Flexible-Modal Face Anti-Spoofing: A Benchmark

Zitong Yu[1], Ajian Liu[2], Chenxu Zhao[3], Kevin H. M. Cheng[4], Xu Cheng[5], Guoying Zhao[4]

[1]Great Bay University, China

[2]Institute of Automation, Chinese Academy of Sciences, China

[3]SailYond Technology, China

[4]University of Oulu, Finland

[5]Nanjing University of Information Science and Technology, China

Abstract

Face anti-spoofing (FAS) plays a vital role in securing face recognition systems from presentation attacks. Benefitted from the maturing camera sensors, single-modal (RGB) and multi-modal (e.g., RGB+Depth) FAS has been applied in various scenarios with different configurations of sensors/modalities. Existing single- and multi-modal FAS methods usually separately train and deploy models for each possible modality scenario, which might be redundant and inefficient. Can we train a unified model, and flexibly deploy it under various modality scenarios? In this paper, we establish the first flexible-modal FAS benchmark with the principle 'train one for all'. To be specific, with trained multi-modal (RGB+Depth+IR) FAS models, both intra- and cross-dataset testings are conducted on four flexible-modal sub-protocols (RGB, RGB+Depth, RGB+IR, and RGB+Depth+IR). We also investigate prevalent deep models and feature fusion strategies for flexible-modal FAS. We hope this new benchmark will facilitate the future research of the multi-modal FAS. The protocols and codes are available at https://github.com/ZitongYu/Flex-Modal-FAS.

1. Introduction

Face recognition has been widely used in many interactive artificial intelligence systems for its convenience (e.g., access control and face payment). However, vulnerability to presentation attacks (e.g., print, video replay, and 3D masks) curtails its reliable deployment. For the reliable use of face recognition systems, face anti-spoofing (FAS) methods [3,13,15,18–20,26,27,32,34,37] are important to detect such presentation attacks (PAs).

In recent years, plenty of hand-crafted feature based [1,2,35] and deep learning based [4,5,14,21–24,29,33,38–40] methods have been proposed for *RGB*-based single-modal FAS. On one hand, some hand-crafted descriptors with fa-
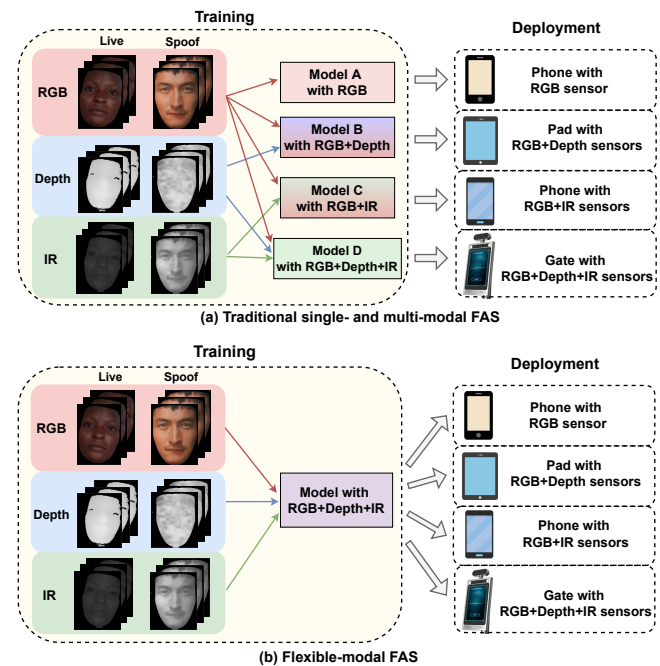

Figure 1. The training and deployment frameworks of the (a) traditional single- and multi-modal FAS; and (b) flexible-modal FAS. The former one aims at separately train and deploy powerful models for each possible modality scenario, while the latter one focuses on training a unified model for all real-world modality scenarios.

cial color texture [1] and physiological signals [35] feature representation are designed based on crucial live/spoof clues (e.g., moiré pattern, noise artifacts, and bio-signal liveness), thus are robust for live/spoof discrimination. On the other hand, deep convolutional neural networks (CNN) [10] and vision transformer (ViT) [6, 8, 35] become mainstream in FAS due to their strong semantic representation capacities to distinguish the bonafide from PAs.

