Paper Review

MLCF

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

Ajian Liu(MAIS, CASIA, China), et al.

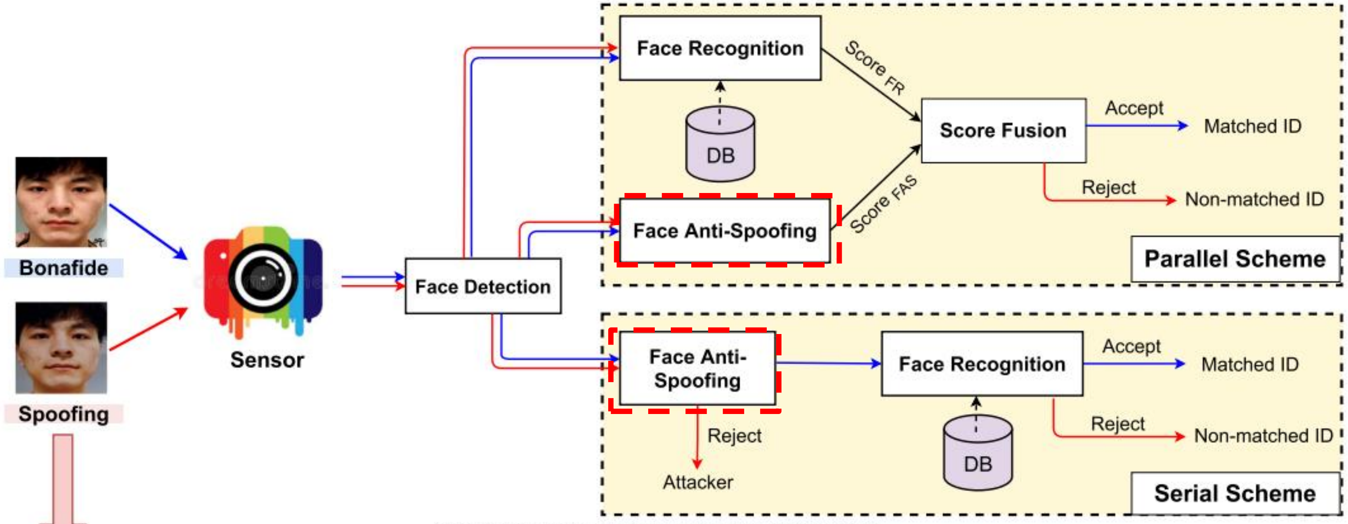
CVPR 2024

Reviewed by Susang Kim

Contents

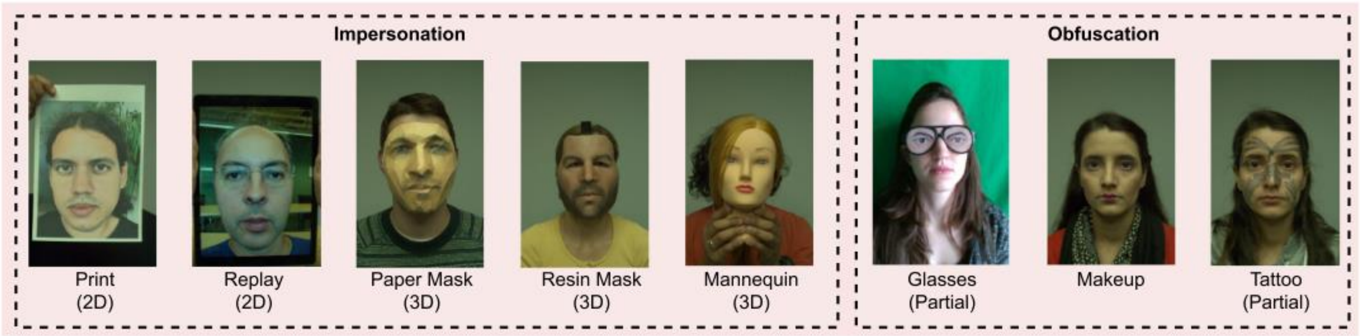
1. Introduction
2. Related Works
3. Methods
4. Experiments
5. Conclusion

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.

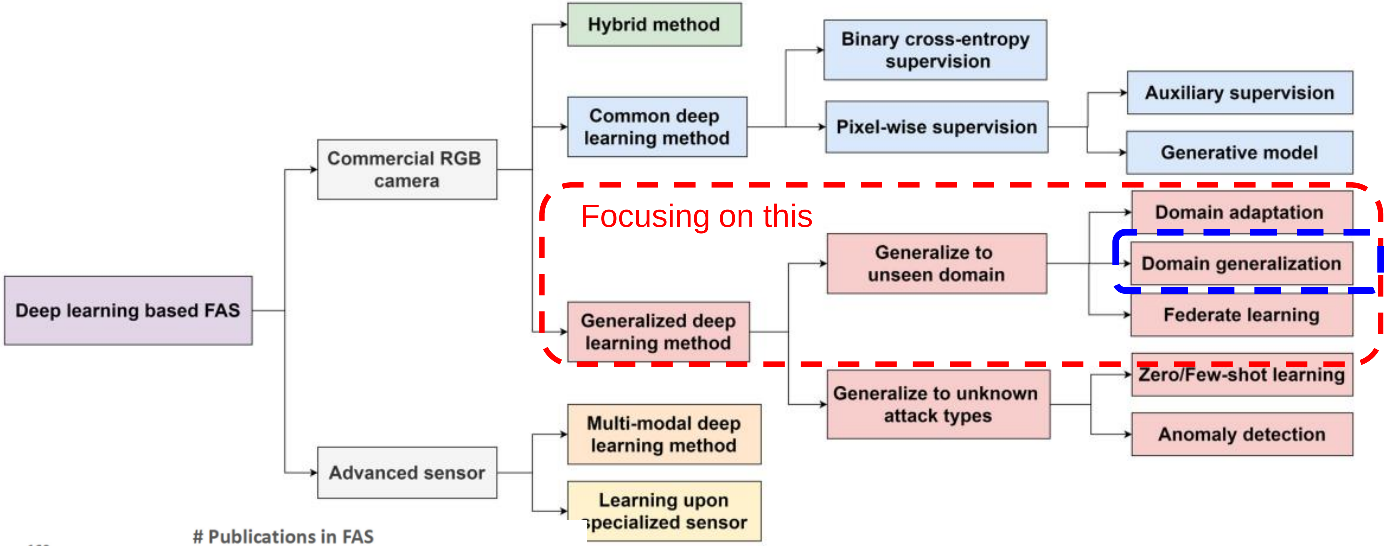


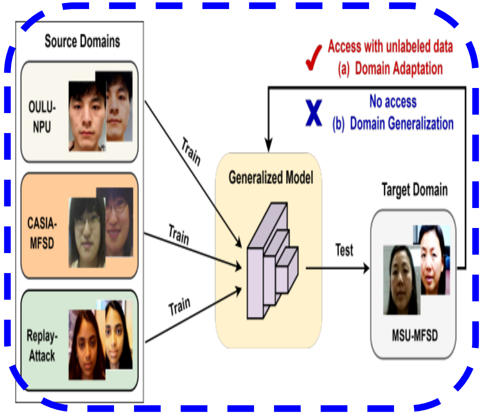
(b) Face spoofing attacks

(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





+Multi-Modal(Text prompt)

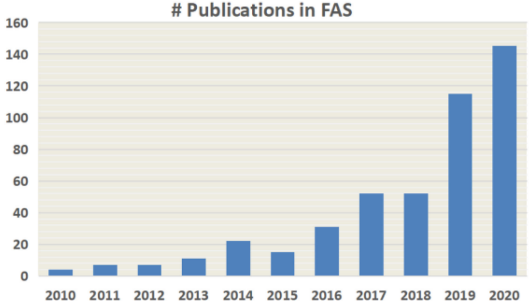


Fig. 1. The increasing research interest in the FAS field, obtained through Google scholar search with key-words: allintitle: "face anti-spoofing", "face presentation attack detection", and "face liveness detection".

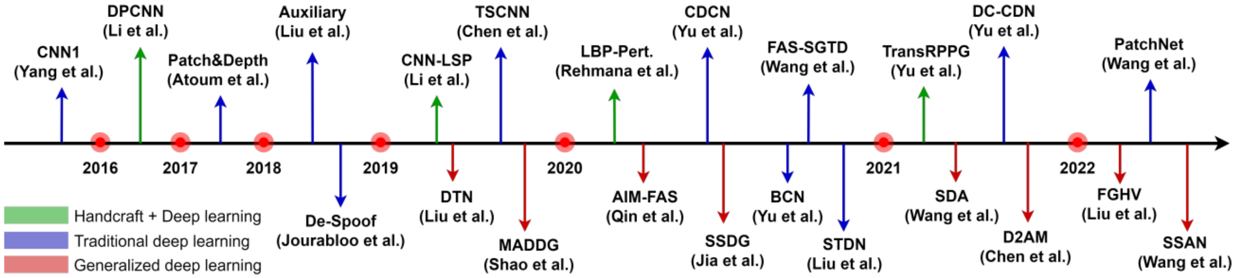


Fig. 6: Chronological overview of the milestone deep learning based FAS methods using commercial RGB camera.

