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

ICCV 2023

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

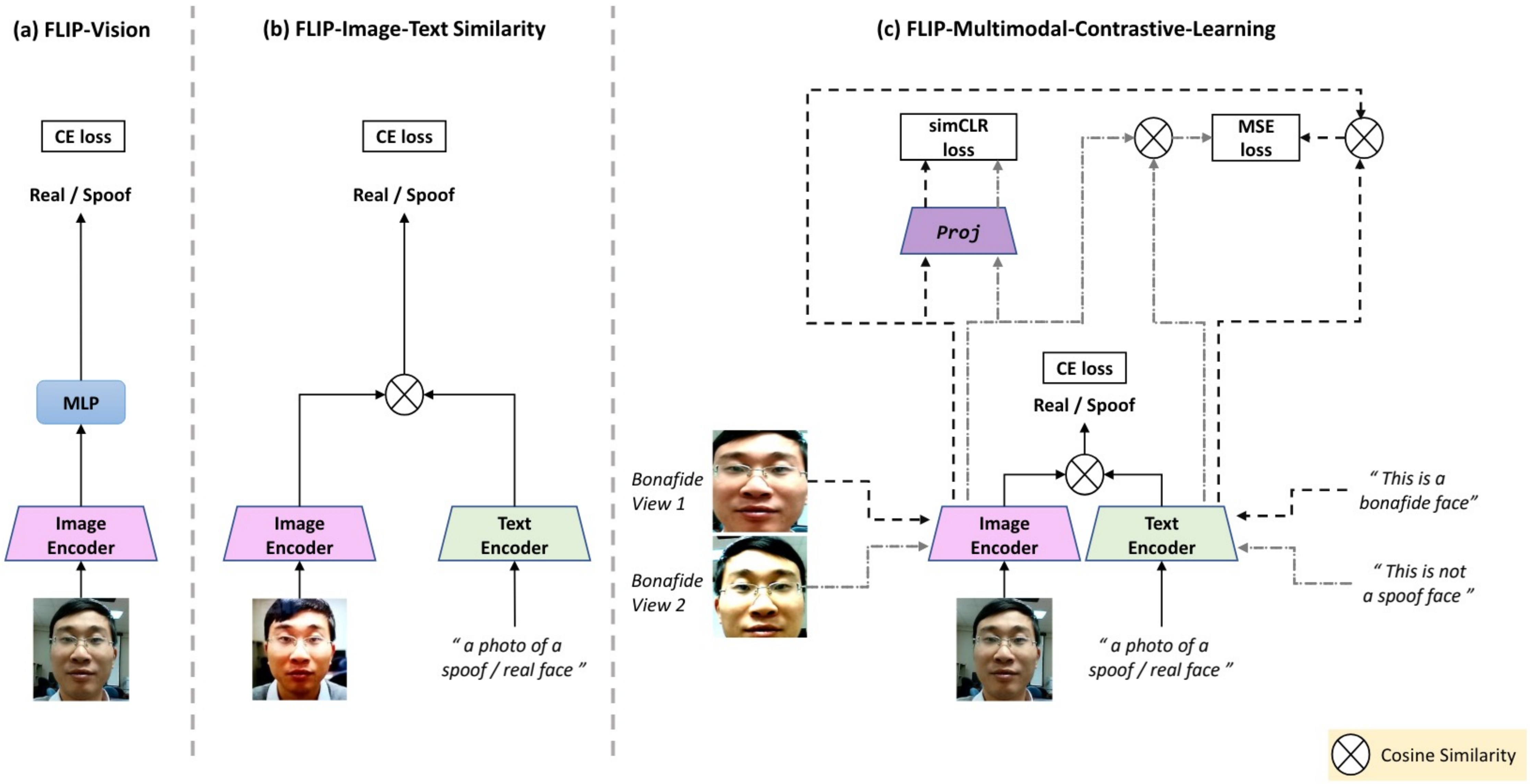
Introduction

- Presentation attack instruments (PAI) such as printed photos, replayed videos, or 3D synthetics masks
- Existing Face anti-spoofing (FAS) methods fail to generalize well
  - (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
  - (b) FAS benchmark datasets have **limited training data**, causing the model to overfit to the source domain
- 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.