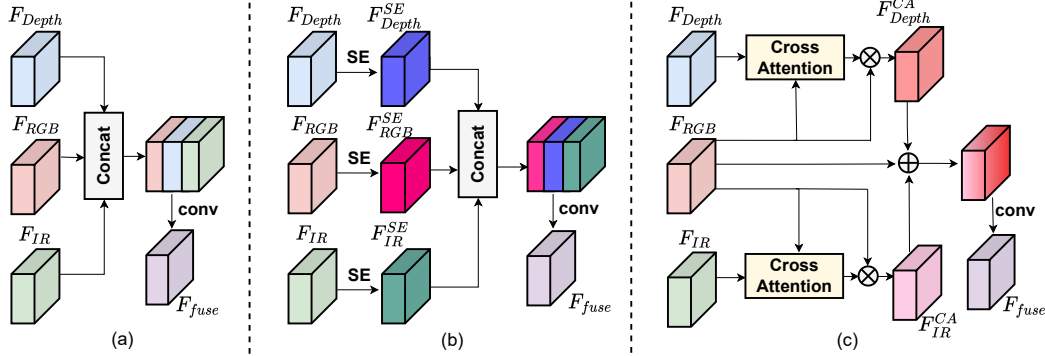With the development of hardware manufacture and in-

Figure 2. Feature fusion modules. (a) Direct concatenation [36]. (b) Squeeze-and-excitation (SE) fusion [41]. (c) Cross-attention fusion.

tegration technology, multi-modal FAS systems with acceptable costs are increasingly used in real-world applications. Meanwhile, a few large-scale multi-modal FAS datasets [9, 17, 41] as well as multi-modal deep learning based FAS methods [7, 18, 25, 28, 36] are proposed. In terms of multi-modal FAS datasets, CASIA-SURF [41] and CeFA [17] are with three modalities (*RGB*, *Depth*, and *Infra-red (IR)*) while WMCA [9] is with four modalities (*RGB*, *Depth*, *IR*, and *Thermal*). To learn intrinsic live/spoof features from multiple modalities, feature-level fusion strategies [16,25,30,31,36] are used. To better leverage contextual modality information and eliminate redundancy, spatial [28] and channel [28,41] attention is applied in multi-modal fusion.

Existing single- and multi-modal FAS methods usually separately train and deploy models for each possible modality scenario (see Fig. 1(a)), which might be redundant and inefficient. A few natural questions occur: Can we train a unified model, and flexibly deploy it under various modality scenarios? How about the performance and efficiency gaps among separate and unified single- and multi-modal models? To explore the questions above, we establish the first flexible-modal FAS benchmark with the principle 'train one for all', focusing on training a unified model for multiple real-world modality scenarios (see Fig. 1(b)). Our contributions include:

- We establish the first flexible-modal FAS benchmark with both intra- and cross-dataset testings under four evaluation modality scenarios (*RGB*, *RGB+Depth*, *RGB+IR*, and *RGB+Depth+IR*).
- We propose an elegant cross-attention fusion module to efficiently mine cross-modal clues for flexible-modal deployment. The proposed cross-Attention module significantly benefits the ViT [6] in both flexible intra- and cross-testings.
- We also investigate prevalent deep models (CDCN [40], ResNet [10], ViT [6]) and feature fusion strategies for flexible-modal FAS. We find

that the modality dropout strategy [25] works well in flex-modal intra-testings but poorly in flex-modal cross-testings.

