences.

Yu, Zitong, et al. "Deep learning for face anti-spoofing: A survey." TPAMI, 2022.

1.Introduction – Deep learning based FAS methods



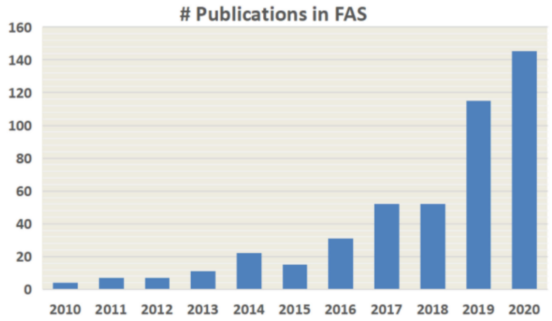


Fig. 1. The increasing research interest in the FAS field, obtained through Google scholar search with key-words: allintitle: "face anti-spoofing", "face presentation attack detection", and "face liveness detection".

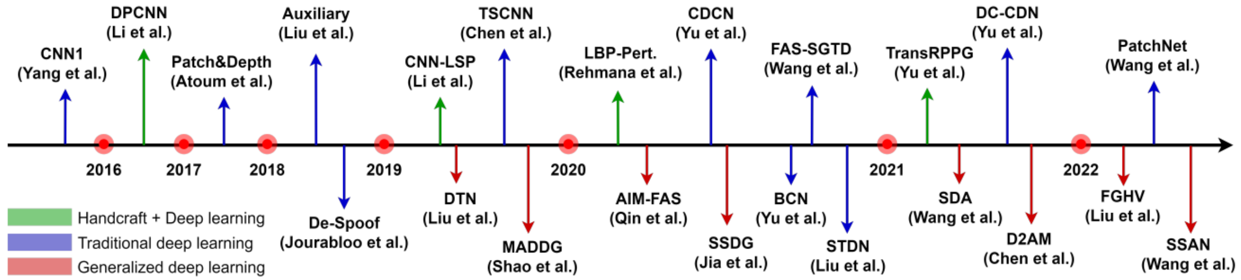
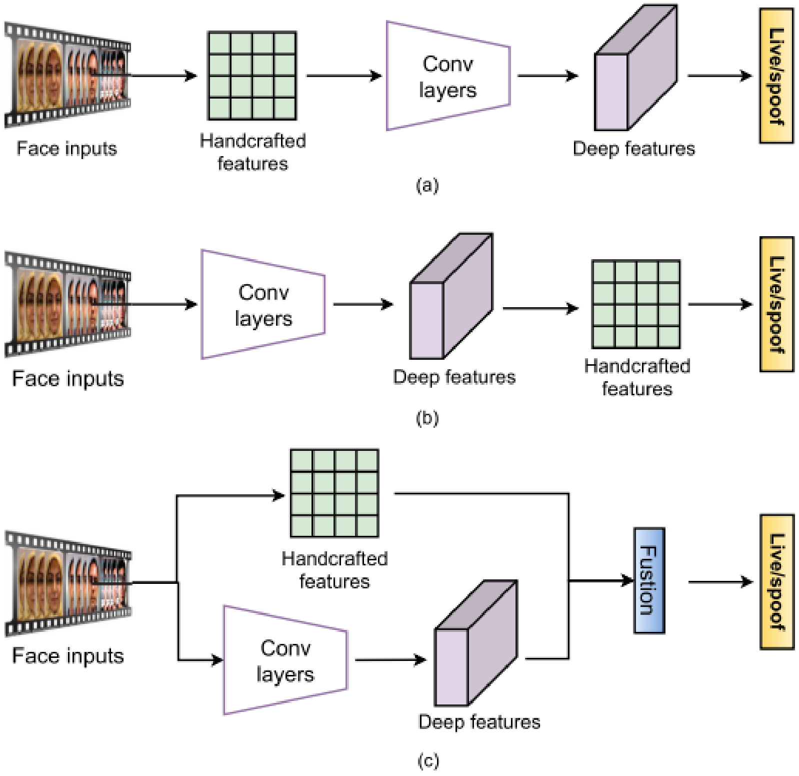
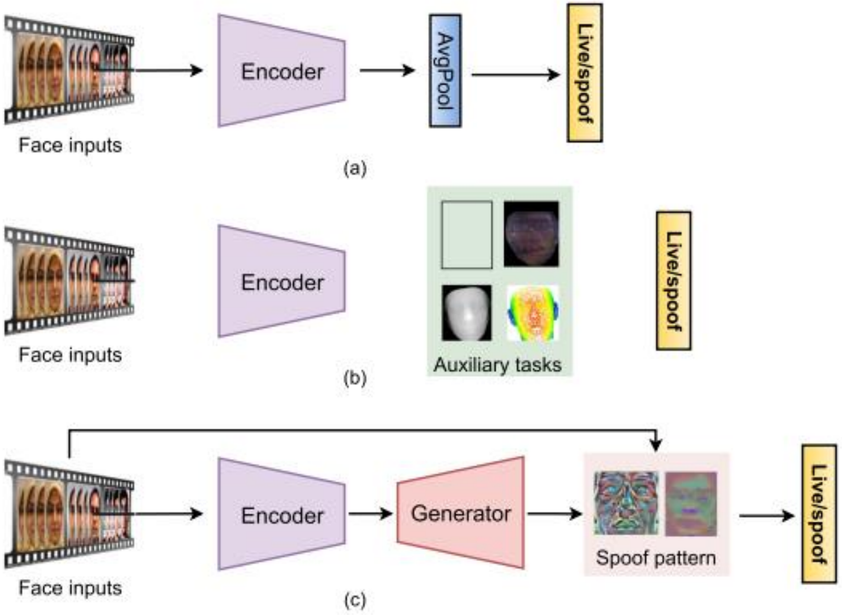


Fig. 6: Chronological overview of the milestone deep learning based FAS methods using commercial RGB camera.

Yu, Zitong, et al. "Deep learning for face anti-spoofing: A survey." TPAMI, 2022.

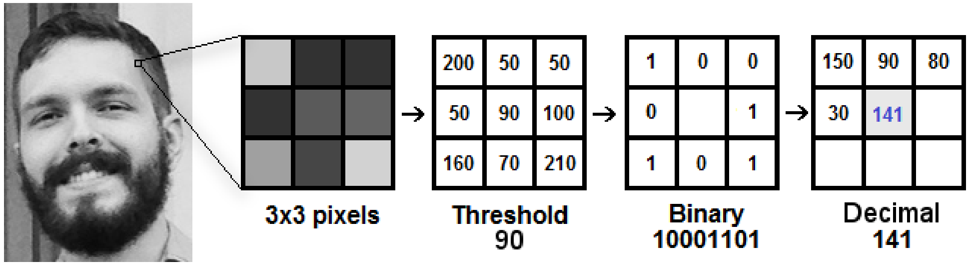
2.Related Works - Hybrid FAS vs Deep Learning(End to End)