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Framework



(a) FLIP-Vision
CE loss
Real / Spoof
MLP
Image Encoder
(b) FLIP-Image-Text Similarity
CE loss
Real / Spoof
Image Encoder
Text Encoder
" a photo of a spoof / real face "
(c) FLIP-Multimodal-Contrastive-Learning
simCLR loss
MSE loss
Proj
CE loss
Real / Spoof
Bonafide View 1
Bonafide View 2
Image Encoder
Text Encoder
" This is a bonafide face"
" This is not a spoof face "
" a photo of a spoof / real face "
Cosine Similarity

Outline

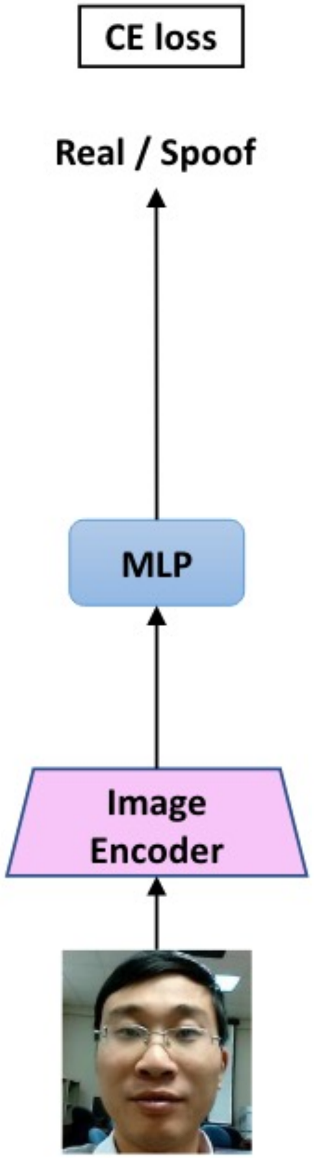
- Preliminary
- Framework
- Method
- Experiment
- Conclusion