2. Multi-Modal Fusion Baselines

For the *RGB*-based single-modal FAS task, given a face input X_{RGB}, the corresponding deep features/descriptors F_{RGB} could be extracted. Then a prediction head h is cascaded for binary live/spoof classification. For the *RGB+Depth+IR* multi-modal FAS task, independent-modality features F_{RGB}, F_{Depth}, and F_{IR} could be captured from face inputs X_{RGB}, X_{Depth}, and X_{IR}, respectively. All these features will be fused to form F_{fuse} first, and then forward the prediction head h. In this paper, we focus on feature-level fusion strategy but there are also e.g., decision-level fusion (a late fusion strategy) for multi-modal scenarios. Here we discuss about three feature-level fusion modules under the scenario with *RGB+Depth+IR*, and it is easily extended to scenarios with less or more modalities.

Direct concatenation fusion. Despite coarse alignment in the spatial domain, the features (F_{RGB}, F_{Depth}, and F_{IR}) in the channel domain have heterogeneous representation. As illustrated in Fig. 2(a), one classical solution is to concatenate these three features in the channel domain first [36], and then aggregate the multi-modal heterogeneous features with a lightweight fusion operator (e.g., convolution). The direct concatenation fusion can be formulated as

$$F_{fuse} = \text{ReLU}(\text{BN}(\text{Conv}(\text{Concat}(F_{RGB}, F_{Depth}, F_{IR})))). \tag{1}$$

Squeeze-and-excitation fusion. To alleviate the feature misalignment among modalities, squeeze-and-excitation (SE) module [11,41] is utilized in each independent modality branch first. With the channel-wise self-calibration via SE module, the refined features (F_{RGB}^{SE}, F_{Depth}^{SE}, and F_{IR}^{SE}) are then concatenated and aggregated. The framework of SE fusion is shown in Fig. 2(b). The SE fusion can be for-

mulated as (where σ denotes the Sigmoid function):

$$F_{RGB}^{SE} = F_{RGB} \cdot \sigma(\text{FC}(\text{ReLU}(\text{FC}(\text{AvgPool}(F_{RGB}))))),$$
$$F_{Depth}^{SE} = F_{Depth} \cdot \sigma(\text{FC}(\text{ReLU}(\text{FC}(\text{AvgPool}(F_{Depth}))))),$$
$$F_{IR}^{SE} = F_{IR} \cdot \sigma(\text{FC}(\text{ReLU}(\text{FC}(\text{AvgPool}(F_{IR}))))),$$
$$F_{fuse} = \text{ReLU}(\text{BN}(\text{Conv}(\text{Concat}(F_{RGB}^{SE}, F_{Depth}^{SE}, F_{IR}^{SE})))),$$
$$(2)$$

Cross-attention fusion. Besides fusion via multi-modal feature concatenation, we also explore the feature addition in the homogeneous space. To this end, we calculate the relationship maps between F_{RGB} and F_{Depth}/F_{IR} via cross-attention (CA), and then the normalized modality-interacted maps are multiplied by F_{RGB} to form cross-attentioned features F_{Depth}^{CA} and F_{IR}^{CA}. Finally, original RGB feature and cross-attentioned features are added and aggregated with an extra convolution. The framework of CA fusion is shown in Fig. 2(c). The CA fusion can be formulated as

$$\bar{F}_{Depth}^{CA} = \text{Softmax}(\bar{F}_{Depth}(\bar{F}_{RGB})^T)\bar{F}_{RGB},$$
$$\bar{F}_{IR}^{CA} = \text{Softmax}(\bar{F}_{IR}(\bar{F}_{RGB})^T)\bar{F}_{RGB},$$
$$F_{fuse} = \text{ReLU}(\text{BN}(\text{Conv}(F_{RGB} + F_{Depth}^{CA} + F_{IR}^{CA}))),$$
$$(3)$$

where F and $\bar{F}$ denote the spatial features and vectorized features, respectively.

3. Flexible-Modal FAS Benchmark

In this section, we introduce the flexible-modal FAS benchmark in terms of datasets, modality-aware protocols, and evaluation metrics. Statistical description is shown in Table 2.

