Effective Presentation Attack Detection Driven by Face Related Task

Wentian Zhang† [1,2,3], Haozhe Liu† [1,2,3,4], Feng Liu✉ [1,2,3], Raghavendra Ramachandra[5], and Christoph Busch[5]

[1] College of Computer Science and Software Engineering, Shenzhen University
[2] SZU Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society
[3] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China
feng.liu@szu.edu.cn
[4] King Abdullah University of Science and Technology, Saudi Arabia
[5] Norwegian Biometrics Laboratory (NBL), Norwegian University of Science and Technology, Gjøvik 2818, Norway

Abstract. The robustness and generalization ability of Presentation Attack Detection (PAD) methods is critical to ensure the security of Face Recognition Systems (FRSs). However, in a real scenario, Presentation Attacks (PAs) are various and it is hard to predict the Presentation Attack Instrument (PAI) species that will be used by the attacker. Existing PAD methods are highly dependent on the limited training set and cannot generalize well to unknown PAI species. Unlike this specific PAD task, other face related tasks trained by huge amount of real faces (e.g. face recognition and attribute editing) can be effectively adopted into different application scenarios. Inspired by this, we propose to trade position of PAD and face related work in a face system and apply the free acquired prior knowledge from face related tasks to solve face PAD, so as to improve the generalization ability in detecting PAs. The proposed method, first introduces task specific features from other face related task, then, we design a Cross-Modal Adapter using a Graph Attention Network (GAT) to re-map such features to adapt to PAD task. Finally, face PAD is achieved by using the hierarchical features from a CNN-based PA detector and the re-mapped features. The experimental results show that the proposed method can achieve significant improvements in the complicated and hybrid datasets, when compared with the state-of-the-art methods. In particular, when training on the datasets OULU-NPU, CASIA-FASD, and Idiap Replay-Attack, we obtain HTER (Half Total Error Rate) of 5.48% for the testing dataset MSU-MFSD, outperforming the baseline by 7.39%. The source code is available at https://github.com/WentianZhang-ML/FRT-PAD.

Keywords: Face, Presentation Attack Detection, Graph Neural Network

† Equal Contribution
This work was done when Haozhe Liu was a visiting student at NTNU, Norway, under the supervision of R. Ramanchandra and C. Busch.

1 Introduction

Face Recognition Systems (FRSs) are widely deployed in authentication applications especially in access control and mobile phone unlocking in our daily life [17,35]. However, recent studies [26,15] demonstrate the existing face recognition systems are lacking robustness, since they are easily spoofed by presentation attacks (PAs), such as photographs, video replays, low-cost artificial masks [3] and facial make-up attacks [7]. Meanwhile, the face images can be easily obtained from social media, which seriously increases the risk of PAs. These issues raise wide concerns about the vulnerability of facial recognition technologies. Consequently, it is crucial to detect PAs to achieve robust and reliable FRS.

To tackle such challenge, many face PAD methods have been proposed, which can be divided into hardware and software based methods. Hardware based solutions [14,25] generally employ specific sensors to acquire presentations with different image modalities to detect PAs. Although, these methods provide strong robustness, their applicability is still limited because of unsatisfying performance to new application scenarios and cost limitations. Software based algorithms usually explore the distinctive features between bona fides (live faces) and PAs, such as hand-crafted features [8,33,24,1] and deep features [34,10]. Due to the advances of deep learning in recent years, deep feature based methods have been widely used in the community, since better performance can be achieved by adopting convolutional neural networks (CNNs) [16,32,12].

However, such learning-based method might not obtain ideal generalization to different unseen attacks, since they often needs comparable bona fide and PA samples for training. In the real scenario, PAs with different materials and instruments are hard to collect, the PAD mechanisms are limited by the unbalanced training data [18,20]. On the contrary, face related tasks (e.g. face, expression and attribute editing) possess strong generalization capability, since they are trained by millions of live faces cross genders, ages and ethnic groups from specific datasets. We argue that face PAD task should share some common patterns with other face related tasks, and the performance of PA detection might be boosted by the features adopted from such tasks.