FLIP-Vision

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

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

(a) FLIP-Vision



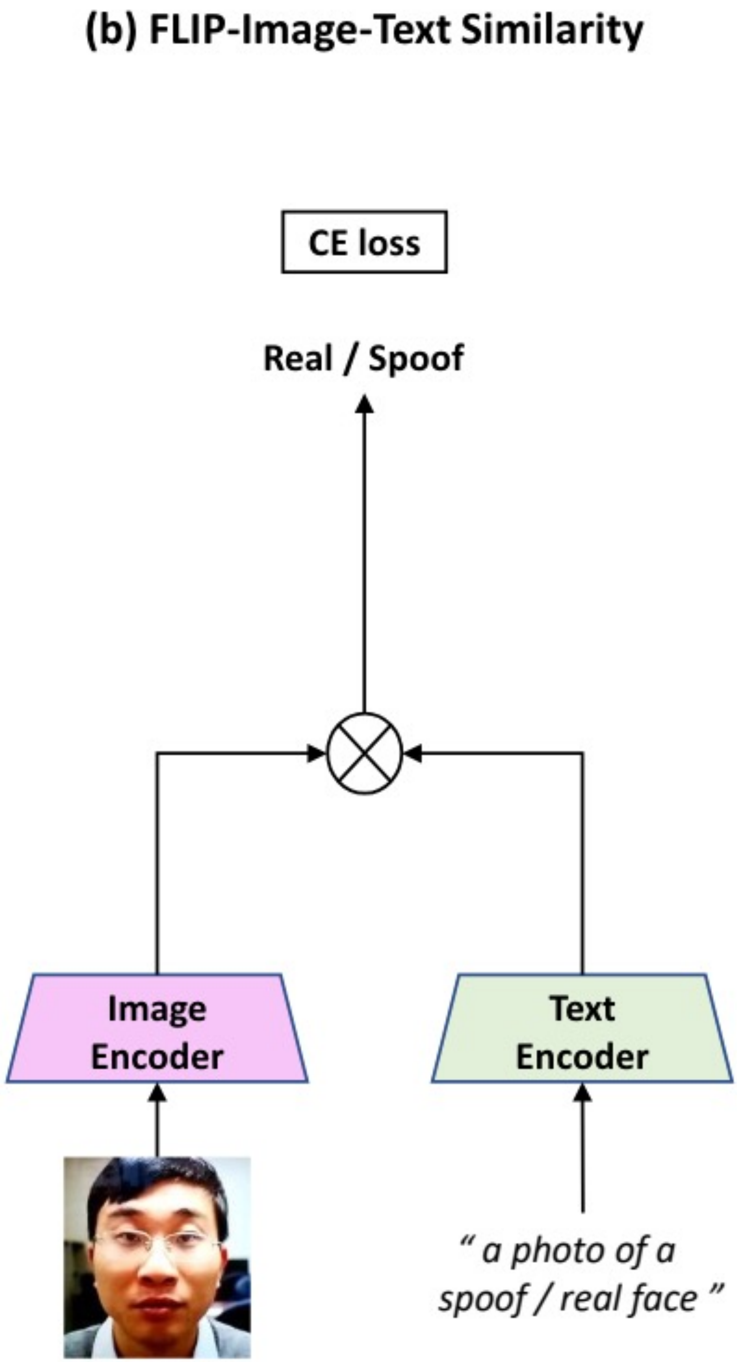
FLIP-Image-Text Similarity

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

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

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

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



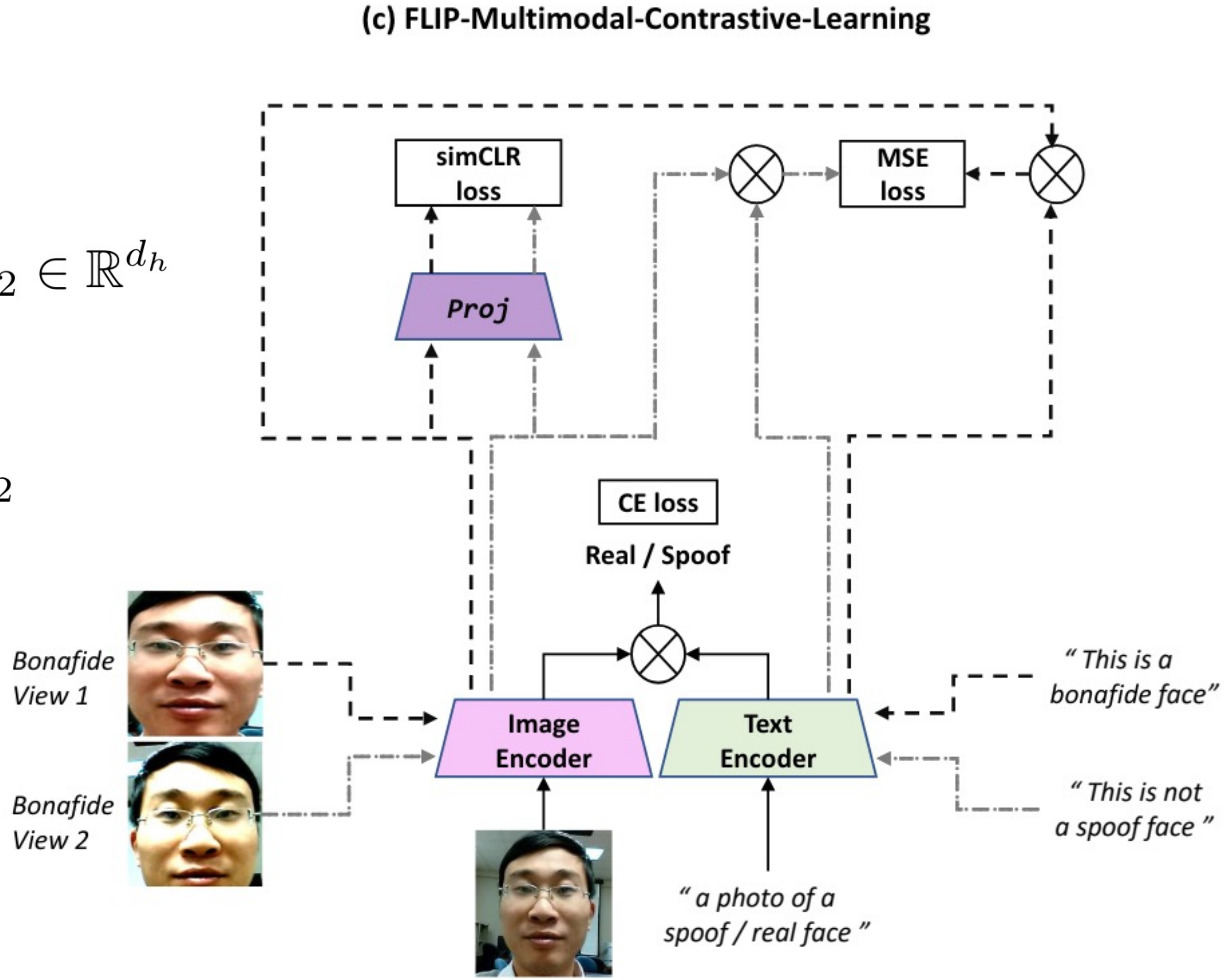
FLIP-Image-Text Similarity

