Flexible-Modal Face Anti-Spoofing: A Benchmark

Zitong Yu\textsuperscript{1}, Ajian Liu\textsuperscript{2}, Chenxu Zhao\textsuperscript{3}, Kevin H. M. Cheng\textsuperscript{4}, Xu Cheng\textsuperscript{5}, Guoying Zhao\textsuperscript{4}

\textsuperscript{1}Great Bay University, China
\textsuperscript{2}Institute of Automation, Chinese Academy of Sciences, China
\textsuperscript{3}SailYond Technology, China
\textsuperscript{4}University of Oulu, Finland
\textsuperscript{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 facial 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-
integration 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

$$F_{fuse} = \text{ReLU}(\text{BN}(\text{Conv}(\text{Concat}(F_{RGB}, F_{Depth}, F_{IR}))))).$$ (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 ($F'^{SE}_{RGB}$, $F'^{SE}_{Depth}$, and $F'^{SE}_{IR}$) 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 $\sigma$ denotes the Sigmoid function):
\[
F_{\text{RGB}}^{SE} = F_{\text{RGB}} \cdot \sigma(\text{FC}(\text{ReLU}(\text{FC}(\text{AvgPool}(F_{\text{RGB}}))))),
\]
\[
F_{\text{Depth}}^{SE} = F_{\text{Depth}} \cdot \sigma(\text{FC}(\text{ReLU}(\text{FC}(\text{AvgPool}(F_{\text{Depth}}))))),
\]
\[
F_{\text{IR}}^{SE} = F_{\text{IR}} \cdot \sigma(\text{FC}(\text{ReLU}(\text{FC}(\text{AvgPool}(F_{\text{IR}}))))),
\]
\[
F_{\text{fuse}} = \text{ReLU}((\text{BN})(\text{Conv}((F_{\text{RGB}}^{SE}, F_{\text{Depth}}^{SE}, F_{\text{IR}}^{SE})))).
\]

(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 $F_{\text{RGB}}$ and $F_{\text{Depth}}$, $F_{\text{IR}}$ via cross-attention (CA), and then the normalized modality-interacted maps are multiplied by $F_{\text{RGB}}$ to form cross-attentioned features $F_{\text{CA Depth}}$ and $F_{\text{CA IR}}$. 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
\[
F_{\text{CA Depth}} = \text{Softmax}(F_{\text{Depth}}(F_{\text{RGB}})^T)F_{\text{RGB}},
\]
\[
F_{\text{CA IR}} = \text{Softmax}(F_{\text{IR}}(F_{\text{RGB}})^T)F_{\text{RGB}},
\]
\[
F_{\text{Fuse}} = \text{ReLU}((\text{BN})(\text{Conv}((F_{\text{RGB}} + F_{\text{CA Depth}} + F_{\text{CA IR}})).
\]

(3)

where $F$ and $F$ 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 PAs, 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 PAs (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 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>
<thead>
<tr>
<th>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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>C.-S.</td>
</tr>
<tr>
<td>#Subject</td>
</tr>
<tr>
<td>#Video</td>
</tr>
<tr>
<td>#PAI</td>
</tr>
<tr>
<td>Protocol 1</td>
</tr>
<tr>
<td>Protocol 2</td>
</tr>
<tr>
<td>Protocol 3</td>
</tr>
<tr>
<td>Protocol 4</td>
</tr>
</tbody>
</table>

6347
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 tests, 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 2:** Results of intra-dataset testings on CASIA-SURF and CeFA datasets with ‘Separate’ and ‘Unified’ settings.

<table>
<thead>
<tr>
<th>Method</th>
<th>Protocol 1</th>
<th>Protocol 2</th>
<th>Protocol 3</th>
<th>Protocol 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TPR@FPR=1%</td>
<td>TPR@FPR=0.1%</td>
<td>TPR@FPR=1%</td>
<td>TPR@FPR=0.1%</td>
</tr>
<tr>
<td>CDCN [40]</td>
<td>52.39</td>
<td>49.3</td>
<td>46.4</td>
<td>43.7</td>
</tr>
<tr>
<td>ResNet50</td>
<td>10.41</td>
<td>45.93</td>
<td>52.1</td>
<td>49.3</td>
</tr>
<tr>
<td>ViT-Base</td>
<td>10.01</td>
<td>31.33</td>
<td>36.9</td>
<td>34.3</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Protocol 1</th>
<th>Protocol 2</th>
<th>Protocol 3</th>
<th>Protocol 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TPR@FPR=1%</td>
<td>TPR@FPR=0.1%</td>
<td>TPR@FPR=1%</td>
<td>TPR@FPR=0.1%</td>
</tr>
<tr>
<td>CDCN [40]</td>
<td>34.35</td>
<td>9.61</td>
<td>27.58</td>
<td>8.17</td>
</tr>
<tr>
<td>ResNet50</td>
<td>36.57</td>
<td>16.14</td>
<td>29.98</td>
<td>12.01</td>
</tr>
<tr>
<td>ViT-Base</td>
<td>39.38</td>
<td>5.09</td>
<td>36.81</td>
<td>11.51</td>
</tr>
<tr>
<td>CDCN</td>
<td>20.59</td>
<td>7.78</td>
<td>25.44</td>
<td>13.83</td>
</tr>
<tr>
<td>ResNet50</td>
<td>43.66</td>
<td>3.8</td>
<td>30.11</td>
<td>4.8</td>
</tr>
<tr>
<td>ViT-Base</td>
<td>36.99</td>
<td>7.1</td>
<td>43.55</td>
<td>1.73</td>
</tr>
<tr>
<td>CDCN</td>
<td>40.61</td>
<td>10.6</td>
<td>32.73</td>
<td>3.5</td>
</tr>
<tr>
<td>ResNet50</td>
<td>35.26</td>
<td>6.53</td>
<td>34.49</td>
<td>7.2</td>
</tr>
<tr>
<td>ViT-Base</td>
<td>46.45</td>
<td>11.53</td>
<td>25.36</td>
<td>19.79</td>
</tr>
<tr>
<td>CDCN</td>
<td>49.52</td>
<td>1.54</td>
<td>28.86</td>
<td>7.88</td>
</tr>
<tr>
<td>ResNet50</td>
<td>45.31</td>
<td>13.26</td>
<td>30.33</td>
<td>13.26</td>
</tr>
<tr>
<td>ViT-Base</td>
<td>39.15</td>
<td>10.66</td>
<td>34</td>
<td>11.24</td>
</tr>
<tr>
<td>CDCN</td>
<td>43.32</td>
<td>4.42</td>
<td>31.94</td>
<td>7.3</td>
</tr>
<tr>
<td>ResNet50</td>
<td>39.06</td>
<td>8.26</td>
<td>30.37</td>
<td>15.95</td>
</tr>
<tr>
<td>ViT-Base</td>
<td>49.95</td>
<td>5.75</td>
<td>30.64</td>
<td>5.09</td>
</tr>
<tr>
<td>CDCN</td>
<td>35.07</td>
<td>26.13</td>
<td>10.07</td>
<td>50.05</td>
</tr>
<tr>
<td>ResNet50</td>
<td>42.38</td>
<td>32.87</td>
<td>33.87</td>
<td>8.17</td>
</tr>
</tbody>
</table>

---

6348
‘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 tests. 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


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.


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