As shown in Fig. 1(a), in traditional face systems, existing PAD mechanism like CNN-based PA detector, always plays a forward in face system and is independent with the following face related tasks. However, it is rarely discussed to contact PAD mechanism with face related tasks in the community. Since the face related tasks have already been trained, we consider that why not directly use these **free acquired** face-related features from face related tasks to serve for PAD. As shown in Fig. 1(b), different from traditional face systems, we attempt to directly implement the trained face related tasks in face systems, and then utilize their extracted task specific features to make PAD mechanism generalize well in real scenarios.

In this paper, we propose a PAD method utilizing feature-level prior knowledge from face related task, denoted as **FRT-PAD** to solve the above mentioned problem. First, we trade the position of PAD mechanism and face related task in a face system. The task specific features, which contain abundant generalization

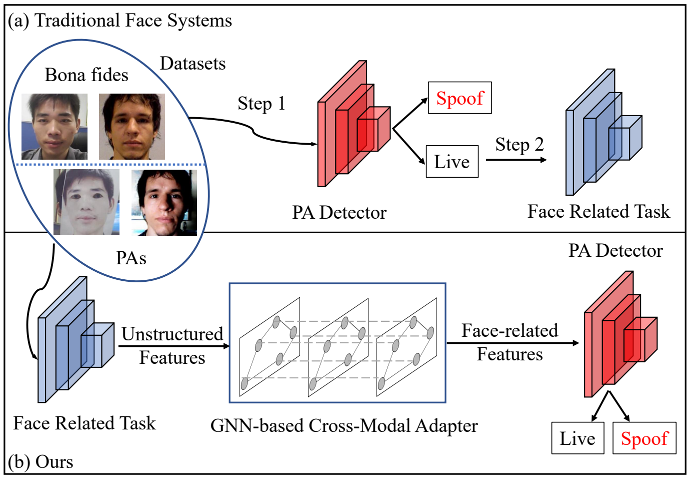


Fig. 1. The diagram of (a) traditional face systems with PA detector and (b) the proposed scheme. In the proposed method, face systems, including face, expression and attribute editing, are directly implemented. We then apply a GNN-based Cross-Modal Adapter to adapt their extracted features, which contain abundant generalization knowledge, to facilitate PAD.

knowledge, are directly obtained from face related tasks. By following a Graph Neural Network (GNN), a Cross-Modal Adapter is put forward to re-map and adapt the task specific features to PAD task. The generalization capability is finally improved by alleviating the problem resulting in limited PA samples for training. The main contributions of this work are concluded as follows,

- Existing PAD mechanism trained on the limited datasets are vulnerable to unseen PAs. To address this problem, we rethink the order of PAD mechanism and face related task in face system and directly introduce the free acquired features from face related tasks in PAD, which can improve generalization ability of PAD model.
- A Cross-Modal Adapter is designed to obtain face-related features, which can adapt task specific features into PAD space.
- The effectiveness and superiority of our method is evaluated on the public datasets. Particularly, when OULU-NPU [3], CASIA-FASD [38], and Idiap Replay-Attack [4] are adopted as training set and MSU-MFSD [33] is used as test set, the HTER (Half Total Error Rate) of the proposed method can outperform the baseline by 7.39%.

2 Related Works

As face related task is for the first time introduced in face PAD, the proposed face PAD scheme is different from existing PAD methods. Hence, our reviews mainly include face PAD methods and face related tasks.

2.1 Face Presentation Attack Detection

Existing face PAD methods can be categorized into hand-crafted and deep learning based methods. Hand-crafted methods employ the algorithms, e.g. LBP [8], IDA [33], SIFT [24], and SURF [1] to extract the features and then adopt traditional classifiers such as LDA and SVM to detect PAs. However, the hand-crafted features can be easily influenced by the variations of imaging quality and illumination. As a result, feature based methods generally can not generalize well to different application scenarios.