Datasets. Three large-scale multi-modal datasets are used in the flexible-modal FAS benchmark. **CASIA-SURF** [41] consists of 1000 subjects with 21000 videos (7000 for RGB, Depth, and IR modality, respectively). There are two kinds of presentation attack instrument (PAI), i.e., print and cut print attacks, in CASIA-SURF. **CeFA** [17] is a cross-ethnicity FAS dataset, covering three ethnicities (Africa, East Asia, and Central Asia), three modalities (RGB, Depth, and IR), 1607 subjects with 7846 videos for each modality. In terms of PAIs, it consists of print, replay, 3D print and 3D silica gel mask attacks. **WMCA** [9] consists of 1941 short video recordings of both bonafide and PAs from 72 different identities. Each video is recorded from several spectrum channels including RGB, depth, IR, and thermal. In addition, there are seven PAIs (i.e., glasses, fake head, print, replay, rigid mask, flexible mask, and paper mask) in WMCA.

Protocols. Towards the principle 'train one for all', four flexible-modal protocols are established. Specifically, after trained on CASIA-SURF and CeFA with

Table 1: Statistics of the flexible-modal FAS benchmark. 'C.-S.' and 'D' are short for 'CASIA-SURF' and 'Depth', respectively. '#Video' indicates the video numbers of each modality.

Partition	Training		Validation		Testing		Testing
							Cross-dataset
			Intra-dataset				Cross-dataset
Dataset	C.-S.	CeFA	C.-S.	CeFA	C.-S.	CeFA	WMCA
#Subject	300	600	100	300	600	699	72
#Video	2100	2400	700	1200	4200	4246	1679
#PAI	2	2	2	2	2	4	7
Modality							
Protocol 1	RGB+D+IR		RGB		RGB		RGB
Protocol 2	RGB+D+IR		RGB+D		RGB+D		RGB+D
Protocol 3	RGB+D+IR		RGB+IR		RGB+IR		RGB+IR
Protocol 4	RGB+D+IR		RGB+D+IR		RGB+D+IR		RGB+D+IR

RGB+Depth+IR modalities, the unified multi-modal model is evaluated on both intra (CASIA-SURF and CeFA) and cross (WMCA) datasets under *RGB*-based (Protocol 1) single-modal, *RGB+ Depth*-based (Protocol 2) or *RGB+IR*-based (Protocol 3) bi-modal, and *RGB+Depth+IR*-based (Protocol 4) tri-modal scenarios. We also compare it with traditional separate training framework in terms of performance and efficiency.

Evaluation metrics. For all experiments, ACER [12] and True Positive Rate (TPR)@False Positive Rate (FPR) are used as evaluation metrics. ACER calculates the mean of Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER). The thresholds are determined by the *Equal Error Rate (EER) threshold* on validation set and set as *0.5* for intra- and cross-dataset testings, respectively. Besides, TPR@FPR=0.1% and TPR@FPR=1% are also utilized for fair comparison.

4. Experiment

4.1. Implementation Details

The experiments are implemented with Pytorch on one NVIDIA V100 GPU. Three deep models (CDCN [40], ResNet-50 [10] and ViT-Base [6]) are used with batch size 16, 64 and 64, respectively. We use the Adam optimizer with learning rate (lr) of 1e-4 for CDCN and ResNet while AdamW optimizer with lr=1e-5 for ViT-Base. CDCN is trained from scratch with 60 epochs while lr halves in the 30th epoch. Instead of supervision with pseudo depth maps [40], we follow [36] to supervise CDCN with simple binary maps. In contrast, ResNet50/ViT-Base is finetuned based on the ImageNet/ImageNet-21K pre-trained models with 30 epochs while lr halves in the 20th epoch. Direct concatenation is adopted as the default fusion method. For the SE fusion, the intermediate channel numbers are reduced to one eighth of original channels. The missing modalities are simply blocked as zeros in the testing phase of Protocols 1,2, and 3. To mimic such scenarios in the training phase, similar to [25], we randomly dropout the

Table 2: Results of intra-dataset testings on CASIA-SURF and CeFA datasets with 'Separate' and 'Unified' settings.