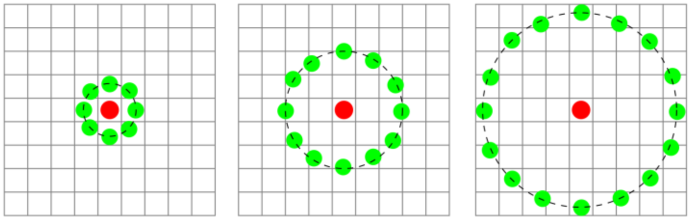


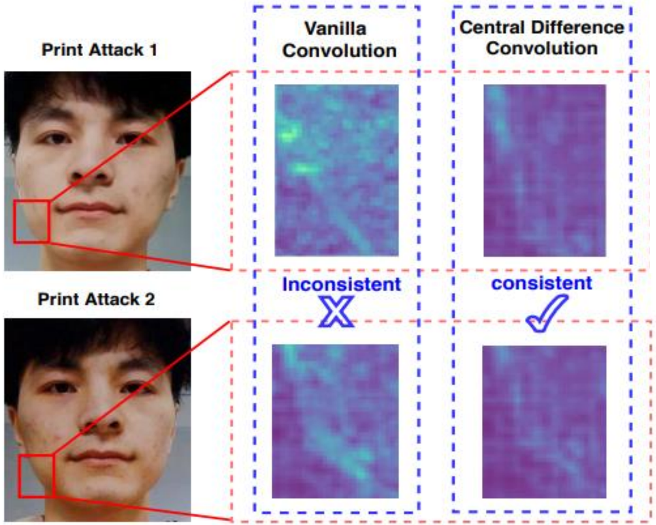


Hybrid frameworks for FAS. (a) Deep features from handcrafted features. (b) Handcrafted features from deep features. (c) Fused handcrafted and deep features.

2.Related Works - From Local Binary Pattern(LBP) to CDCN (CVPR 2020)







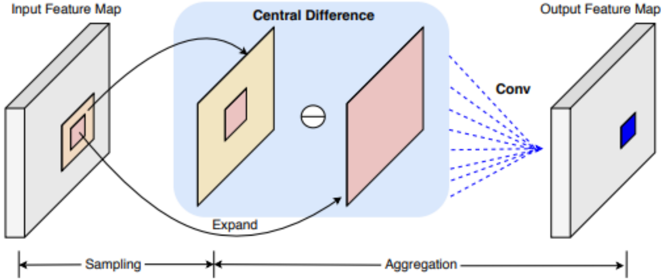
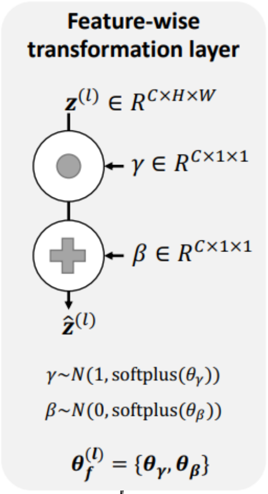


Figure 2. Central difference convolution.

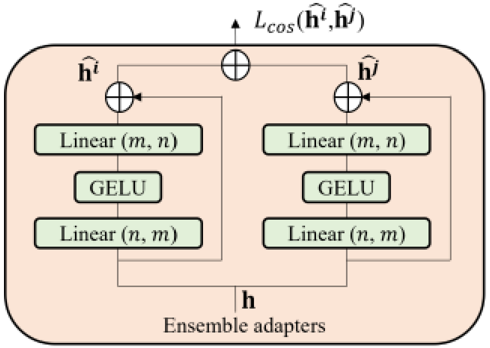
$$y(p_0) = \underbrace{\theta \cdot \sum_{p_n \in \mathcal{R}} w(p_n) \cdot (x(p_0 + p_n) - x(p_0))}_{\text{central difference convolution}} + \underbrace{(1-\theta) \cdot \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n)}_{\text{vanilla convolution}},$$

YU, Zitong, et al. Searching central difference convolutional networks for face anti-spoofing. CVPR 2020.

2.Related Works - Adaptive vision transformers (ViT) for FAS (ECCV 2022)

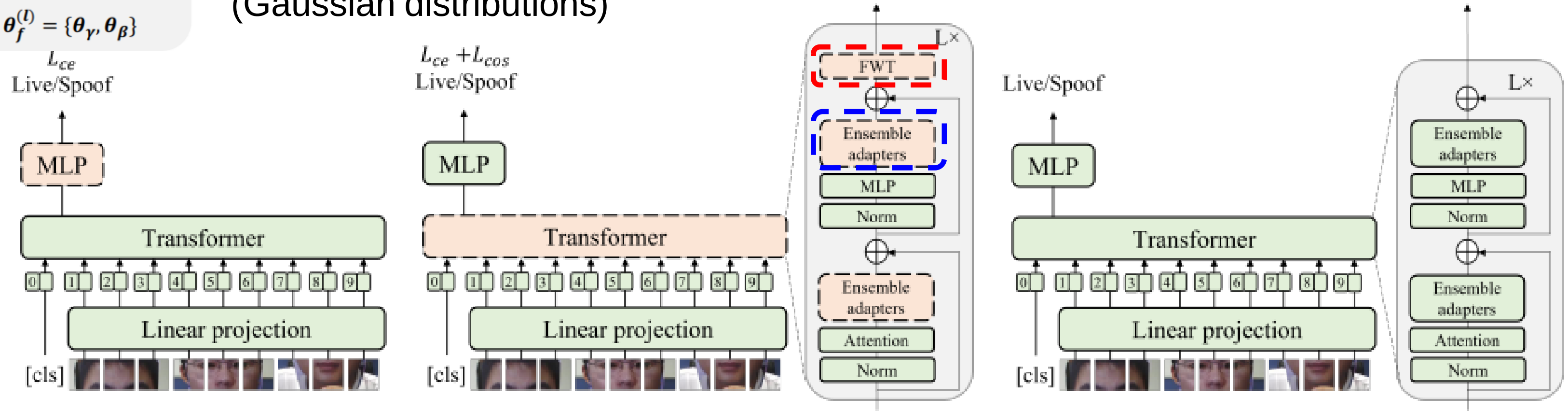


Improve performance by leveraging diverse modalities
(Incorporating separate encoders for each modality)
e.g., RGB+Reflection+Depth



FWT integrates a feature-wise transformation to augment the intermediate feature activations with affine transformations. (Gaussian distributions)

Ensemble adapters is adapted for face anti-spoofing task



(a) Pre-training (b) Fine-tuning (c) Testing

HUANG, Hsin-Ping, et al. “Adaptive transformers for robust few-shot cross-domain face anti-spoofing.” ECCV 2022.
Tseng, Hung-Yu, et al. "Cross-domain few-shot classification via learned feature-wise transformation.“ ICLR 2020.