To address such challenges, deep learning models are then proposed for face PAD. Yang et al. [34] proposed to use CNNs to extract deep discriminative features for face PAD. Nguyen et al. [23] designed a multi-task learning model, which locates the most important regions of the input to detect PAs. Yu et al. [36] proposed a Central Difference Convolution (CDC) structure to capture intrinsic detailed patterns for face PAD and then used Neural Architecture Search (NAS) in CDC based network to achieve a better result. Besides applying only RGB images, auxiliary information of face, e.g. face depth, are considered to establish a more robust detector. Liu et al. [21] explored face depth as auxiliary information and estimated rPPG signal of RGB images through a CNN-RNN model for face PAD. George et al. [12] introduced a cross-modal loss function to supervise the multi-stream model, which extracted features from both RGB and depth channels. Liu et al. [19] proposed a self-supervised learning based method to search the better initialization for face PAD. Although such deep learning methods can achieve better PAD performance, their dependence on training data would inevitably leads a bias when accessible data is limited. In particular, numerous studies [11,28,30] have shown that, PA detectors trained on one dataset can not generalize to other datasets effectively.

To improve the generalization, researchers have further proposed one-class and domain generalization methods. A one-class multi-channel CNN model [11] was proposed to learn the discriminative representation for bona fides within a compact embedding space. Different from one-class methods, domain generalization based methods pay more attention to the disparities among the domains. Shao et al. [28] proposed a multi-adversarial deep domain generalization framework to learn a generalized feature space within the dual-force triplet-mining constraint. Since the learned feature space is discriminative and shared by multiple source domains, the generalization to new face PAs can be ensured effectively. Wang et al. [30] disentangled PAD informative features from subject-driven features and then designed a multi-domain learning based network to learn domain-independent features cross different domains for face PAD. Generally speaking, when training data is adequate for the aforementioned method, the detector can achieve very competitive performance. However, unlike other face related tasks, Presentation Attacks (PAs) are hard to collect and the types/instruments of attacks are consistently increasing. Due to such open challenges, we propose a PAD mechanism based on face related task in a common face system to decrease the dependence of the PA detector on data scale. In particular, we obtain face-related features from face related tasks for a more robust representation with better generalization capability. Benefit from extensive samples collected in other tasks, face PA detector can achieve significant improvement within limited data.

2.2 Face Related Tasks

With the advances in deep learning, face related tasks including face recognition, face expression recognition, face attribute editing, etc, have become a very active field [6,31,5]. In this section, we briefly introduce some representative tasks adopted in this work due to the limited scope of the paper. For large-scale face recognition, Deng et al. [6] proposed an Additive Angular Margin Loss (ArcFace) , which has a clear geometric interpretation, to obtain highly discriminative features. ArcFace is a solid work evaluated on the various face recognition benchmarks, including image datasets with trillions of pairs and a large-scale video dataset. In the terms of facial expression recognition, Wang et al. [31] proposed a Self-Cure Network (SCN) to address uncertainties in facial expressions. By combining self-attention and relabelling mechanism, such method can prevent deep networks from over-fitting uncertain facial images. Choi1 et al. [5] proposed a scalable approach called StarGAN, to perform image-to-image translations for multiple domains and achieve facial attribute transfer. StarGAN can flexibly learn reliable features universally applicable to multiple domains of images with different facial attribute values, which is always set as a famous baseline in generative task. As the representative works for the corresponding tasks, the aforementioned solutions are conducted in this paper to investigate the relationship between PAD and face related tasks.

3 Proposed Method

In this paper, we propose a face-related-task based PAD mechanism (**FRT-PAD**) to improve the generalization capability of PAD model. As shown in Fig. 2, the proposed **FRT-PAD** method consists of two branches, including a CNN based PA detector and an auxiliary branch. The CNN based PA detector disentangles disparities between bona fides and PAs by directly extracting features from image space. The auxiliary branch aims to extract face-related features from a model trained by face related tasks. In such auxiliary branch, we firstly hierarchically obtain the task-specific features from multiple layers of the trained model. Then, we design a Cross-Modal Adapter based on GNN to adapt the features to PAD. The features from both branches are fused comprehensively for the final PAD. In the following sections, we will present the detailed discussion on the proposed method.