Method		Protocol 1 ACER(%) ↓	TPR(%)@FPR=0.1% ↑	Protocol 2 ACER(%) ↓	TPR(%)@FPR=0.1% ↑	Protocol 3 ACER(%) ↓	TPR(%)@FPR=0.1% ↑	Protocol 4 ACER(%) ↓	TPR(%)@FPR=0.1% ↑
Separate	CDCN [40]	32.49	4.7	6.22	47.4	43.98	0.93	6.46	55.43
	ResNet50 [10]	10.41	45.93	1.7	88.23	41.26	3.13	3.07	62.33
	ViT-Base [6]	10.81	31.33	1.44	91.27	26.34	6.5	3.82	80.87
Unified	CDCN	32.69	2.43	5.08	68.17	41.32	0.6	6.46	55.43
	CDCN w/ DropModal	36.49	1.47	7.91	40.83	35.89	0.97	11.36	30.1
	CDCN_SE	35.13	1.03	4.6	61	37.09	0.3	8.81	46.8
	CDCN_SE w/ DropModal	46.5	0.3	25.88	19.6	46.8	0.23	31.15	16.4
	CDCN_CA	34.99	1.93	31.55	2.37	35.18	1.83	34.33	2.2
	CDCN_CA w/ DropModal	40.58	1.23	38.9	1.43	39.38	1.2	40.94	1.2
	ResNet50	27.03	15.37	2.67	80.93	34.17	2.63	3.07	62.33
	ResNet50 w/ DropModal	14.24	23.23	9.32	61.5	18.18	22.77	8.1	56.27
	ResNet50_SE	20.59	6.87	2.05	78.37	27.67	2.3	2.48	55.3
	ResNet50_SE w/ DropModal	14.57	32.6	3.66	82.9	13.58	31.77	5.29	77.57
	ResNet50_CA	28.46	4.87	13.95	18.47	28.33	4.07	14.75	9.83
	ResNet50_CA w/ DropModal	18	9.8	13.72	26.1	16	7.47	13.21	20.37
	ViT-Base	20.33	4.07	2.5	84.27	29.51	3.1	3.82	80.87
	ViT-Base w/ DropModal	7.87	40.73	2.59	80.97	9.06	30.37	4.97	79.43
	ViT-Base_SE	23.28	1.5	1.87	92.67	37.38	1.93	2.7	88.2
	ViT-Base_SE w/ DropModal	8.58	36.6	3.01	77.73	10.18	30.8	3.03	78.6
	ViT-Base_CA	19.4	13.27	1.75	87.63	14.84	15.07	2.43	78.75
	ViT-Base_CA w/ DropModal	6.04	57.53	3.85	73.3	5.97	57.5	3.91	71.17

Table 3: Results of cross-dataset testings on WMCA when trained on CASIA-SURF and CeFA.