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

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

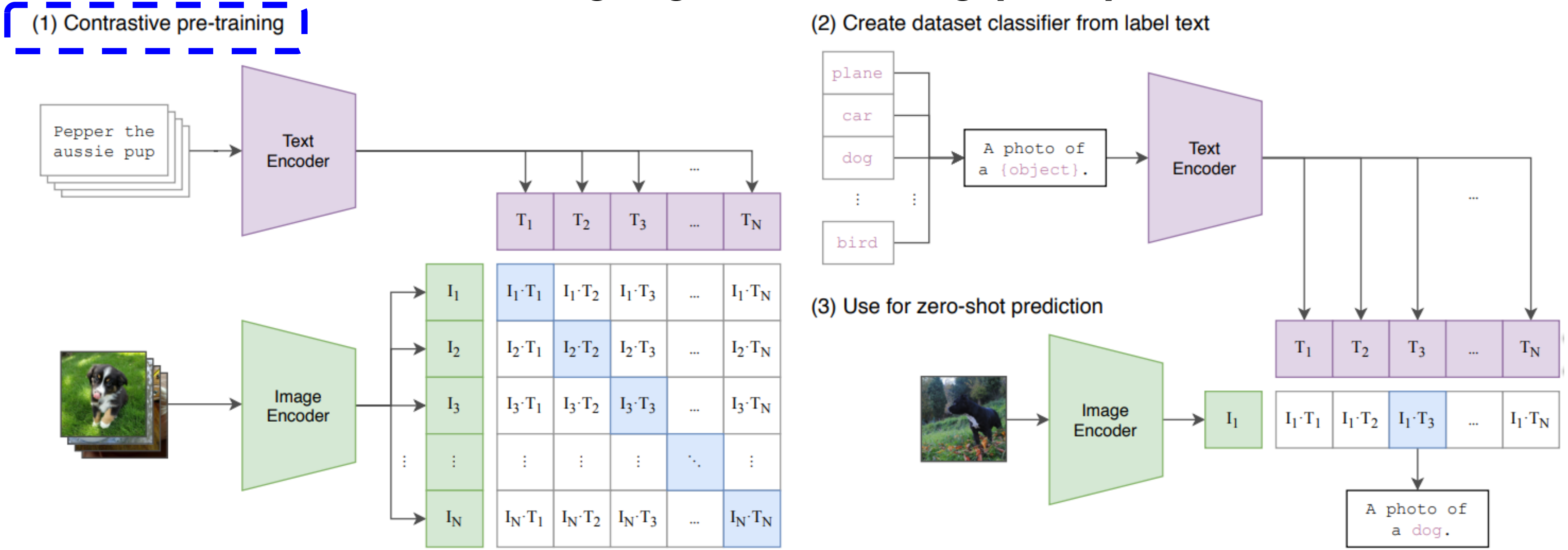


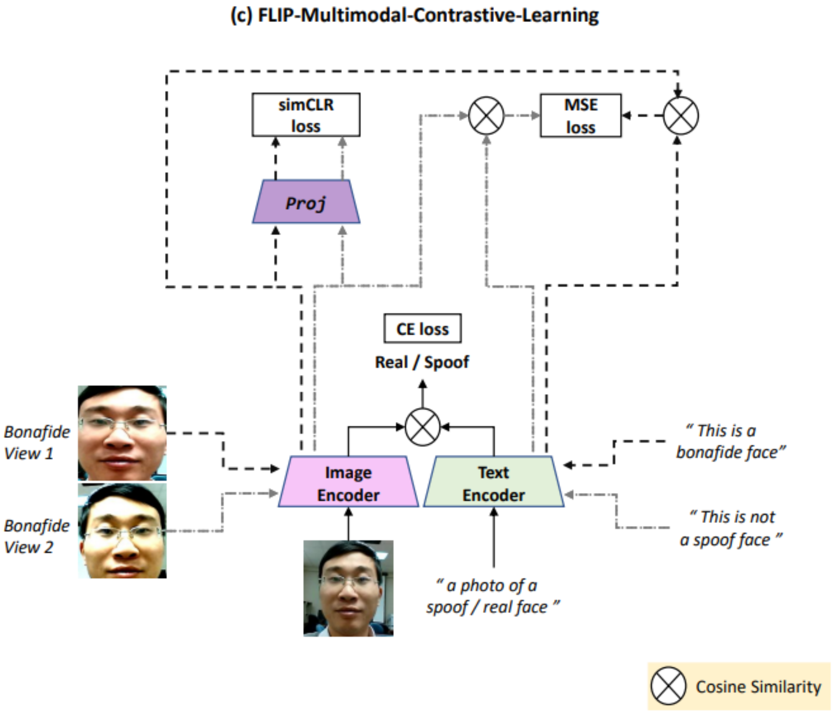
Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the 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.

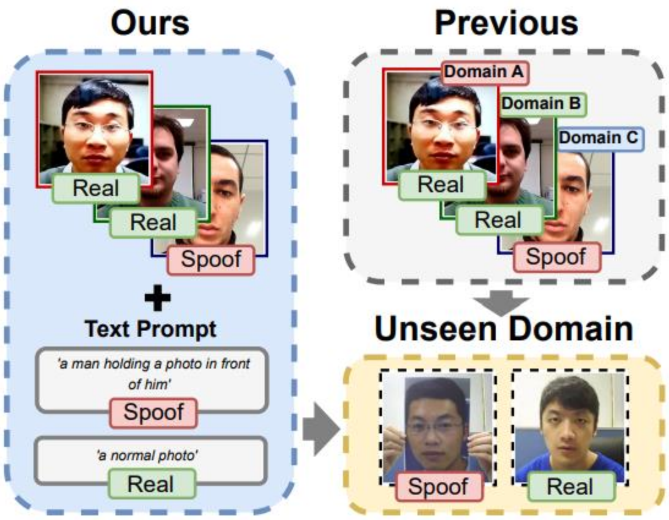
Radford, Alec, et al. "Learning transferable visual models from natural language supervision." ICML 2021.

2.Related Works - Spoofing with Vision Language Model (CLIP)

Textual information to improve the generalization ability of FAS



FLIP framework for cross-domain face anti-spoofing.



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

Srivatsan et al, FLIP: Cross-domain Face Anti-spoofing with Language Guidance. ICCV 2023
MU, Lianrui, et al. TeG-DG: Textually Guided Domain Generalization for Face Anti-Spoofing. arXiv 2023.11.30

2.Related Works - Learning to Prompt for Vision-Language Models (CVPR 2022)

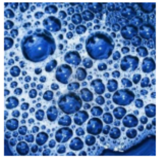
Prompt engineering vs Context Optimization (CoOp)

Caltech101 Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

(a)

Flowers102 Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**94.51**

(b)

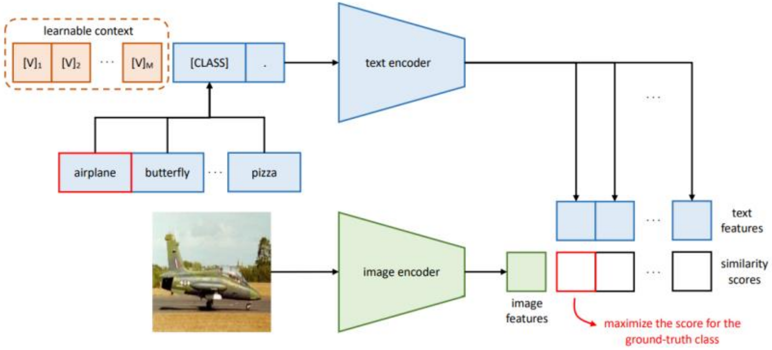


Describable Textures (DTD) Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**63.58**

(c)