2.Related Works – Vision Language Pre-training (CLIP)

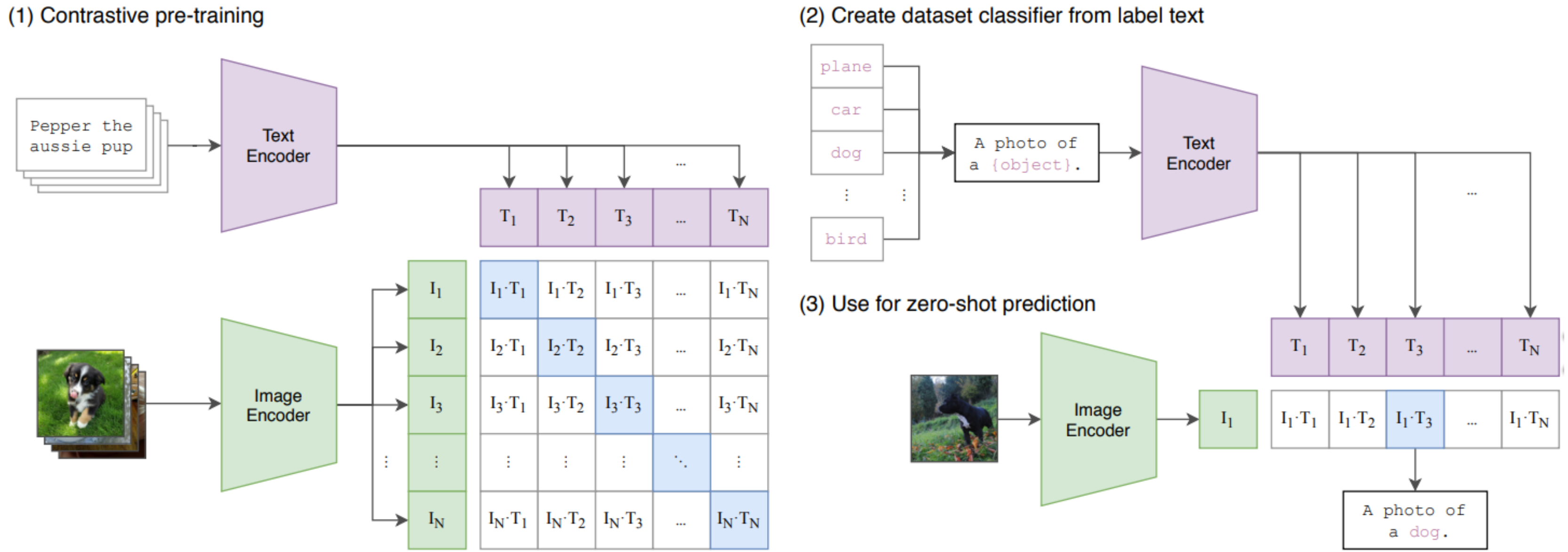


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet.
Contrastive learning and Self-supervised learning.

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." ICML 2021.

3.Methods – Multimodal & Contrastive learning

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

Multimodal contrastive learning strategy to boost generalization with CLIP encoder.

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

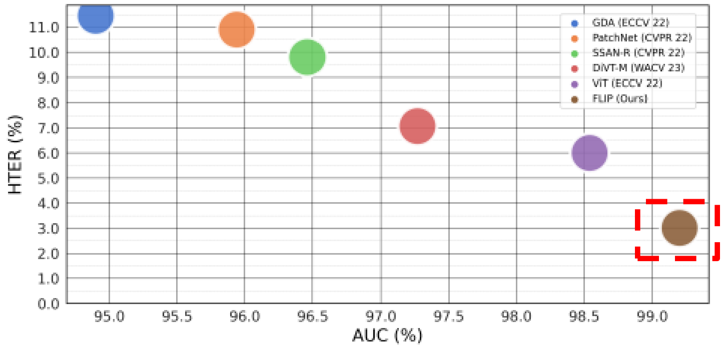
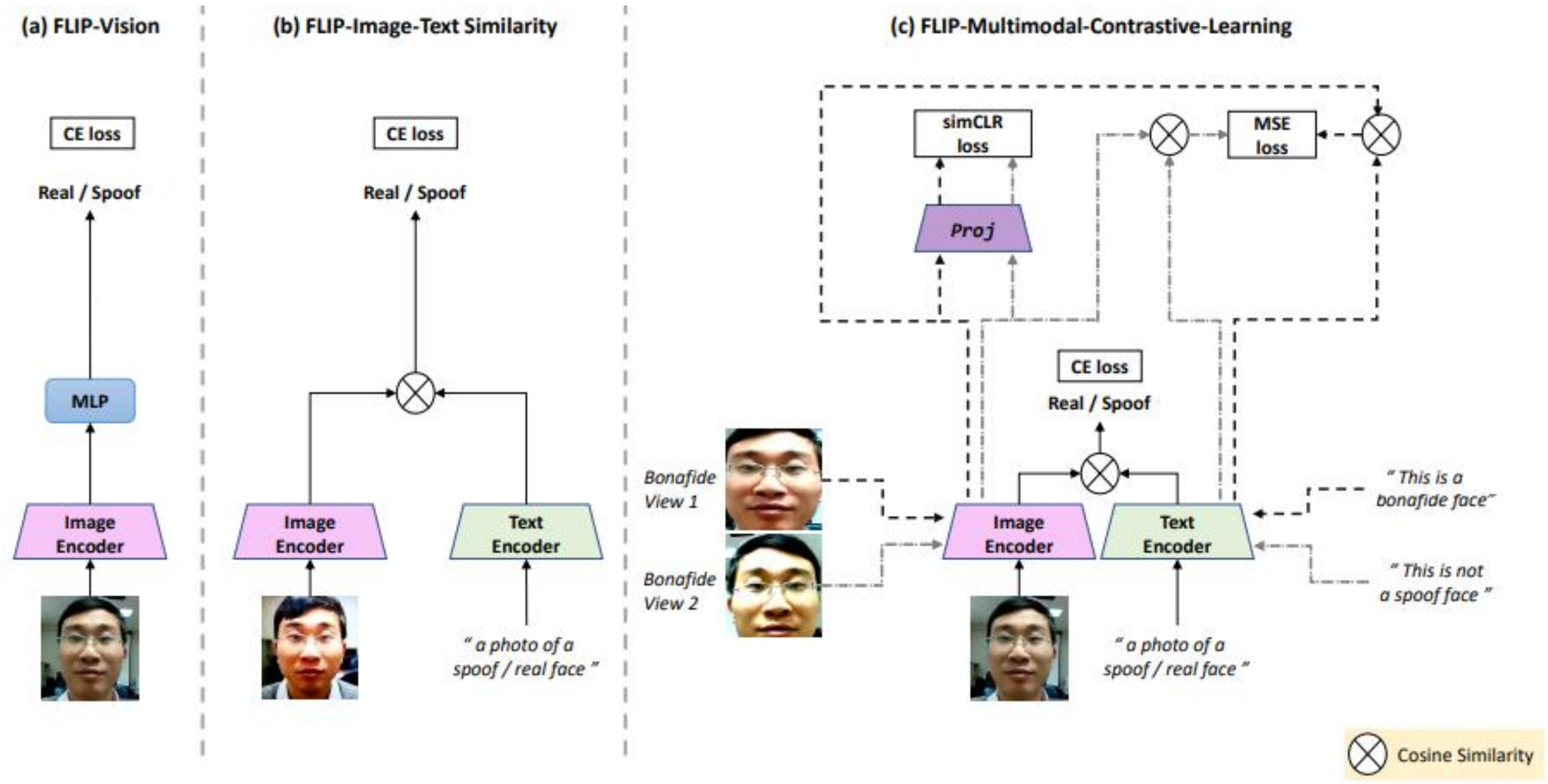
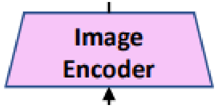


Figure 1. Area Under ROC Curve (AUC %) and Half Total Error Rate (HTER %) comparison between our proposed method and state-of-the-art (SOTA). Our method achieves the highest AUC (↑) performance with the lowest HTER (↓) for cross-domain face anti-spoofing on MCIO datasets, surpassing all the SOTA methods.