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

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

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

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



Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

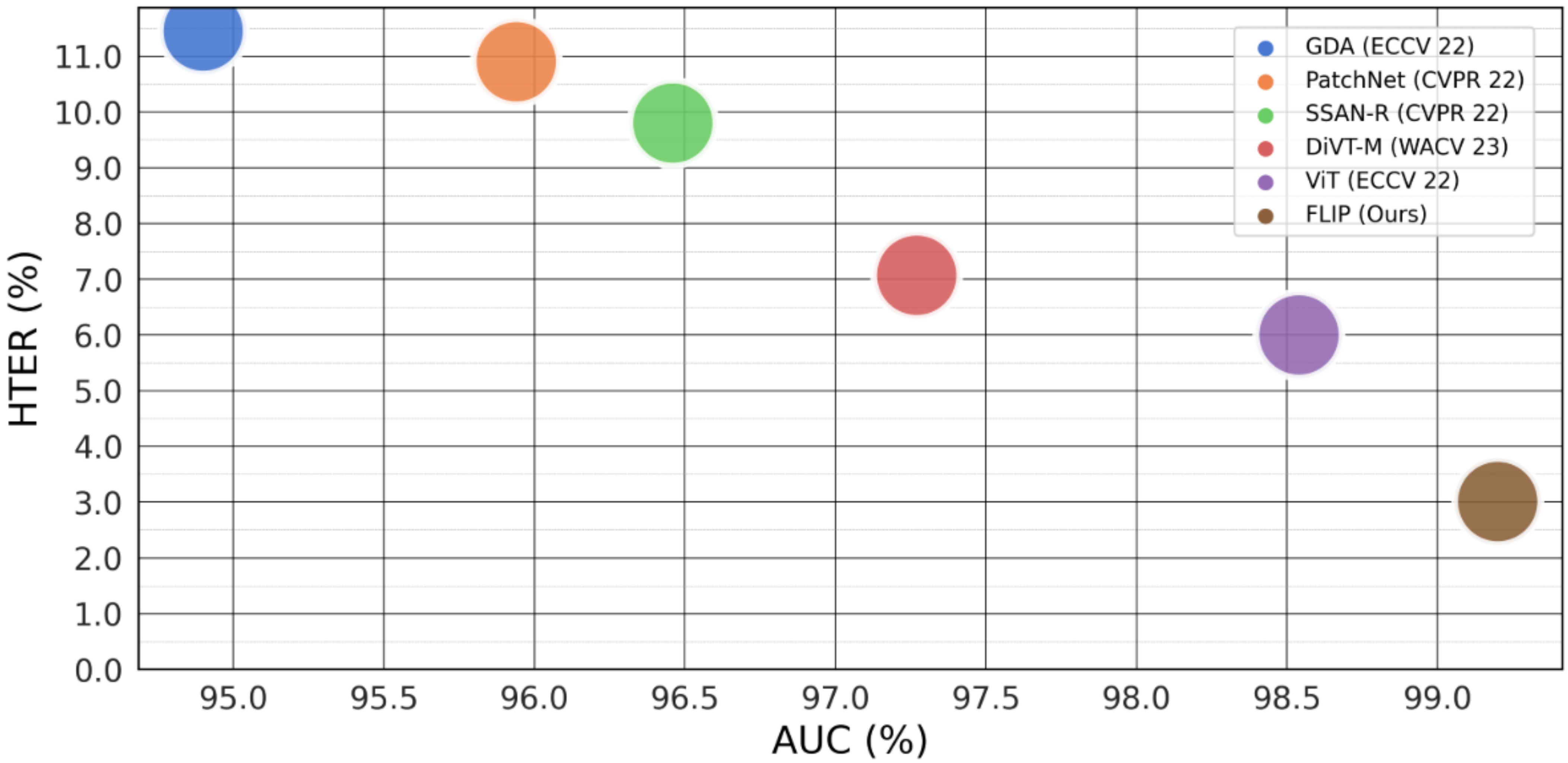
Experimental Setup

- Each dataset is considered as a domain
- Protocol 1
  - 3-source-to-single-target domain
- Protocol 2
  - Large scale 3-source-to-single-target domain
- Protocol 3
  - single-source-to-single-target domain