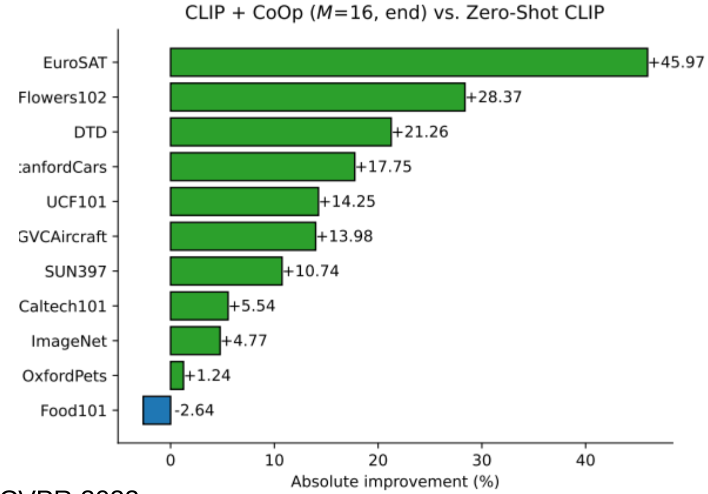
EuroSAT Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**83.53**

(d)



$$t = [\mathrm{V}]_1[\mathrm{V}]_2 \ldots [\mathrm{V}]_M[\mathrm{CLASS}], \quad (2)$$

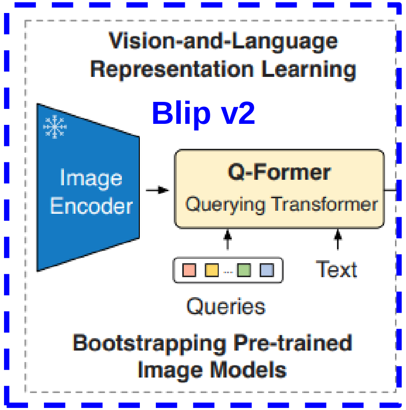
where each $[\mathrm{V}]_m$ $(m \in \{1, \ldots, M\})$ is a vector with the same dimension as word embeddings (i.e., 512 for CLIP), and M is a hyperparameter specifying the number of context tokens.

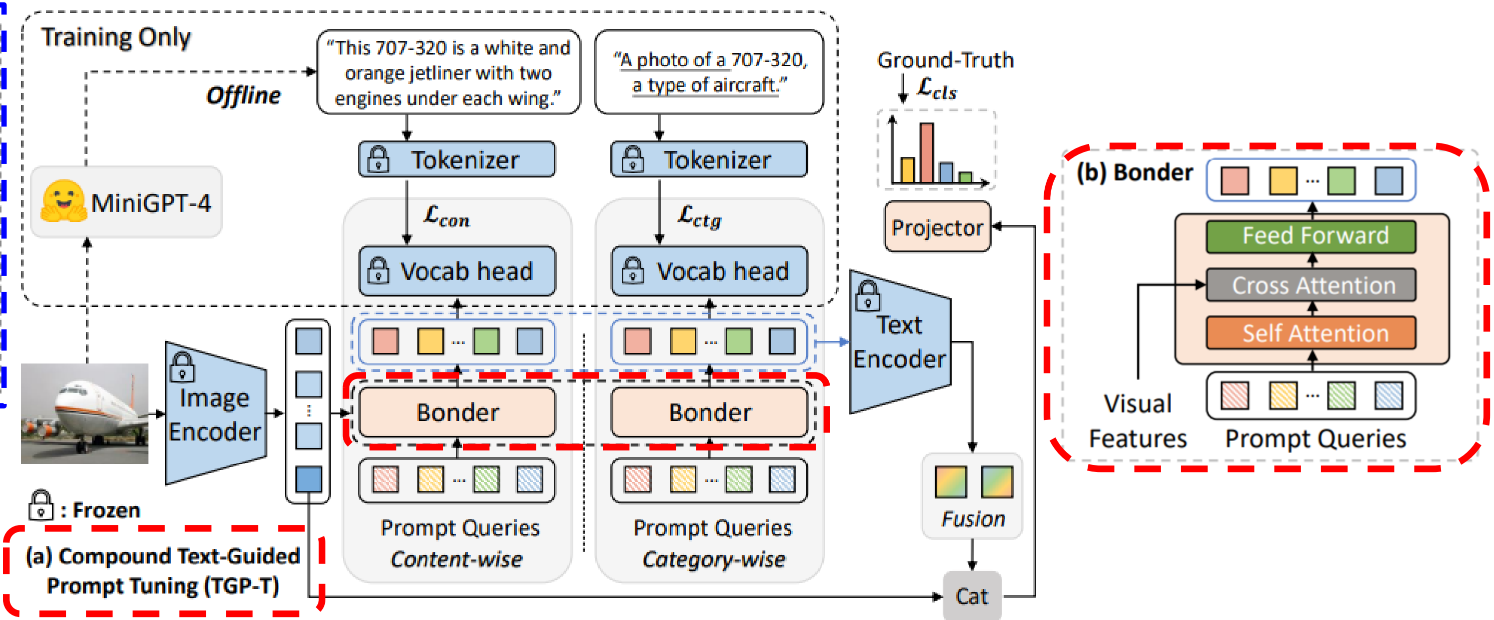


Zhou, Kaiyang, et al. "Learning to prompt for vision-language models.“ IJCV 2022.
Zhou, K., Yang, J., Loy, C. C., & Liu, Z. Conditional prompt learning for vision-language models. CVPR 2022.

2.Related Works - Learning to Prompt for Vision-Language Models (CVPR 2022)

We propose BLIP-2, a new vision-language pre-training method that bootstraps from frozen pre-trained unimodal models. **In order to bridge the modality gap, we propose a Querying Transformer (Q-Former) for vision-language representation learning**.

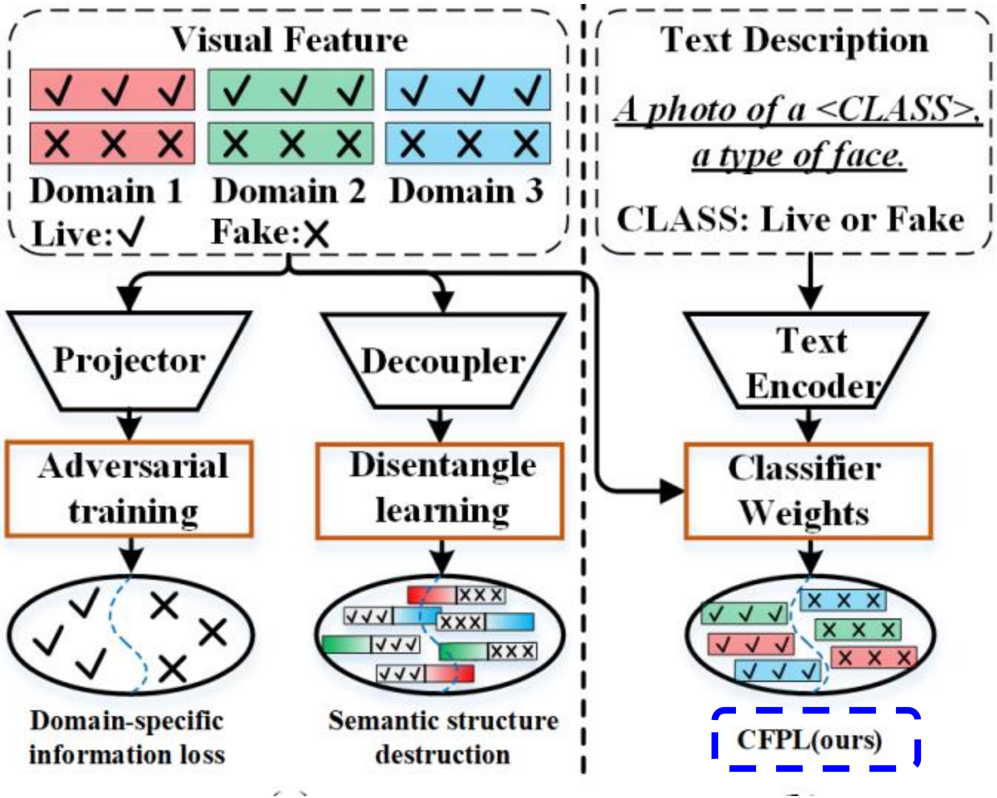




We found that compound text supervisions, i.e., **category-wise and content-wise**, are highly effective. Since they provide inter-class separability and capture intra-class variations, respectively.

LI, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." ICML 2023.
Tan, Hao, et al. "Compound text-guided prompt tuning via image-adaptive cues." AAAI 2024.