3.Method - Overview of the proposed FLIP framework



3.Method - Image Encoder



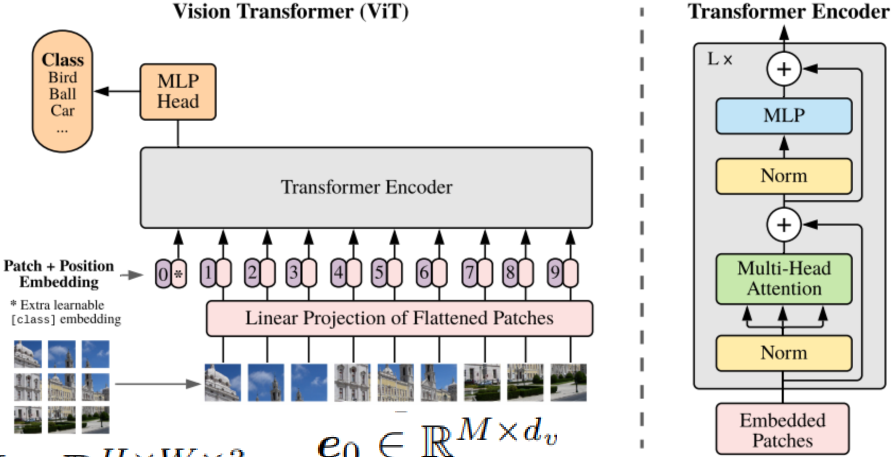
$\mathcal{V}$ consisting of K transformer blocks $\{\mathcal{V}_k\}_{k=1}^{K}$

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

class token c_{k-1} Patch embeddings $\boldsymbol{e}_{k-1}$

The final image representation x is obtained by linearly projecting the class token Ck from the last transformer block $(\mathcal{V}_K)$

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



$I \in \mathbb{R}^{H \times W \times 3}$ $\qquad e_0 \in \mathbb{R}^{M \times d_v}$

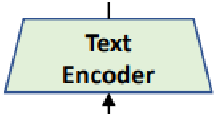
M : fixed-size patches
d : patch embeddings

```
class feature_generator_clip(nn.Module):

  def __init__(self):
    super(feature_generator_clip, self).__init__()
    self.vit, _ = clip.load("ViT-B/16", device='cuda')
    self.vit = self.vit.visual

  def forward(self, input):
    feat = self.vit.forward_full(input)
    return feat
```

3.Method - Text Encoder



Word embedding $\boldsymbol{w}_0 = [w_0^1, w_0^2, \cdots, w_0^Q] \in \mathbb{R}^{Q \times d_l}.$

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

transformer block $(\mathcal{L}_k)$

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

The final text representation z is obtained by projecting the text embeddings corresponding to the last token of the last transformer block $(\mathcal{L}_K)$

A Base size we use a 63M-parameter 12- layer 512-wide model with 8 attention heads. (GPT-2)

The transformer operates on a lower-cased Byte Pair Encoding (BPE : subword tokenizer) representation of the text with a 49,152 vocab size.

```
def encode_text(self, text):
    x = self.token_embedding(text).type(self.dtype)
    x = x + self.positional_embedding.type(self.dtype)
    x = x.permute(1, 0, 2)  # NLD -> LND
    x = self.transformer(x)
    x = x.permute(1, 0, 2)  # LND -> NLD
    x = self.ln_final(x).type(self.dtype)
    x = x[torch.arange(x.shape[0]), text.argmax(dim=-1)] @ self.text_projection

    return x
```

3.Method – Contrastive Loss (SimCLR)

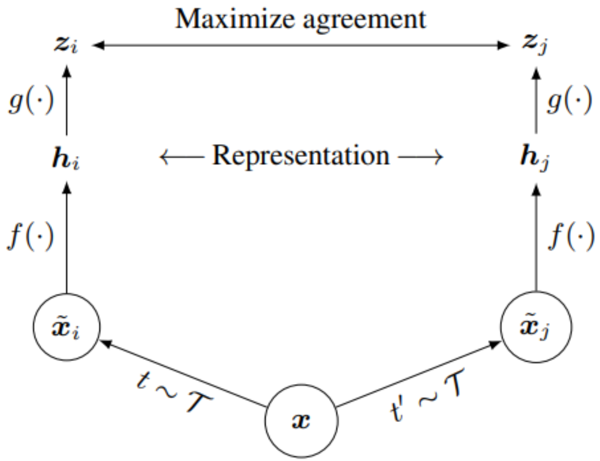
Contrastive representation learning for images has found that contrastive objectives can learn better representations than their equivalent predictive objective.
The InfoNCE (Nagative Constrastive) loss was adapted for contrastive (text, image) representation learning.

Given a set $X = \{x_1, \ldots, x_N\}$ of N random samples containing one positive sample from $p(x_{t+k}|c_t)$ and $N-1$ negative samples from the 'proposal' distribution $p(x_{t+k})$, we optimize:

$$\mathcal{L}_N = -\mathbb{E}_X \left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)} \right]$$

Optimizing this loss will result in $f_k(x_{t+k}, c_t)$ estimating the density ratio, which is:

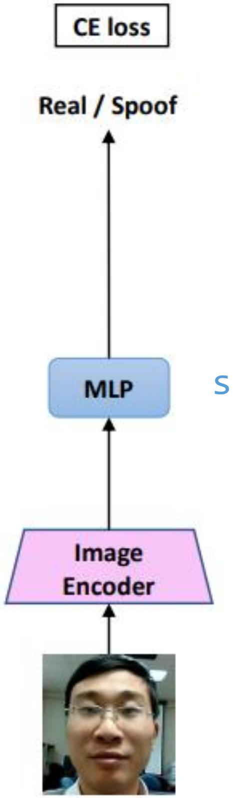
$$f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k}|c_t)}{p(x_{t+k})}$$



SimCLR - two separate data augmentation operators are sampled from the same family of augmentations (t ~ T and t' ~ T)

Oord, A. V. D., Li, Y., & Vinyals, O. Representation learning with contrastive predictive coding. arXiv preprint.
CHEN, Ting, et al. A simple framework for contrastive learning of visual representations. ICML 2020.

3.Method – FLIP Vision



$$L_{\mathrm{CE}} = -\sum_{i=1}^{n} t_i \log(p_i), \quad \text{for n classes,}$$

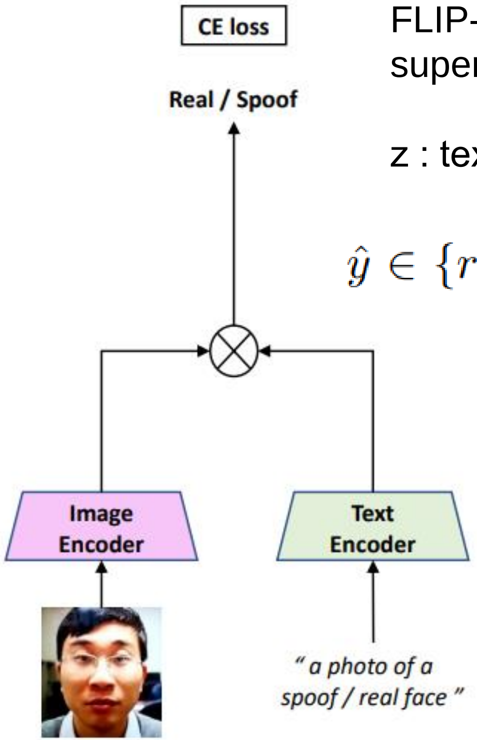
```
criterion = {'softmax': nn.CrossEntropyLoss().cuda()}
classifier_label_out , feature = net1(input_data, True)
cls_loss = criterion['softmax'](classifier_label_out.narrow(0, 0,
input_data.size(0)), source_label)
```

```
self.classifier_layer = nn.Linear(512, 2)
```

Pre-trained CLIP model and use only its image encoder V and discard the text encoder L

Representations produced by CLIP have shown impressive out-of-the-box performance for many downstream vision applications based on natural images such as classification, object detection, and segmentation.