Evaluation of cross-domain performance

	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**

Evaluation of cross-domain performance



Ablation Studies

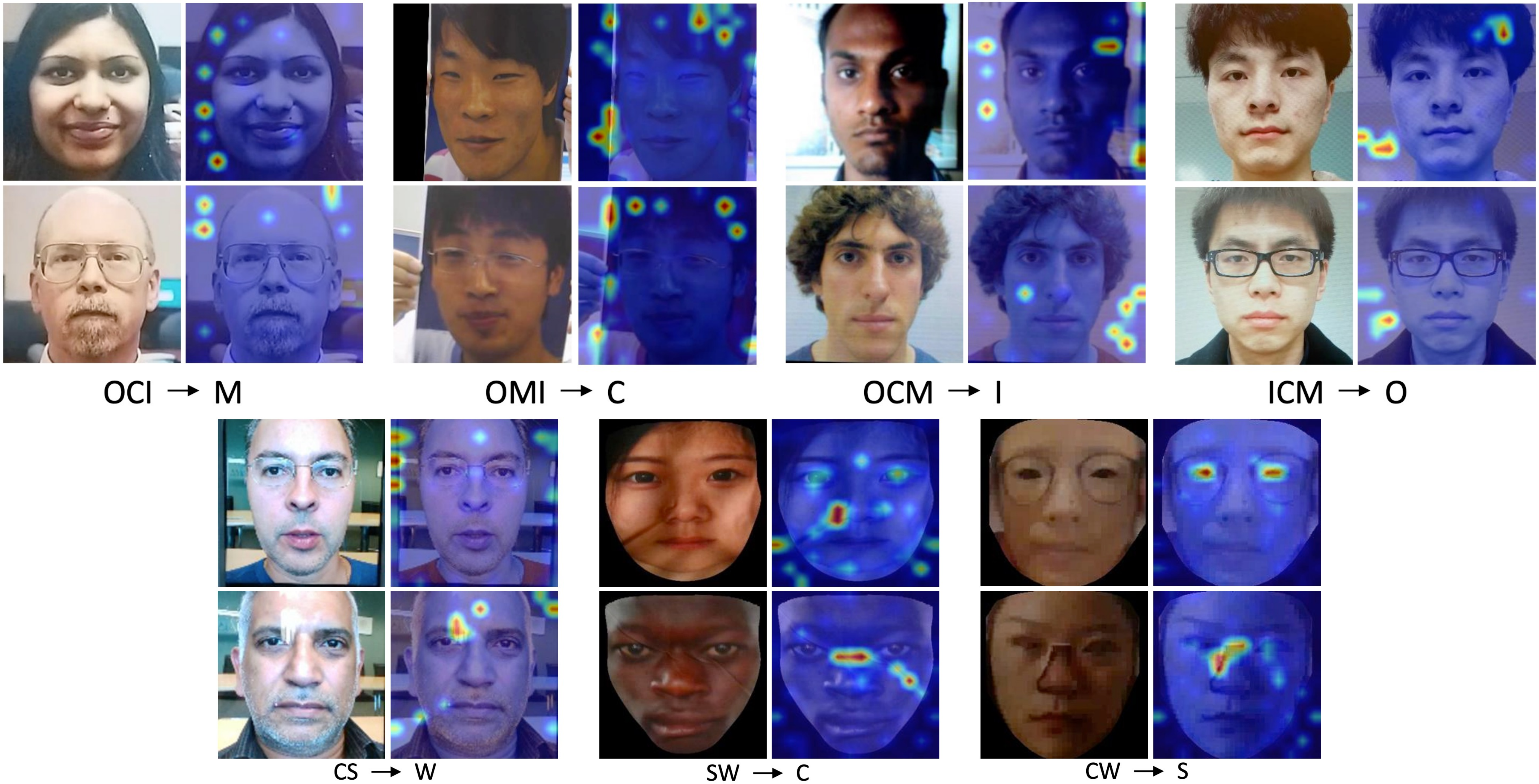
Method	OCI → M		OMI → C		OCM → I		ICM → O		Avg.
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC	HTER
Scratch	18.32	87.36	40.05	61.13	19.22	88.15	29.72	73.66	25.86
BeIT [1]	4.73	98.46	7.86	96.62	13.51	92.42	15.19	91.95	8.70
ImageNet [16]	**1.58**	**99.68**	5.70	98.91	9.25	97.15	7.47	98.42	6.00
CLIP (FLIP-V)	3.79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