3.Method - Comparison with existing DG FAS methods

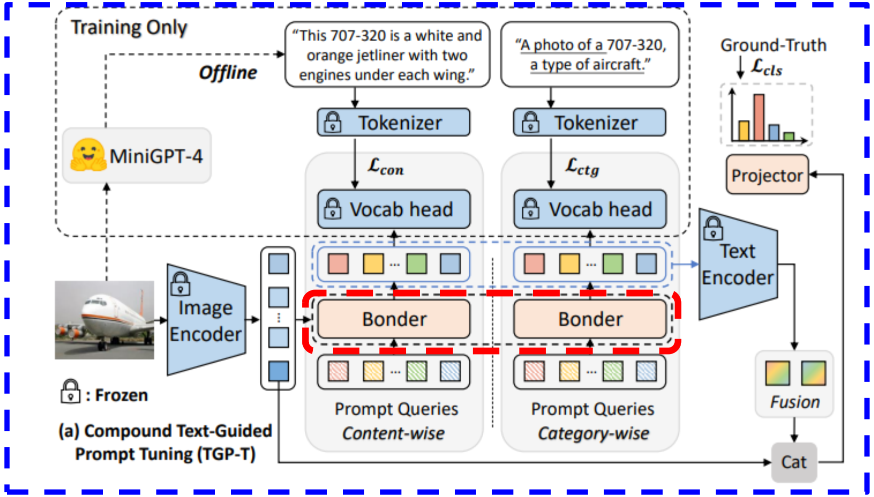


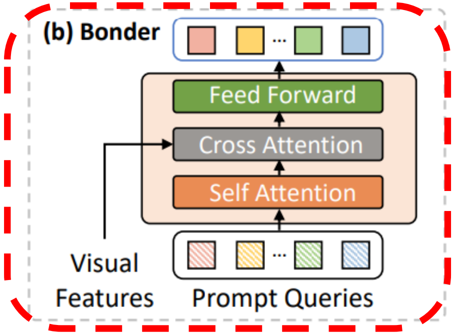
The previous methods either rely on a projector to align domain-invariant feature spaces with adversarial training or disentangle generalizable features from the whole sample with a decoupler, which inevitably leads to the distortion of semantic structures and achieves limited generalization.

CFPL framework is built on CLIP to learn generalized visual features by using the text features as weights of the classifier.

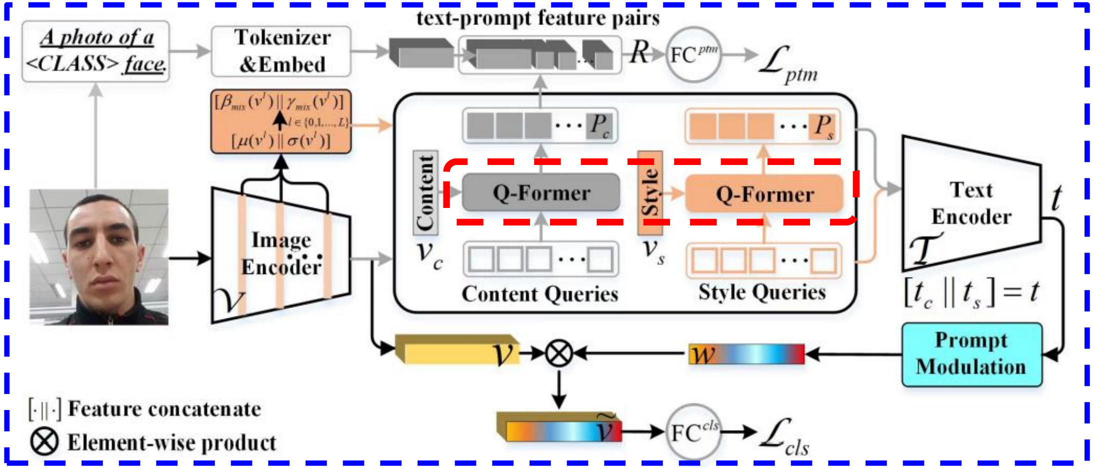
Text : A photo of a {real/fake}, a type of face.

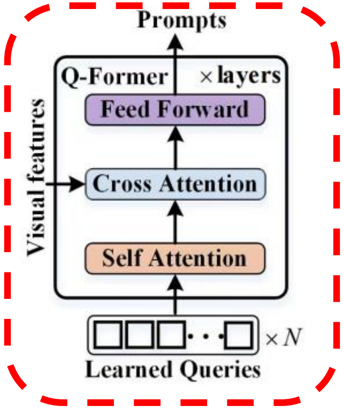
3.Method - Inter-class separability and capture Intra-class variations



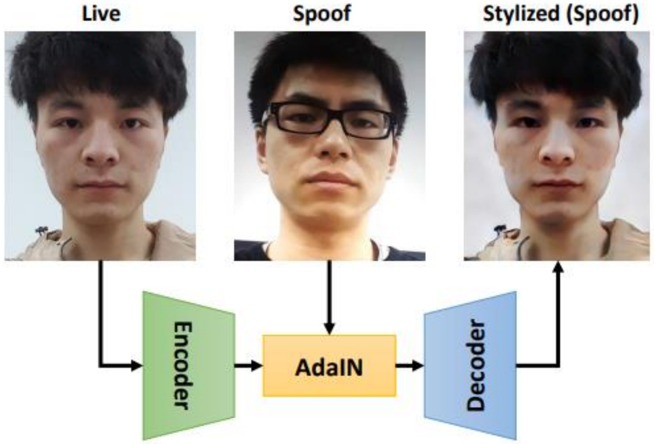


<u>Inspired by BLIP-2 and TGPT</u>, we design two lightweight transformers CQF and SQF, to learn the expected prompts conditioned on content and style features by using a set of learnable query vectors, respectively.





3.Method - Visual Content and Style features



Content Style

$$\mathrm{AdaIN}(x, y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y)$$

μ(·) and σ(·) represent channel-wise mean and standard deviation

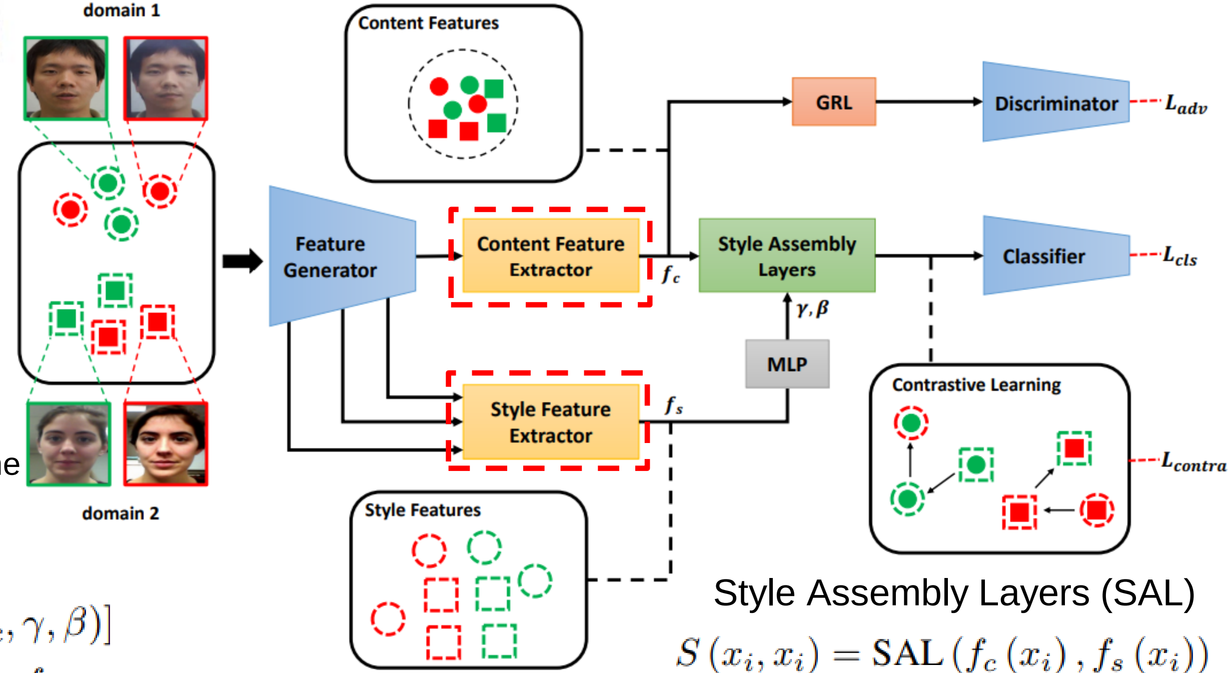
K1 and K2 = 3 × 3 convolution kernels, ⊗ is the convolution, z =intermediate variable

$$\gamma, \beta = \mathrm{MLP}\left[\mathrm{GAP}\left(f_s\right)\right],$$

$$z = \mathrm{ReLU}\left[\mathrm{AdaIN}(K_1 \otimes f_c, \gamma, \beta)\right]$$

$$\mathrm{SAL}\left(f_c, f_s\right) = \mathrm{AdaIN}(K_2 \otimes z, \gamma, \beta) + f_c,$$

Content information is semantic features and physical attributes. Style information describes domain-specific and liveness-related style information. Thus, content and style features are captured in the two-stream paths separately in our network.