3.Method – FLIP IT(Image-Text Similarity)



FLIP-Image-Text similarity, we obtain the prediction with the help of language supervision instead of using the MLP head.

z : text representation, x : image representation, τ : temperature parameter

$$\hat{y} \in \{r, s\}$$

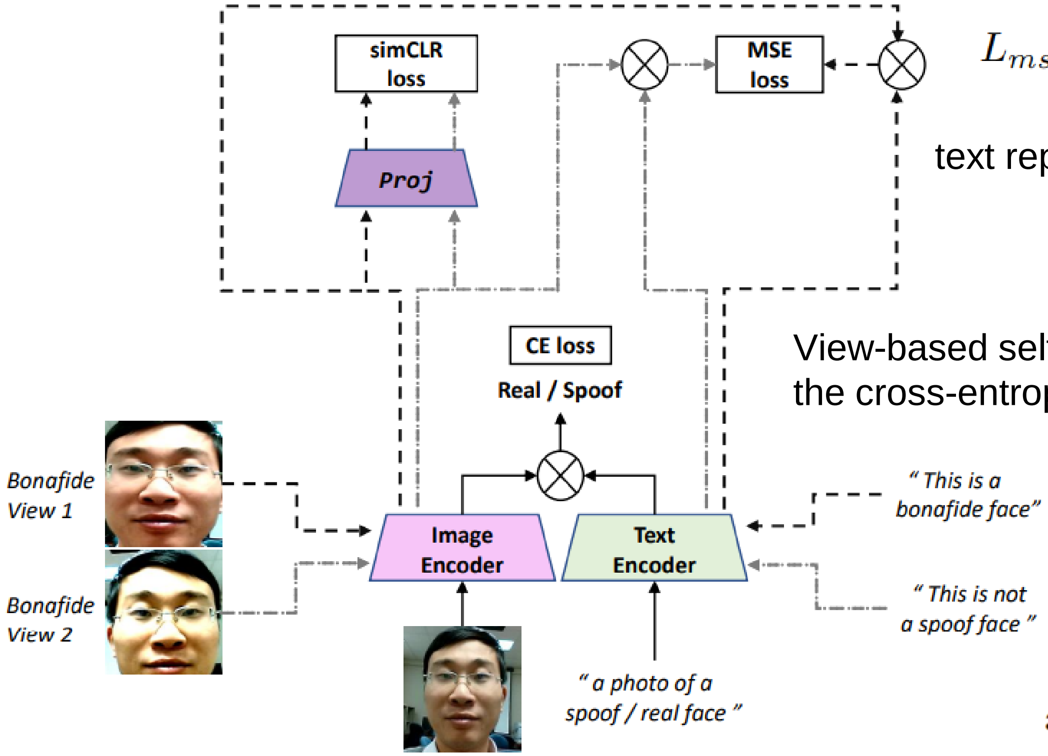
$$p(\hat{y}|x) = \frac{\exp(sim(\boldsymbol{x}, \boldsymbol{z}_{\hat{y}})/\tau)}{\exp(sim(\boldsymbol{x}, \boldsymbol{z}_r)/\tau) + \exp(sim(\boldsymbol{x}, \boldsymbol{z}_s)/\tau)},$$

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.

Aligning the image with a multitude of natural language class descriptions enables the model to learn class specific clues.

3.Method – FLIP-MCL(Multimodal-Contrastive-Learning)



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

text representations $\boldsymbol{z}^{v_1}$ and $\boldsymbol{z}^{v_2}$

View-based self-supervised learning methods
the cross-entropy loss applied on the cosine similarity logits

image encoder V
non-linear projection network H

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

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

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

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

4. Experiments - Datasets and DG Protocols

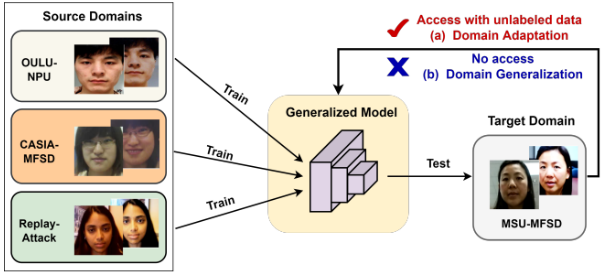
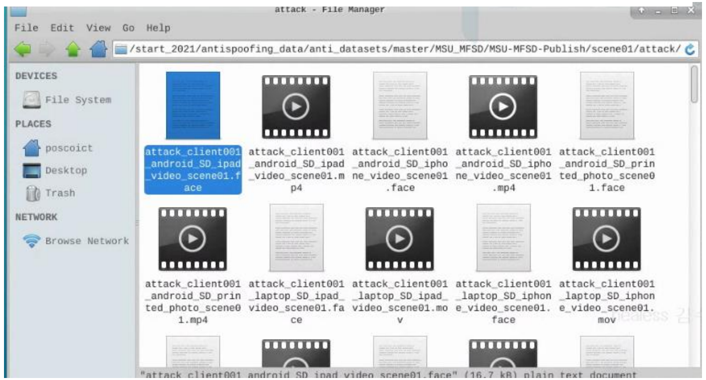
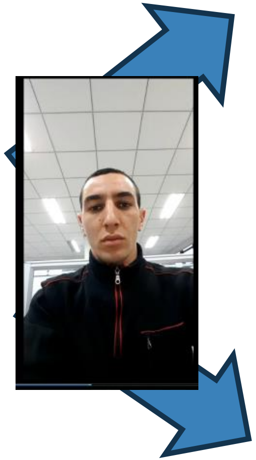


Table 5. **Four datasets for Leave-One-Out test.**

Dataset	Live/Spoof	Attack Types
CASIA-MFSD [83]	150/450	Print, Replay
REPLAY-ATTACK [8]	200/1000	Print, Replay
MSU-MFSD [73]	70/210	Print, Replay
OULU-NPU [6]	720/2880	Print, Replay





Real

Attack

many redundant images.

Video(RGB) -> Frame sampling -> Crop Face 256x256 (MTCNN)

4. Experiments - Implementation Details

Protocol 1 : The widely used cross-domain FAS benchmark datasets,
MSU-MFSD (**M**)[1], CASIA-MFSD (**C**)[2], Idiap Replay Attack (**I**)[3], and OULU-NPU (**O**) [4].
OCI (source domains) → M (target domain)

Protocol 2 : The large-scale FAS datasets,
WMCA (W), CASIA-CeFA (C), and CASIA-SURF (S).
CS (source domains) → W (target domain)

Protocol 3 : The low-data regime as a single-source-to-single-target. (a total of 12 different scenarios.
C → I ,C → M, C → O, I → C, I → M, I → O, M → C, M → I, M → O, O → C, O → I, O → M

For all protocols, we incorporate CelebA-Spoof as supplementary training data to enhance the diversity of training samples,

[1] Di Wen, et al. Face spoof detection with image distortion analysis. IEEE Transactions on Information Forensics and Security, 2015.
[2] Zhiwei Zhang, et al. A face antispoofing database with diverse attacks. IAPR International Conference on Biometrics (ICB), 2012.
[3] Ivana Chingovska, et al. On the effectiveness of local binary patterns in face antispoofing. (BIOSIG), 2012.
[4] Zinelabinde Boulkenafet, et al. Oulu-npu: A mobile face presentation attack database with real-world variations. IEEE International Conference on Automatic Face & Gesture Recognition 2017.