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

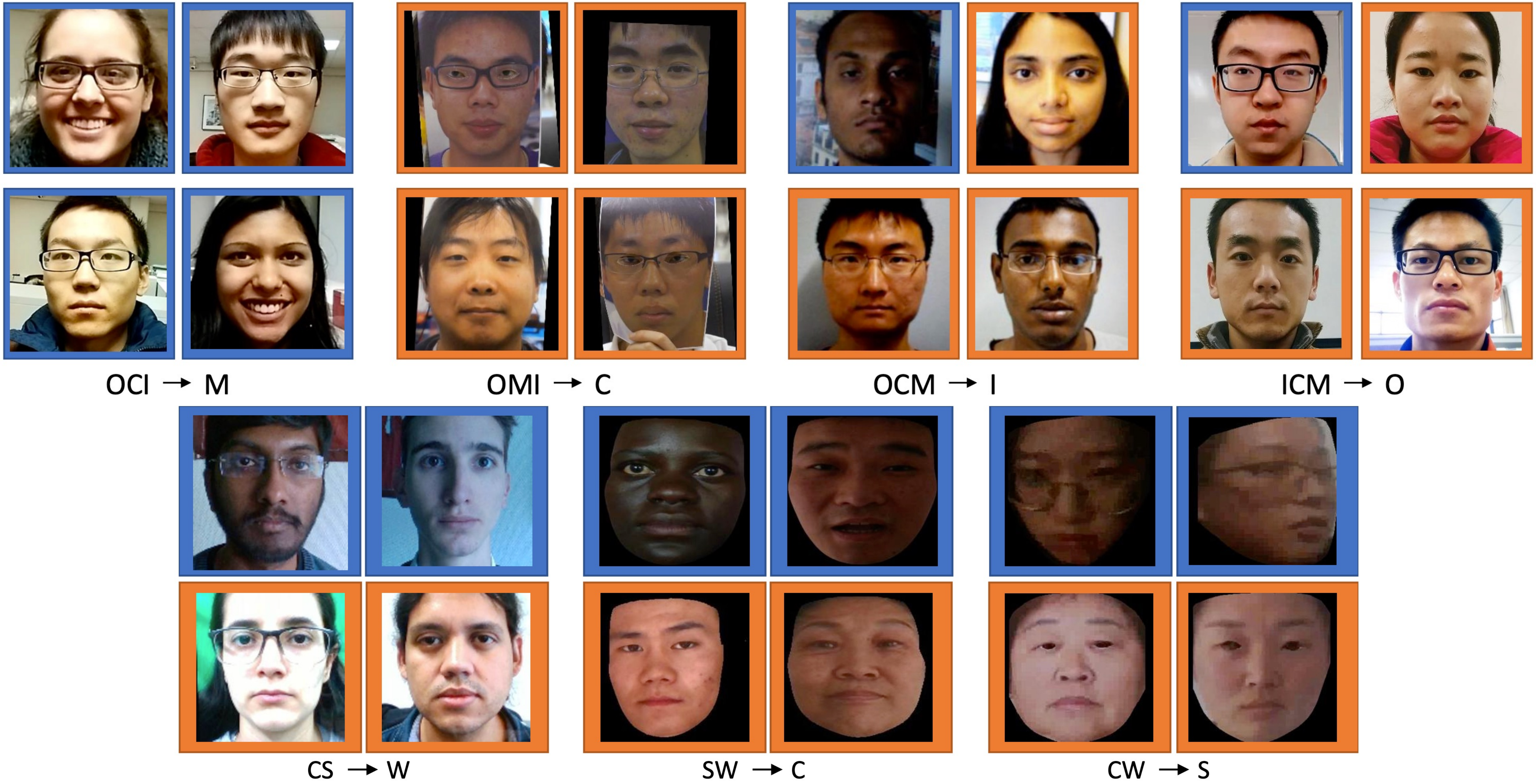
$$L_{mcl} = \alpha L_{ce} + \beta L_{simCLR} + \gamma L_{mse}$$

(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

Visualization of Attention maps



Visualization of Mis-match class



Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Conclusion

- Vision-language pre-training (e.g., CLIP) have **excellent generalization ability** for the face anti- spoofing task

- **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.