Style Assembly Layers (SAL)

$$S\left(x_i, x_i\right) = \mathrm{SAL}\left(f_c\left(x_i\right), f_s\left(x_i\right)\right)$$

Wang, Zhuo, et al. "Domain generalization via shuffled style assembly for face anti-spoofing." CVPR 2022.

3.Method - Semanticized Prompts Generation (by Visual Content and Style features)

Content Q-Former (CQF) and Style Q-Former (SQF) generate content and style prompts conditioned on corresponding visual features.

$$\boldsymbol{v} \in \mathbb{R}^{d}$$

$$\mathrm{AdaIN}(x, y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y)$$

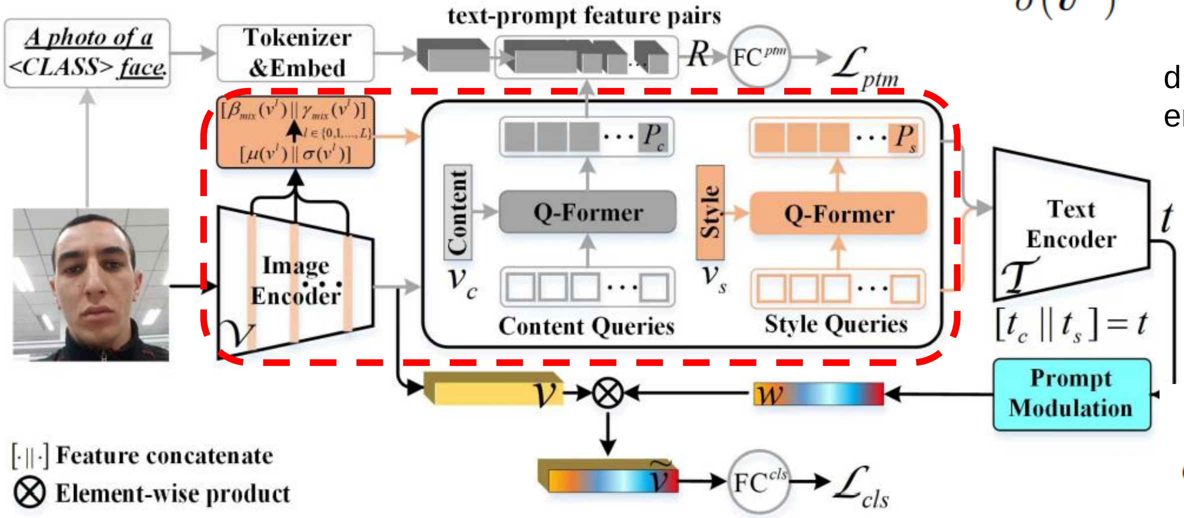
Style feature (statistics of layer)

$$\boldsymbol{v}_s = \frac{\sum_{l=1}^{L} \boldsymbol{v}_s^l}{L}, \boldsymbol{v}_s^l = \left[\mu(\boldsymbol{v}^l) \,||\, \sigma(\boldsymbol{v}^l)\right], \boldsymbol{v}_s \in \mathbb{R}^{1\times 2d}$$

Content feature (output of the image encoder)

$$\boldsymbol{v}_c = \frac{\boldsymbol{v}^L - \mu(\boldsymbol{v}^L)}{\sigma(\boldsymbol{v}^L)}, \boldsymbol{v}_c \in \mathbb{R}^{n\times d}$$

d = 512 (same dimension with multi-modal embedding space)



N learnable query embeddings

$$\boldsymbol{Q} = \left\{\boldsymbol{q}^1, \boldsymbol{q}^2, \cdots, \boldsymbol{q}^N\right\} \in \mathbb{R}^{N\times d}$$

$$\boldsymbol{Q}' = \boldsymbol{Q} + \mathrm{MSA}(\mathrm{LN}(\boldsymbol{Q})), \boldsymbol{Q}' \in \mathbb{R}^{N\times d}$$

$$\boldsymbol{Q}'' = \boldsymbol{Q}' + \mathrm{MCA}(\mathrm{LN}(\boldsymbol{Q}'), \mathrm{LN}(\boldsymbol{v})), \boldsymbol{Q}'' \in \mathbb{R}^{N\times d}$$

$$\boldsymbol{P} = \boldsymbol{Q}'' + \mathrm{MLP}(\mathrm{LN}(\boldsymbol{Q}'')), \boldsymbol{P} \in \mathbb{R}^{N\times d}$$

Content / Style prompt

$$\boldsymbol{P} = \left\{\boldsymbol{p}^1, \boldsymbol{p}^2, \ldots, \boldsymbol{p}^N\right\} \in \mathbb{R}^{N\times d}$$

3.Method - Generalized Prompt Optimization

Due to the lack of semantics for CLIP in the FAS categories, it is not suitable to align queries and text representations with the concept of maximizing their mutual information. So, the model is asked to predict whether a prompt-text pair is matched (PTM)

$$\boldsymbol{S} = \mathtt{Embed}(\mathtt{Tokenizer}(\boldsymbol{T})), \boldsymbol{S} \in \mathbb{R}^{B\times 77\times d},$$

$$\boldsymbol{S} = \mathtt{Mean\&Expand}(\boldsymbol{S}), \boldsymbol{S} \in \mathbb{R}^{B\times N\times d},$$

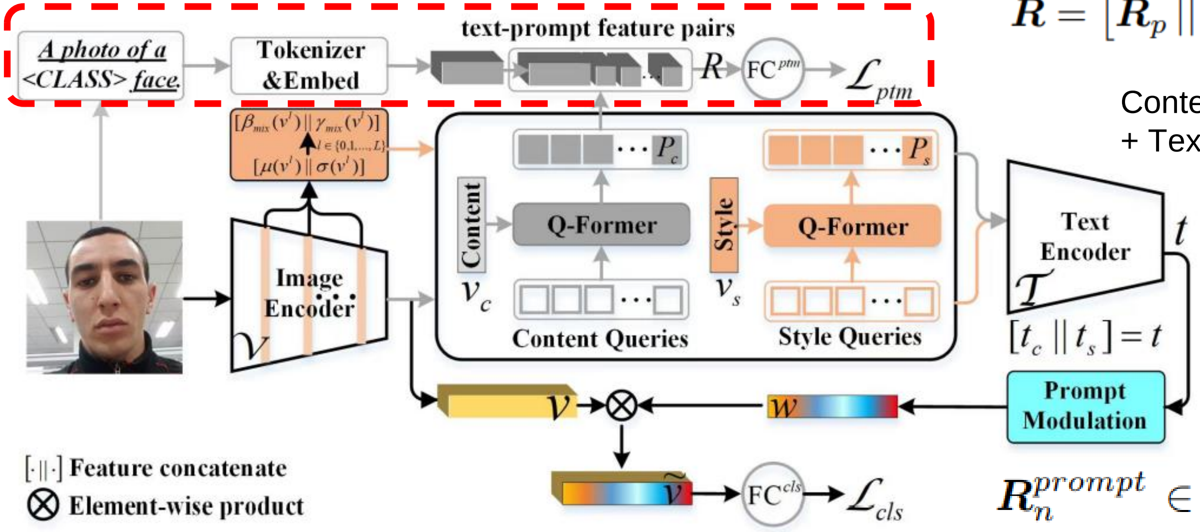
$$\boldsymbol{R}_p = [\boldsymbol{P}\,||\,\boldsymbol{S}]_2\,, \boldsymbol{R}_p \in \mathbb{R}^{B\times N\times 2d},$$

$$\boldsymbol{R} = \left[\boldsymbol{R}_p\,||\,\boldsymbol{R}_n^{prompt}\,||\,\boldsymbol{R}_n^{text}\right]_0, \boldsymbol{R} \in \mathbb{R}^{3B\times N\times 2d}$$

T = “a photo of a ⟨Rea/Fake⟩ face.”

cross-entropy loss

$$\mathcal{L}_{ptm} = \sum_{i=1}^{3B} \mathcal{H}(\boldsymbol{y}_i^{ptm}, \mathtt{Mean}(\mathtt{FC}^{ptm}(\boldsymbol{R}_i))) \qquad \boldsymbol{y}^{ptm} \in \{0, 1\}$$



Content / Style prompt
+ Text description(+class) -> Positive + Negative

Content / Style prompt

$$\boldsymbol{P} = \{\boldsymbol{p}^1, \boldsymbol{p}^2, \ldots, \boldsymbol{p}^N\} \in \mathbb{R}^{N\times d}$$

content prompt $\boldsymbol{P}_c \in \mathbb{R}^{B\times N\times d}$

negative feature pairs(prompt and text)