3.1 Task Specific Features from Face Related Tasks

As some face related tasks are trained from huge amount of faces, features extracted by such trained networks have better generalization capability. The hypothesis of this work is that face related tasks share some common patterns in

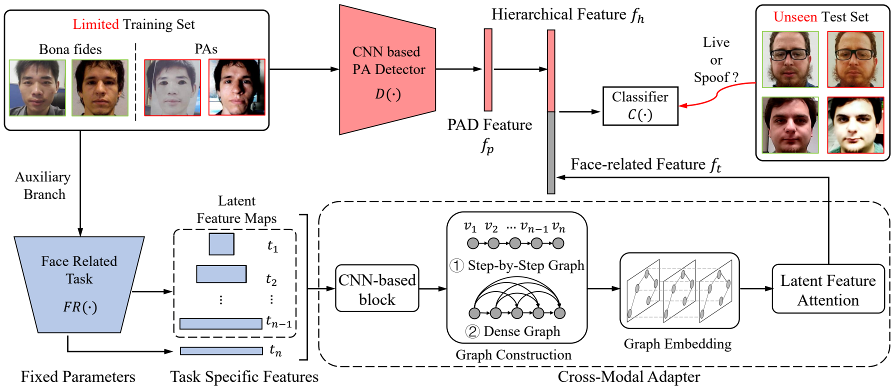


Fig. 2. The pipeline of our proposed PAD mechanism based on face related task in a common face system (**FRT-PAD**). In auxiliary branch, the task specific features t_i can be extracted by the parameter-fixed model $FR(\cdot)$, which has been trained from face related tasks. Then, a CNN-based block is used to transform t_i to graph vertexes. We construct a Graph Attention Netwok (GAT) to re-map t_i to fit PAD. By following latent feature attention, the face-related feature f_t is obtained, and finally fused with the main branch to achieve face PAD.

face feature learning. For example, expression recognition requires the model to localize the action unit of the face, which also serves as a potential feature for PAD. Through transferring the knowledge contained in trained tasks, the dependence of PA detector on a large training data can be reduced. Hence features from face related tasks not only perform strong generalization in the trained task, but can also boost the performance of PA detector. Let x refers to a face in training set, and denote $FR(\cdot)$ as the network trained by face related tasks, e.g. face recognition, expression recognition and attribute editing. As shown in Fig. 2, $FR(\cdot)$ embeds x to a task specific feature $T = \{t_i | i \in [1, n]\}$, where t_i refers to the feature map extracted from the i-th layer. As a multi-level representation of x, the features from different layer represent different properties of the face. To prevent from the loss of information, the proposed method regards such non-structure feature map as the input of the auxiliary branch.

3.2 Cross-Modal Representation Using GAT

As $FR(\cdot)$ is trained by face related tasks, the extracted features t_i might contain information unrelated to the face PAD task. To alleviate the potential negative influence of the irrelevant information, we propose a Cross-Modal Adapter to re-map them for PAD. Considering that task specific features are non-structural, a Graph Neural Network (GNN), denoted as $G(V, E)$, is employed to process T. $v_i \in V$ denotes vertex feature of graph. E is the edge matrix of graph to connect