Method		Protocol 1 ACER(%) ↓	TPR(%)@FPR=1% ↑	Protocol 2 ACER(%) ↓	TPR(%)@FPR=1% ↑	Protocol 3 ACER(%) ↓	TPR(%)@FPR=1% ↑	Protocol 4 ACER(%) ↓	TPR(%)@FPR=1% ↑
Separate	CDCN [40]	34.35	9.61	27.58	8.17	40.62	3.27	28.28	4.71
	ResNet50 [10]	36.57	16.14	29.98	12.01	50	0.77	30.72	13.74
	ViT-Base [6]	39.38	5.96	36.81	11.53	49.34	2.88	33.08	3.07
Unified	CDCN	50	7.11	27.06	6.63	50	2.93	28.28	4.71
	CDCN w/ DropModal	32.25	7.78	25.44	13.83	32.95	6.24	26.92	11.82
	CDCN_SE	43.66	3.8	30.11	4.8	43.3	2.4	28.18	4.51
	CDCN_SE w/ DropModal	36.99	7.1	43.55	1.73	42.86	4.29	42.6	2.02
	CDCN_CA	40.61	1.06	32.73	3.5	41.26	1.06	40.74	2.5
	CDCN_CA w/ DropModal	35.26	6.53	34.49	7.2	35.56	7.2	34.44	5.48
	ResNet50	46.02	4.71	35.81	7.49	42.22	2.21	30.72	13.74
	ResNet50 w/ DropModal	46.45	11.53	25.36	19.79	46.53	10.28	19.43	19.79
	ResNet50_SE	49.52	1.34	28.86	7.88	43.58	0.67	32.55	7.49
	ResNet50_SE w/ DropModal	45.31	13.26	30.88	13.26	42.22	17.58	27.41	15.95
	ResNet50_CA	39.15	10.66	34	11.24	36.9	8.45	34.76	7.3
	ResNet50_CA w/ DropModal	43.32	4.42	31.94	7.3	47.65	5	37.17	8.17
	ViT-Base	48.52	9.03	44.42	7.59	50	0.38	33.08	3.07
	ViT-Base w/ DropModal	39.06	8.26	30.37	15.95	40.61	8.36	29.51	17.2
	ViT-Base_SE	49.95	3.75	30.64	5.09	50	1.83	41.03	5.48
	ViT-Base_SE w/ DropModal	33.48	9.51	30.67	9.41	35.96	5.19	31.33	8.55
	ViT-Base_CA	35.07	26.13	10.07	50.05	24.57	16.04	20.87	36.5
	ViT-Base_CA w/ DropModal	42.38	6.15	33.87	8.17	37.81	6.72	33.59	7.88

Depth and *IR* inputs (called DropModal).

4.2. Intra Testing

The experimental results of flexible-modal intra-dataset testing on CASIA-SURF and CeFA datasets is shown in Table 1. We can see from the first block and first rows in last three blocks that 1) the separated trained ResNet50 and ViT models perform obviously better than the unified counterparts while CDCN performs the opposite; and 2) ViT has higher TPR@FPR=0.1% than ResNet50 and CDCN on Protocols 2, 3, and 4 with both separated and unified settings, indicating the excellent multi-modal modeling capacities based on global self-attentioned features.

Impact of fusion modules. It can be seen from the 'Unified' block in Table 1, compared with directly concatenation fusion, the SE fusion [41] has no gains for CDCN, ResNet50, and ViT-Base. In contrast, we can find from the results of 'ViT-Base' and 'ViT-Base_CA' that the proposed CA module improves the ViT-Base remarkably (with gains 9.2%, 3.36% and 11.93% TPR@FPR=0.1% for Protocols 1, 2 and 3, respectively). Despite benefits from CA for ViT-Base backbone, the CA module still generalizes poorly

across other architectures (e.g., CDCN and ResNet50). It is still an open question to design architecture-agnostic fusion methods for the flexible-modal FAS benchmark.

Impact of DropModal. The multi-modal learning is easily dominated by partial-modal features (e.g., *Depth* modality) but neglecting other modalities with relatively weak clues (e.g., *IR* modality). The results of all variants of ResNet50 and ViT-Base on Protocols 1 and 3 are sharply improved with 'DropModal', indicating the augmentation with random modality dropout [25] alleviates the modality overfitting issue in intra testings.

4.3. Cross Testing

Table 3 shows the results of flexible-modal cross-dataset testing on WMCA. Due to the domain shifts (e.g., from sensors and sessions) and unseen PAIs, the performance of both separate and unified models are unsatisfactory (ACER>10%).

Impact of fusion modules. Similar to the intra-dataset testings, the results with SE fusion [41] cannot bring obvious benefits for CDCN, ResNet50, and ViT-Base on cross-dataset testings. As can be seen from the results of

Table 4: Results of 'shared' and 'unshared' multi-modal settings on intra-dataset testings on CASIA-SURF and CeFA.