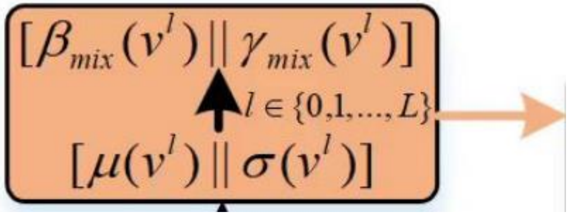
$$\boldsymbol{R}_n^{prompt} \in \mathbb{R}^{B\times N\times 2d} \text{ and } \boldsymbol{R}_n^{text} \in \mathbb{R}^{B\times N\times 2d}$$

positive feature pairs $\boldsymbol{R}_p \in \mathbb{R}^{B\times N\times 2d}$

3.Method - Diversified Style Prompt

Due to the indescribability of the sample style, we are unable to complete this task using text supervision. Implicitly, we borrow a strategy from MixStyle that mixes style feature statistics between instances to achieve diversification of style prompts.
λ is an instance-specific, random weight sampled from the beta distribution, λ ~ Beta(α, α). α is set to 0.1

$$\boldsymbol{v}_s = \frac{\sum_{l=1}^{L} \boldsymbol{v}_s^l}{L}, \boldsymbol{v}_s^l = \left[\mu(\boldsymbol{v}^l) \,||\, \sigma(\boldsymbol{v}^l)\right], \boldsymbol{v}_s \in \mathbb{R}^{1\times 2d}$$

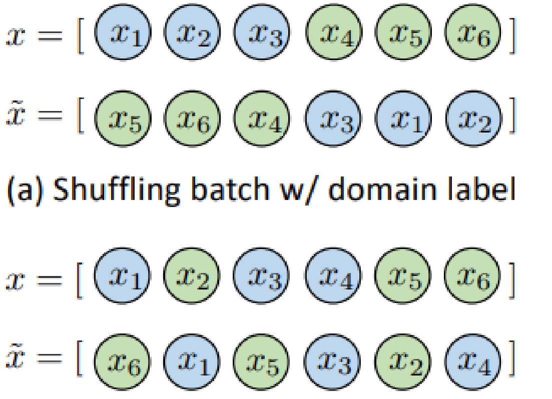


$$\lambda \in \mathbb{R}^B \quad \lambda \sim Beta(\alpha, \alpha) \quad \alpha \in (0, \infty)$$

$$\gamma_{mix} = \lambda\sigma(\boldsymbol{v}) + (1-\lambda)\sigma(\hat{\boldsymbol{v}}),$$

$$\beta_{mix} = \lambda\mu(\boldsymbol{v}) + (1-\lambda)\mu(\hat{\boldsymbol{v}})$$

$$\mathrm{MixStyle}(x) = \gamma_{mix}\frac{x-\mu(x)}{\sigma(x)} + \beta_{mix}$$



(a) Shuffling batch w/ domain label

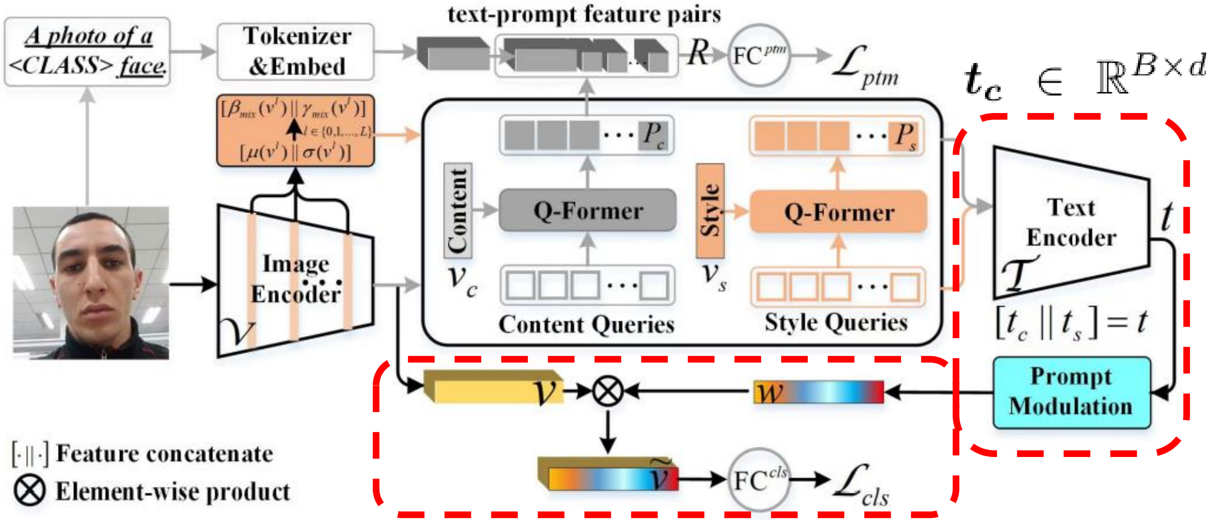
(b) Shuffling batch w/ random shuffle

Figure 2: A graphical illustration of how a reference batch is generated. Domain label is denoted by color.

ZHOU, Kaiyang, et al. Domain generalization with mixstyle. ICLR, 2021.

3.Method - Prompt Modulation on Visual Features

Due to the content and style prompts are generated based on sample instances, they are more suitable as a set of fine-tuning factors (class free) for adaptively recalibrating channel-wise visual feature responses, compared to using them as classifier's weights (with class) to predict visual feature.



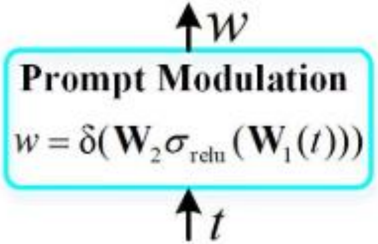
$t_c \in \mathbb{R}^{B\times d}$ and $t_s \in \mathbb{R}^{B\times d}$

$\mathbb{R}^{B\times 2d}$

$\mathbf{W}_1 \in \mathbb{R}^{\frac{d}{r}\times 2d}$

$\mathbf{W}_2 \in \mathbb{R}^{d\times \frac{d}{r}}$

$$\boldsymbol{w} = \delta(\mathrm{g}_e(\boldsymbol{t}, \mathbf{W})) = \delta(\mathbf{W}_2\sigma_{\text{relu}}(\mathbf{W}_1\boldsymbol{t})),$$



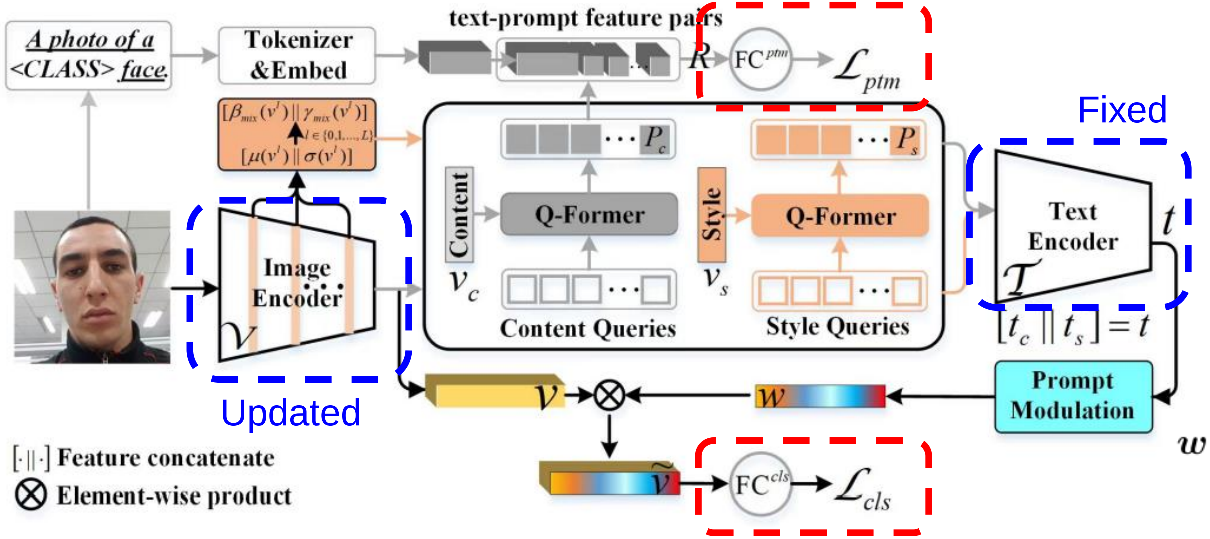
$$\tilde{\boldsymbol{v}} = \left[\tilde{\boldsymbol{v}}^1, \tilde{\boldsymbol{v}}^2, \dots, \tilde{\boldsymbol{v}}^d\right], \tilde{\boldsymbol{v}}^c = \boldsymbol{w}^c \cdot \boldsymbol{v}^c$$

$$y^{cls} \in \{0,1\} \qquad \mathcal{L}_{cls} = \sum_{i=1}^{B} \mathcal{H}(\boldsymbol{y}_i^{cls}, \mathrm{FC}^{cls}(\tilde{\boldsymbol{v}}_i)), \tilde{\boldsymbol{v}} \in \mathbb{R}^{B\times d}$$