4. Experiments - Implementation Details

Crop and resize the face images to $224 \times 224 \times 3$ and split them into a patch size of 16×16.

Adam optimizer and set the initial learning rate to 10^6 and weight decay to 10^6

batch size of 3 in Protocol 1, Protocol 3.
batch size of 8 in Protocol 2.

FLIP-V uses a two-layer MLP head containing fully connected layers of dimensions 512 and 2.

Dimensionality of image representation is 768
Dimension of the shared vision-language embedding space is 512

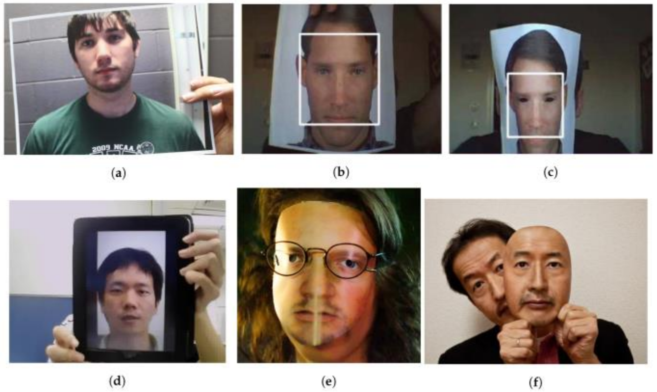
train for 4000 iterations

FLIP-V update all the layers of the image encoder and MLP.
FLIP-IT update all the layers of the image and text encoders.
FLIP-MCL update all the layers of the image encoder, text encoder, and the non-linear projection network H.
H consists of 3 linear layers of dimensions 512, 4096, and 256, and the first two layers are followed by BatchNorm and ReLU.

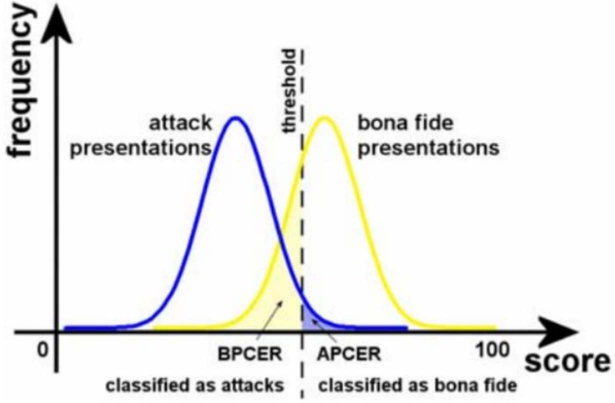
4. Experiments – Evaluation metric

Table 1. Comparison of existing face PAD databases. (* indicates the dataset only contains images. AS: Asian, A: Africa, U: Caucasian, I: Indian, E: East Asia, C: Central Asia.)

Dataset	Year	#Subject	#Num	Attack	Modality	Device	Ethnicity
Replay-Attack [9]	2012	50	1200	Print,Replay	RGB	RGB Camera	-
CASIA-FASD [46]	2012	50	600	Print,Cut,Replay	RGB	RGB Camera	-
3DMAD [12]	2014	17	255	3D print mask	RGB/Depth	RGB Camera/Kinect	-
MSU-MFSD [41]	2015	35	440	Print,Replay	RGB	Cellphone/Laptop	-
Replay-Mobile [11]	2016	40	1030	Print,Replay	RGB	Cellphone	-
Msspoof [10]	2016	21	4704*	Print	RGB/IR	RGB/IR Camera	-
OULU-NPU [8]	2017	55	5940	Print,Replay	RGB	RGB Camera	-
SiW [24]	2018	165	4620	Print,Replay	RGB	RGB Camera	AS/A/U/I
CASIA-SURF [45]	2019	1000	21000	Print,Cut	RGB/Depth/IR	Intel Realsense	E
CeFA (Ours)	2019	1500	18000	Print, Replay	RGB/Depth/IR	Intel Realsense	A/E/C
		99	5346	3D print mask			
		8	192	3D silica gel mask			
		Total: **1607** subjects, **23538** videos					



Print attack, Replay/video attack ,3D mask attack



$$APCER = \frac{\#\text{ of accepted attacks}}{\#\text{ of attacks}} \qquad BPCER = \frac{\#\text{ of rejected real attempts}}{\#\text{ of real attempts}}$$

$$ACER(\tau) = \frac{APCER(\tau) + BPCER(\tau)}{2} \quad [\%]$$

Attack Presentation Classification Error Rate (APCER)
Normal Presentation Classification Error Rate (NPCER)
Average Classification Error Rate (ACER)

LIU, Ajian, et al. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. WACV 2021.

4. Experiments - Cross-domain FAS Performance

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**

4. Experiments - Cross-domain FAS Performance

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**

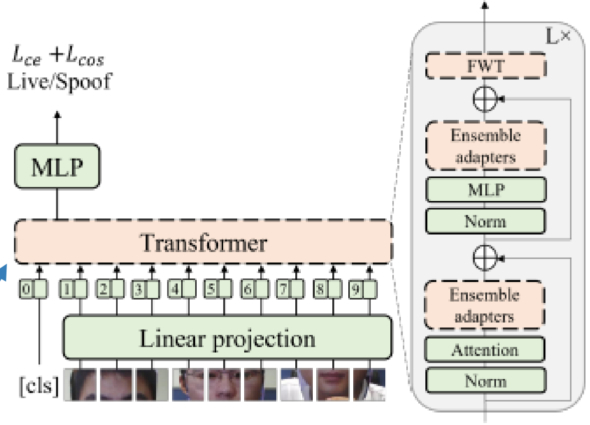
Table 4. Evaluation of cross-domain performance in Protocol 3, for all the 12 different combinations between MSU-MFSD (**M**), CASIA-MFSD (**C**), Replay Attack (**I**) and OULU-NPU (O). We run each experiment 5 times under different seeds and report the mean HTER.

	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**

4. Experiments - Ablation Studies

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**



(b) Fine-tuning

[1] BAO, Hangbo, et al. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254, 2021.
[16] Hsin-Ping Huang, et al. Adaptive transformers for robust few-shot cross-domain face anti-spoofing. ECCV 2022.

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

Similarly, the performance degrades when β = 0 or γ = 0, verifying that the self-supervised losses indeed facilitate better generalization.