Method		Protocol 1	Protocol 2	Protocol 3	Protocol 4	Overall	
		TPR(%)@FPR=0.1% ↑	TPR(%)@FPR=0.1% ↑	TPR(%)@FPR=0.1% ↑	TPR(%)@FPR=0.1% ↑	#Param. (M)	#FLOPs (G)
Separate	ViT-Base	31.33	91.27	6.5	80.87	362.01	132.31
	ViT-Base (unshared)	43.4	91.73	1.7	88.63	688.58	132.31
Unified	ViT-Base	4.07	84.27	3.1	80.87	96.83	199.92
	ViT-Base (unshared)	0.03	89.83	0.37	88.63	260.12	199.92

'ResNet50_SE', 'ResNet50_CA', 'ViT-Base_SE', and 'ViT-Base_CA' in Table 3, compared with SE fusion, the proposed CA is highly compatible with multi-modal ViT and ResNet50 architectures (especially on Protocols 1, 2, and 3), and improves the cross-dataset testing results dramatically.

Impact of DropModal. It is reasonable to find in Table 3 that 'DropModal' benefits the cross-testing performance of direct and SE concatenation fusions for all three kinds of models. However, the results of 'ResNet50_CA w/ DropModal' and 'ViT-Base_CA w/ DropModal' indicate that 'DropModal' degrades the cross-testing performance for CA fusion for both ResNet50 and ViT-Base backbones. It indicates that training CA with dropped-modality features limit the overall domain generalization capacity.

4.4. Efficiency Analysis

Both performance and efficiency are important in flexible-modal benchmark. Here we analyze the efficiency based on ViT-Base with two kinds of settings: separate/unified models and shared/unshared modality branches. As shown on the right part of Table 4, compared with separate models, the unified models save more than 50% parameters but require a bit more FLOPs (due to the fixed tri-modal branch setting in the testing phase) overall 4 protocols. Besides, using unshared backbone for independent modality branch usually brings slight performance improvement but introducing extra huge parameters. Overall, it will be a good tradeoff if the unified models with modality-shared backbones could achieve satisfactory performance.

5. Conclusion and Future Work

In this paper, we establish the first flexible-modal FAS benchmark with both intra- and cross-dataset testings. We also provide sufficient baselines on prevalent deep models and feature fusion strategies for flexible-modal FAS. However, from the experimental results in Tables 1, 3 and 4, there are still observable performance gaps between separate and unified models in some protocols, as well as the efficiency issue (#FLOPs). One limitation and assumption of the benchmark is that it needs all three modalities simultaneously in the training stage but RGB modality usually dominates while partial NIR and Depth modalities are missing under real-world scenarios.

References

[1] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *ICIP*, 2015.

[2] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE SPL*, 24(2):141–145, 2017.

[3] Rizhao Cai, Yawen Cui, Zhi Li, Zitong Yu, Haoliang Li, Yongjian Hu, and Alex Kot. Rehearsal-free domain continual face anti-spoofing: Generalize more and forget less. *arXiv preprint arXiv:2303.09914*, 2023.

[4] Rizhao Cai, Haoliang Li, Shiqi Wang, Changsheng Chen, and Alex C Kot. Drl-fas: A novel framework based on deep reinforcement learning for face anti-spoofing. *TIFS*, 16:937–951, 2020.

[5] Rizhao Cai, Zhi Li, Renjie Wan, Haoliang Li, Yongjian Hu, and Alex C Kot. Learning meta pattern for face anti-spoofing. *IEEE TIFS*, 2022.

[6] 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 *ICLR*, 2021.

[7] Anjith George and Sebastien Marcel. Cross modal focal loss for rgbd face anti-spoofing. In *CVPR*, 2021.

[8] Anjith George and Sébastien Marcel. On the effectiveness of vision transformers for zero-shot face anti-spoofing. In *IJCB*, 2021.

[9] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *TIFS*, 2019.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[12] international organization for standardization. Iso/iec jtc 1/sc 37 biometrics: Information technology biometric presentation attack detection part 1: Framework. In *https://www.iso.org/obp/ui/iso*, 2016.

[13] Xuan Li, Jun Wan, Yi Jin, Ajian Liu, Guodong Guo, and Stan Z Li. 3dpc-net: 3d point cloud network for face anti-spoofing. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2020.