3.Method - Model Training and Inference

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

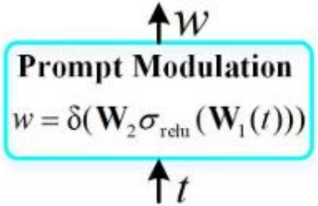
$$\mathcal{L}_{ptm} = \sum_{i=1}^{3B} \mathcal{H}(\boldsymbol{y}_i^{ptm}, \texttt{Mean}(\texttt{FC}^{ptm}(\boldsymbol{R}_i)))$$



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

$$\boldsymbol{w} = \delta(\mathrm{g}_e(\boldsymbol{t}, \mathbf{W})) = \delta(\mathbf{W}_2\sigma_{\text{relu}}(\mathbf{W}_1\boldsymbol{t})),$$

$$\mathcal{L}_{cls} = \sum_{i=1}^{B} \mathcal{H}(\boldsymbol{y}_i^{cls}, \texttt{FC}^{cls}(\tilde{\boldsymbol{v}}_i)), \tilde{\boldsymbol{v}} \in \mathbb{R}^{B\times d}$$

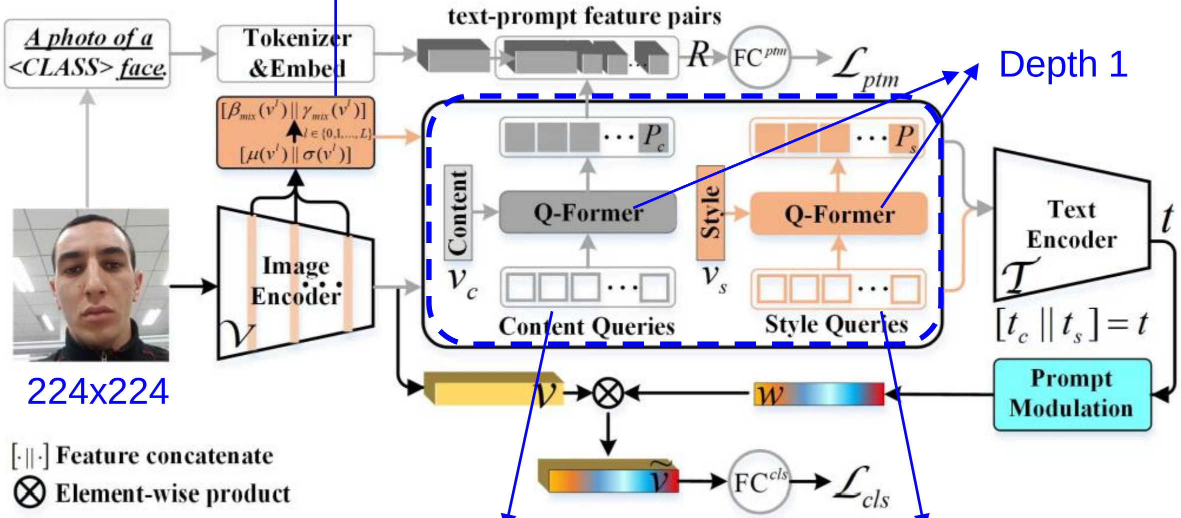


4.Experiments - Implementation Details

Style prompt diversification is activated in the training phase with a probability of 0.5 and does not participate in the test phase

Depth 1

augmented with random resized cropping and horizontal flipping



224x224

Length of style and content queries to 16
Each query has a dimension of 512
num_heads=8 (bonder code)

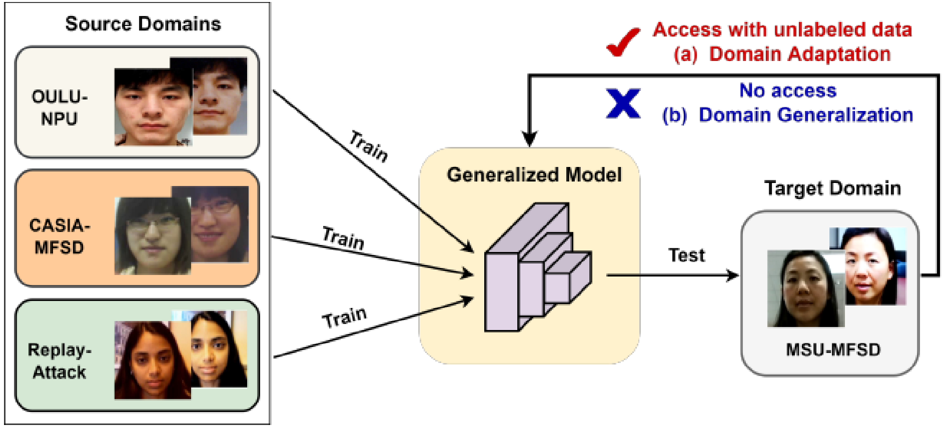
batch size of 12, Adam optimizer with a weight decay of 0.05.
The minimum learning rate at the second stage is 1e − 6. train all models with 500 epochs.

4.Experiments – Datasets and Evaluation Metrics

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)

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



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

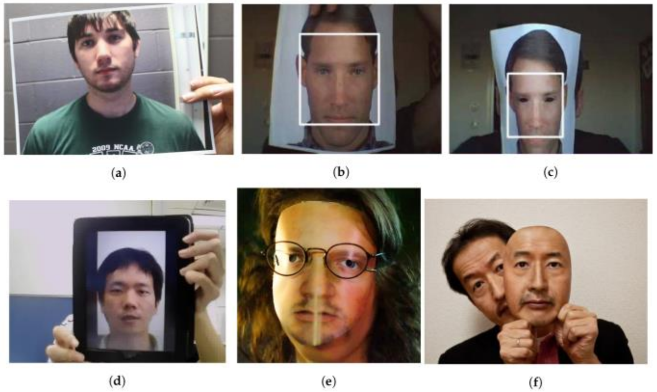
For pair comparison, CelebA-Spoof as supplementary training data to enhance the diversity of training samples.

[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.

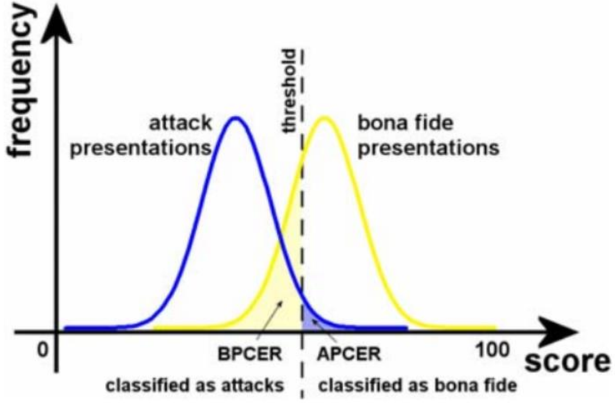
4.Experiments – Evaluation metric

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



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

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

HTER (Half Total Error Rate)

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 Results

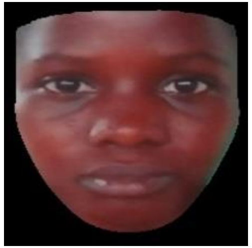
Method	OCI→M			OMI→C			OCM→I			ICM→O			avg.
	HTER↓	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER	AUC	TPR@ FPR=1%	HTER
MADDG [40]	17.69	88.06	-	24.50	84.51	-	22.19	84.99	-	27.98	80.02	-	23.09
DR-MD-Net [47]	17.02	90.10	-	19.68	87.43	-	20.87	86.72	-	25.02	81.47	-	20.64
RFMeta [41]	13.89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