4. Experiments - Ablation Studies

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

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

4. Experiments – Visualization (Attention maps on spoof images)

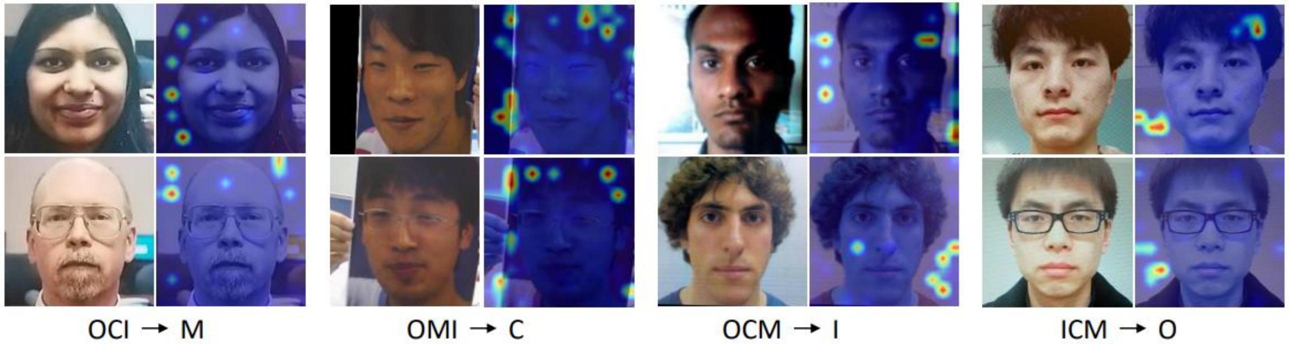


Figure 3. **Attention maps on spoof images from different scenarios in Protocol 1:** We observe that the attention highlights are on the spoof-specific clues such as paper texture (M), edges of the paper (C), and moire patterns (I and O).

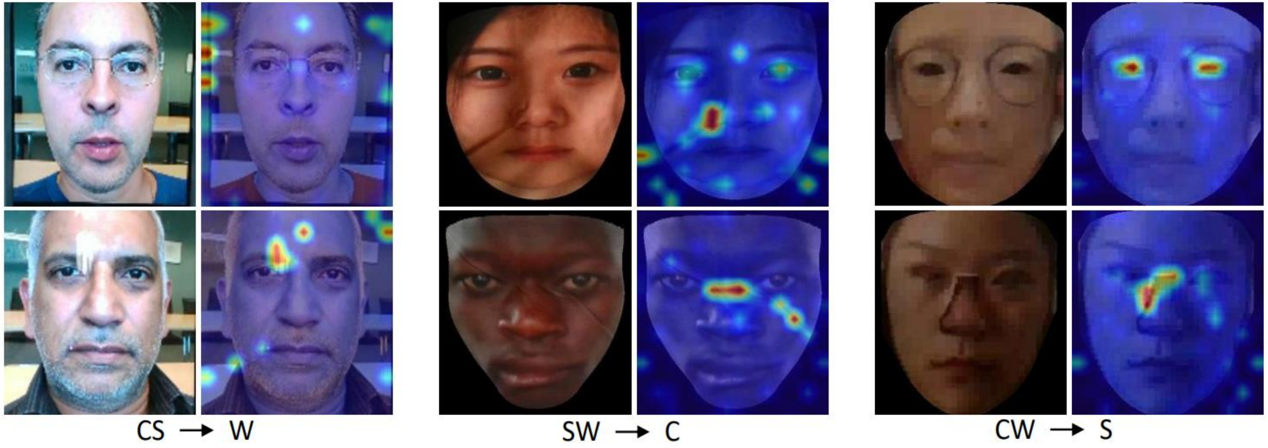
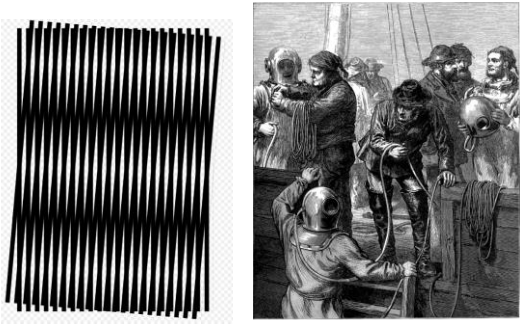


Figure 5. **Attention maps on spoof images from different scenarios in Protocol 2:** We observe that the attention highlights are on the spoof-specific clues such as screen edges/ screen reflection (W), wrinkles in printed cloth (C), and cut-out eyes/nose (S).

FLIP-MCL model on the spoof samples in Protocol 1 and Protocol 2.

Protocol 1 only print and replay attacks - attention highlights are on the spoof-specific clues such as paper texture (M), edges of the paper (C), and moire patterns (I and O).

Protocol 2 focuses on spoof clues such as the edges of the paper/screen or the reflection on the screen.



https://en.wikipedia.org/wiki/Moir%C3%A9_pattern

4. Experiments – Visualization (Mis-Classified examples)

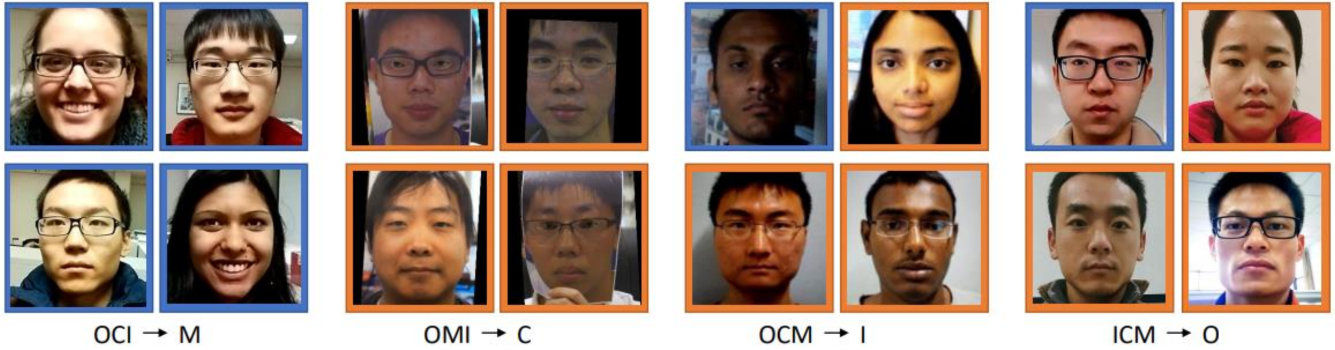


Figure 4. **Mis-Classified Examples in Protocol 1:** Blue boxes indicate real faces mis-classified as spoof. Orange boxes indicate spoof faces mis-classified as real.