[14] Zhi Li, Rizhao Cai, Haoliang Li, Kwok-Yan Lam, Yongjian Hu, and Alex C Kot. One-class knowledge distillation for face presentation attack detection. *IEEE TIFS*, 2022.

[15] 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*, 2021.

[16] Ajian Liu and Yanyan Liang. Ma-vit: Modality-agnostic vision transformers for face anti-spoofing. In *IJCAI*, 2022.

[17] Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multimodal cross-ethnicity face anti-spoofing. In *WACV*, 2021.

[18] Ajian Liu, Zichang Tan, Jun Wan, Yanyan Liang, Zhen Lei, Guodong Guo, and Stan Z Li. Face anti-spoofing via adversarial cross-modality translation. *TIFS*, 2021.

[19] Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zichang Tan, Qi Yuan, Kai Wang, Chi Lin, Guodong Guo, Isabelle Guyon, et al. Multi-modal face anti-spoofing attack detection challenge at cvpr2019. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[20] Ajian Liu, Chenxu Zhao, Zitong Yu, Anyang Su, Xing Liu, Zijian Kong, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zhen Lei, et al. 3d high-fidelity mask face presentation attack detection challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 814–823, 2021.

[21] Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, et al. Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 17:2497–2507, 2022.

[22] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, 2018.

[23] Yunxiao Qin, Zitong Yu, Longbin Yan, Zezheng Wang, Chenxu Zhao, and Zhen Lei. Meta-teacher for face anti-spoofing. *TPAMI*, 2021.

[24] 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 *AAAI*, 2020.

[25] Tao Shen, Yuyu Huang, and Zhijun Tong. Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing. In *CVPRW*, 2019.

[26] Jun Wan, Sergio Escalera, Hugo Jair Escalante, Guodong Guo, and Stan Z Li. Special issue on face presentation attack detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):282–284, 2021.

[27] Jun Wan, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Multi-modal face presentation attack detection. *Synthesis Lectures on Computer Vision*, 9(1):1–88, 2020.

[28] Guoqing Wang, Chuanxin Lan, Hu Han, Shiguang Shan, and Xilin Chen. Multi-modal face presentation attack detection via spatial and channel attentions. In *CVPRW*, 2019.

[29] Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In *CVPR*, 2020.

[30] Zitong Yu, Rizhao Cai, Yawen Cui, Xin Liu, Yongjian Hu, and Alex Kot. Rethinking vision transformer and masked autoencoder in multimodal face anti-spoofing. *arXiv preprint arXiv:2302.05744*, 2023.

[31] Zitong Yu, Rizhao Cai, Zhi Li, Wenhan Yang, Jingang Shi, and Alex C Kot. Benchmarking joint face spoofing and forgery detection with visual and physiological cues. *arXiv preprint arXiv:2208.05401*, 2022.

[32] Zitong Yu, Jukka Komulainen, Xiaobai Li, and Guoying Zhao. Review of face presentation attack detection competitions. In *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, pages 287–336. Springer, 2023.

[33] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Face anti-spoofing with human material perception. In *ECCV*, 2020.

[34] Zitong Yu, Xiaobai Li, Jingang Shi, Zhaoqiang Xia, and Guoying Zhao. Revisiting pixel-wise supervision for face anti-spoofing. *IEEE TBIOM*, 2021.

[35] Zitong Yu, Xiaobai Li, Pichao Wang, and Guoying Zhao. Transrppg: Remote photoplethysmography transformer for 3d mask face presentation attack detection. *IEEE SPL*, 2021.

[36] Zitong Yu, Yunxiao Qin, Xiaobai Li, Zezheng Wang, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Multi-modal face anti-spoofing based on central difference networks. In *CVPRW*, 2020.

[37] Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Deep learning for face anti-spoofing: A survey. *IEEE TPAMI*, 2022.

[38] Zitong Yu, Yunxiao Qin, Hengshuang Zhao, Xiaobai Li, and Guoying Zhao. Dual-cross central difference network for face anti-spoofing. In *IJCAI*, 2021.

[39] 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.

[40] 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 *CVPR*, 2020.

[41] 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 *CVPR*, 2019.