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

4.Experiments - Cross-domain Results

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

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

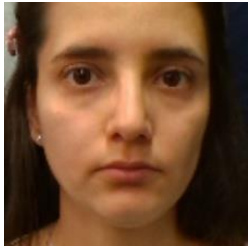

Cefa Fake


Cefa Real


WMCA Fake


WMCA Real


Surf Fake


Surf Real

4.Experiments - Ablation Study

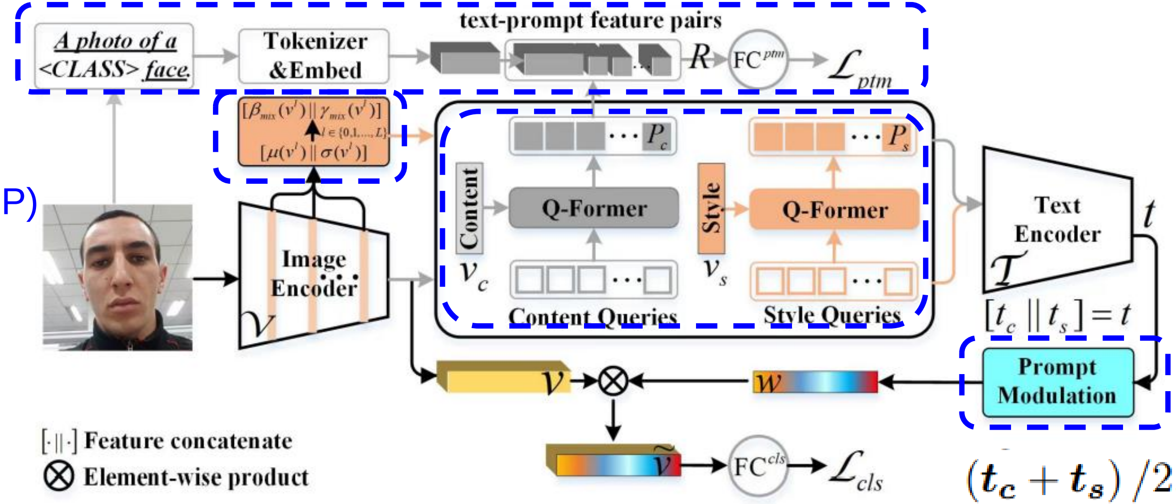
Baseline	PTM	DSP	PM	HTER(%)↓	AUC(%)	TPR(%) @FPR=1%
CoOp [61]	-	-	-	8.78	94.77	43.71
✓	-	-	-	8.11	96.09	51.59
✓	✓	-	-	7.50	96.39	54.78
✓	✓	✓	-	7.08	96.79	57.61
✓	✓	✓	✓	6.72	97.09	60.35

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

Baseline : two lightweight transformers CQF and SQF

Text Supervision (PTM)

Diversification of Style Prompt (DSP)



Prompt Modulation (PM)

$$\left(\boldsymbol{t_c} + \boldsymbol{t_s}\right)/2, \boldsymbol{w} \in \mathbb{R}^{B\times d}.$$

4.Experiments - Ablation Study

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

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

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

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

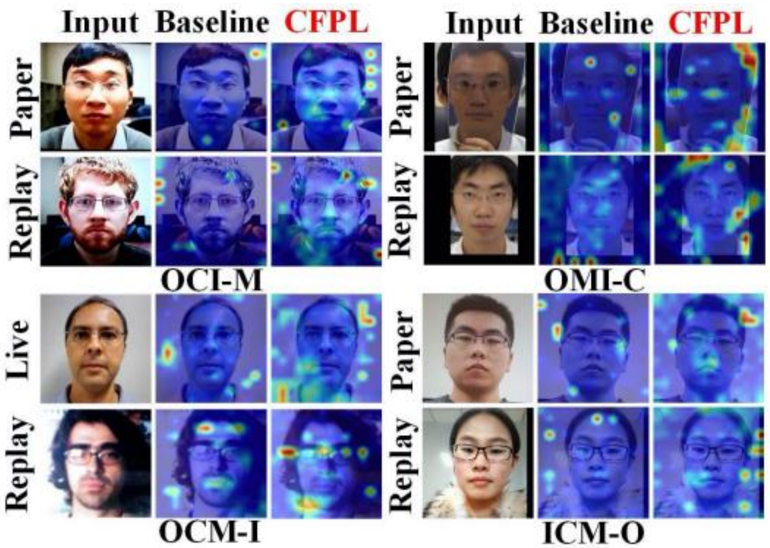


Figure 4. Using visualization tool [2], the attention maps on all sub-protocols from Protocol 1, where the Baseline caused classification errors due to its failure to detect spoofing regions, and our CFPL correctly classifies these samples by correcting the region of interest.

4.Experiments - Ablation Study

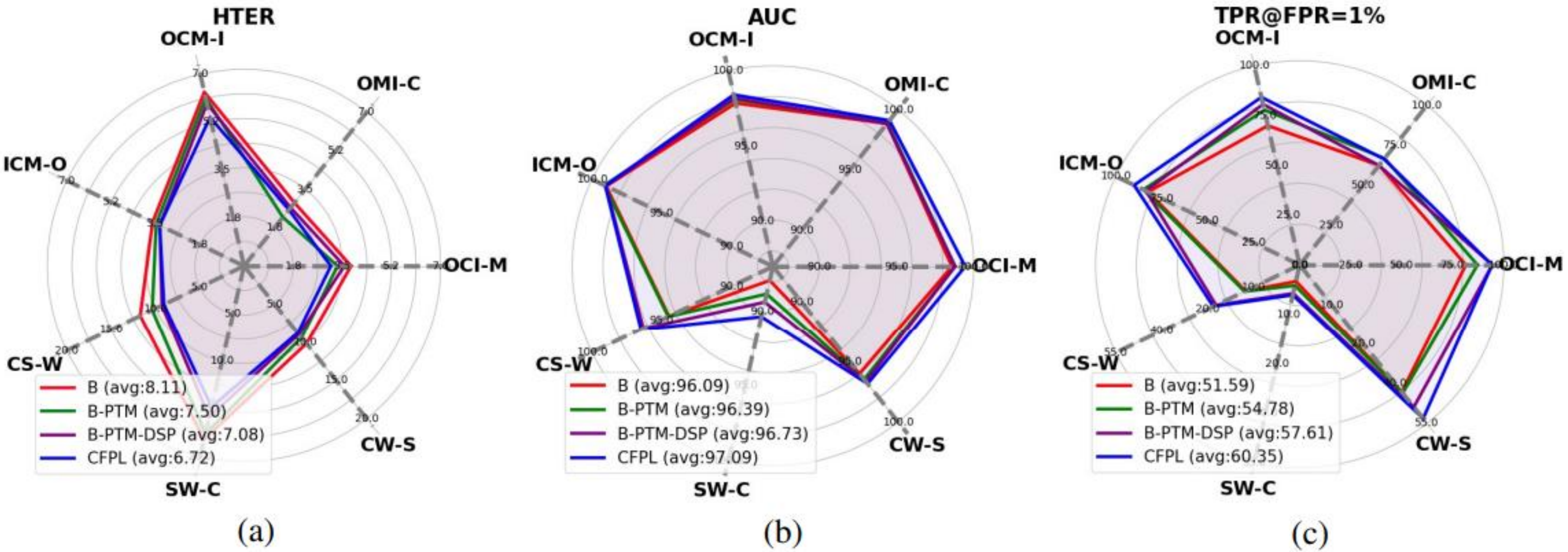


Figure 3. The results of each method on three metrics across all sub-protocols, where the red line represents the Baseline, and the blue line represents our CFPL. For the HTER metric, the smaller area enclosed by lines, the better performance of the corresponding methods. The opposite conclusion applies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

5.Conclusion

(+) Instead of directly manipulating visual features, it is the first work to explore DG FAS via textual prompt learning, which allows a broader semantic space to adjust the visual features to generalization.

(+) Diversifying style and content by text prompt modulation to promote the generalization.

(+) Propose two lightweight transformers, CQF and SQF, to learn the different semantic prompts conditioned on content and style features

(+) CFPL(PTM, DSP, PM) is effective and outperforms SOTA methods by an undeniable margin.
- PTM : Prompt-Text Matched, DSP : Diversified Style Prompt, PM : Prompt Modulation.

(-) It follows the TGPT (BLIP v2) architecture and adapted for the spoofing task.
(-) Although it involves related text supervision, it does not provide a detailed explanation of the specific spoofing cues.(“a photo of a ⟨CLASS⟩ face.”,)
(-) It is based on the CLIP architecture, and has an additional module attached to it.(CFPL)

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
FLIP-MCL* [42]	4.95	98.11	74.67	0.54	99.98	100.00	4.25	99.07	84.62	2.31	99.63	92.28	3.01
CFPL*(Ours)	1.43	99.28	98.57	2.56	99.10	66.33	5.43	98.41	85.29	2.50	99.42	94.72	2.98

Thanks
Any Questions?

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