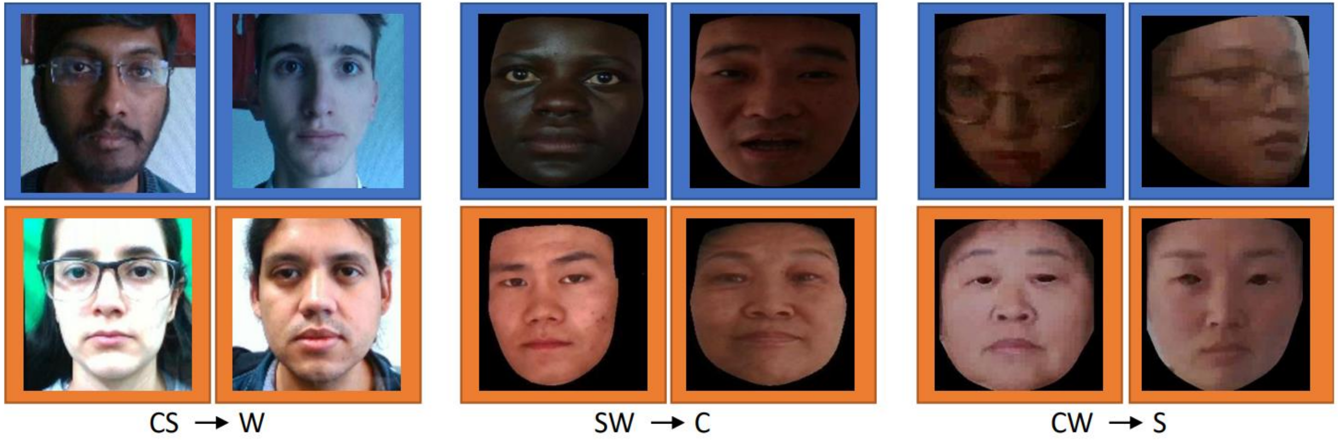


Figure 6. **Mis-Classified Examples in Protocol 2**: Blue boxes indicate real faces mis-classified as spoof. Orange boxes indicate spoof faces mis-classified as real.

Some of the bonafide samples are mis-classified as spoof due to low image resolution and lighting variations.

For the spoof samples, the mis-classification could be attributed to the adverse change in lighting conditions.

Samples in O have higher resolution compared to the other datasets as shown, and this could be attributed to mis-classifying spoof as real.

The real samples being mis-classified as spoof is either due to **a) Pixelization, b) extreme pose changes, or c) darker lighting conditions**

5.Conclusion

(+) Vision-language pre-training (e.g., CLIP) have excellent generalization compared to their counterparts trained only on images. (ability for the face anti spoofing task)

(+) 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.

(+) Text encoder further boosts generalizability.

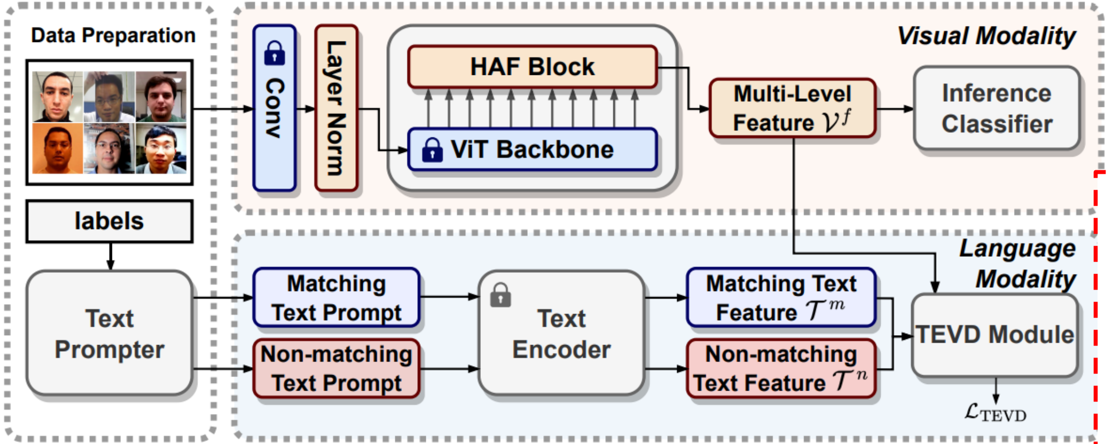
(+) Multimodal contrastive learning also enhances the generalizability across data regimes and domain gaps

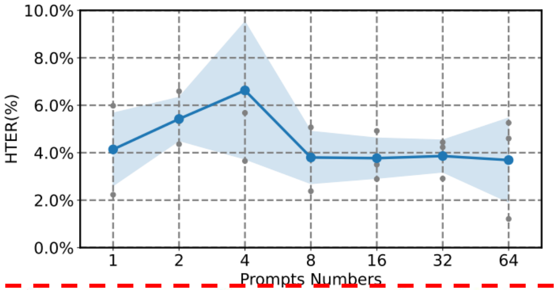
(-) The additional computational overhead involved in invoking the text encoder during training

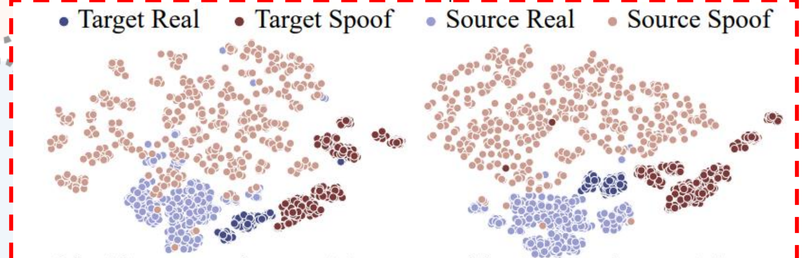
(-) To explore if these conclusions hold for other VLP foundation models

(-) Prompt learning is also a potential way to further improve performance

※ FAS with prompt engineering







(a) without textual supervision. (b) with texual supervision.

Text Prompts For Real Faces In Prompt Library

1. A real photo of a person
2. A genuine image of the person
3. An actual snapshot of the individual
4. A real-life photograph of the person
5. A true-to-life photo of the person
6. An authentic photograph of the individual
7. A bona fide picture of the person
8. An unedited photograph of the individual
9. A legitimate snapshot of the person
10. An original photo depicting the individual
11. A veritable image of the person
12. An unaltered photograph of the individual

52. An untouched, true-to-form photo
53. A straightforward, unenhanced sı
54. An unedited, authentic image of tl
55. A clear-cut, unadulterated photo
56. A non-altered, genuine representa
57. A raw, unfiltered capture of the pe
58. An unadorned, straightforward pi
59. A pure, unvarnished image of the
60. An honest, unprocessed photograp
61. A direct, undistorted snapshot of
62. An unedited, clear depiction of th
63. A truthful, unaltered photo of the
64. An unenhanced, natural picture of

Text Prompts For Print Attacks In Prompt Library

1. A printed photo
2. A print attack photo
3. A printed photo with is blur and lack of details
4. A photo of an A4 paper with a face printed on it
5. An image of an A4 sheet of paper bearing a printed face
6. A photo of a paper printed with an image of a person
7. A hard copy of a photograph
8. A printed image on paper
9. A blurred and indistinct printed photo
10. A photograph depicting a face on an A4 sheet
11. An A4 paper with a facial image printed upon it
12. A paper bearing a printed photograph of an individual
13. A printout of a photo with reduced clarity and detail

51. A photo print showing ink smears
52. A printed image with a noticeable paper texture
53. A photograph print with uneven ink distribution
54. A printed photo, showing reduced dynamic range
55. A face printout on textured paper
56. A photo print with a yellowish tint
57. A printed image of a person, cropped awkwardly
58. A facial photo printed with low ink levels
59. A digitally printed face with artifacts
60. A printout of a photo with a watermark
61. A printed photograph, slightly torn at the edge
62. A print of a digital photo, with color bleeding
63. A photo printed on thin, low-quality paper
64. A printout of a face, showing digital noise

Text descriptions across different domains can be leveraged to bridge the gap between various visual domains

MU, Lianrui, et al. TeG-DG: Textually Guided Domain Generalization for Face Anti-Spoofing. arXiv 2023.11.30

Thanks

Any Questions?

You can send mail to
Susang Kim(healess1@gmail.com)