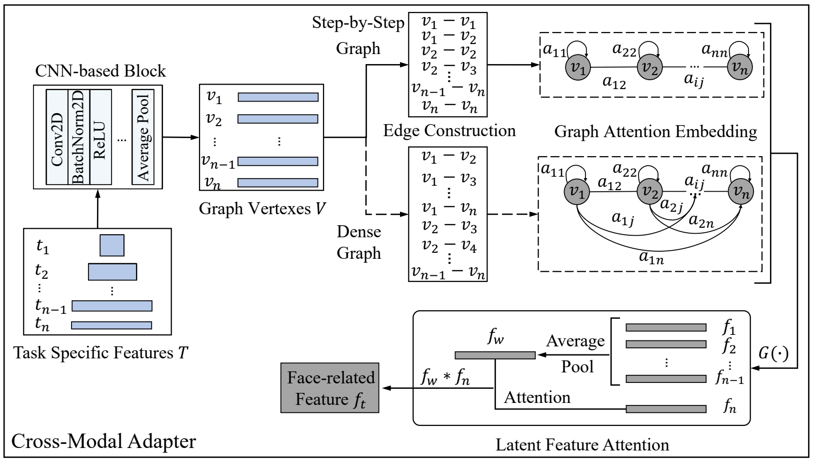


Fig. 3. The pipeline of Cross-Modal Adapter. To transform the task specific features $t_i \in T$ to graph vertexes, t_i are reconstructed to one-dimension vectors $v_i \in V$ using a CNN-based block. We design two different graphs, including Step-by-Step Graph and Dense Graph. In both graphs, attention mechanisms are used to specify the connection strength between different v_i. Through graph embedding $G(\cdot)$, v_i can be re-mapped to f_i. Then, the latent feature f_i is used to compute latent feature attention f_w by an average pool operation. The final face-related feature f_t is obtained by $f_w * f_n$.

neighboring vertexes given by:

$$E = \begin{cases} e_{ij} = 1 \ v_i \rightarrow v_j \\ e_{ij} = 0 \ v_i \overset{\times}{\rightarrow} v_j \end{cases} \tag{1}$$

Given two graph vertexes v_i and v_j, $e_{ij} = 1$ presents an undirected edge existing between them. To construct neighboring v_i properly in E, the relationship among v_i is needed to be exploited. Since t_i is extracted by $FR(\cdot)$ step by step, T can perform as a sequence. Based on such observation, we propose two potential graph structures, including Step-by-Step Graph and Dense Graph, to investigate the reasonable utilization of T.

However, matrix E can only reflect the connection of each v_i. The importance of different v_i is unknown. Thus, we adopt an attention mechanism in GNN (i.e. GAT) embedding to find out the contribution of neighboring vertex to the central vertex. More important vertex features will obtain larger weights. Formally, the relative weights of the connected graph vertexes can be computed by:

$$A = E * (q_1 W V + (q_2 W V)^T) \tag{2}$$

where $Q = \{q_1, q_2\}$ is set as the shared attention parameter. A shared learnable weight matrix W is set to achieve graph convolutional operation, which can map each vertex v_i to a high-level feature. A stands for the attention matrix of graph vertexes V. Then, we use softmax to normalize $a_{ij} \in A$ by following the rule of connection:

$$A_s(i,j) = \frac{exp(a_{ij})}{\sum_{k \in [1,n]} exp(a_{ik}))} \tag{3}$$

Given two connected points v_i and v_j , $A_s(i,j)$ and $A_s(j,i)$ measures the connection strength coefficient of them. For one-layer attention, the vertex features with attention weights can be obtained:

$$V^{'} = A_s W V \tag{4}$$

As shown in Fig. 3, task specific features t_i are firstly embedded into graph vertex features v_i by a CNN-based block. Then, for Step-by-Step Graph, two graph vertexes v_i and v_{i+1} are sequentially connected by a single edge. Different graph vertexes are fully connected in Dense Graph. Through the multi-layer graph embedding $G(V^{'},E)$, the transformed features f_i are obtained. We regard latent features as the attention weights to strengthen the representation of f_n. Through an average pooling operation along with $f_i, i \in [1, n-1]$, the latent feature attention f_w can be computed. The final face-related feature f_t is represented by $f_w * f_n$.

3.3 Face-related-task based Presentation Attack Detection

To adopt the re-mapped face-related feature in face PAD, the proposed method introduces a CNN based PA detector $D(\cdot)$ to learn the PAD feature f_p. Then,

Algorithm 1 Presentation Attack Detection Using FRT-PAD

```
Input:
  Training Set $\mathcal{X}_T$; CNN based PA detector $D(\cdot)$; Face recognition network $FR(\cdot)$;
  Classifier $C(\cdot)$; Graph embedding $G(\cdot)$;
Output:
  Trained $D(\cdot)$, $G(\cdot)$ and $C(\cdot)$;
Fixed parameters of Trained $FR(\cdot)$;
for $x_j$ in $X_T$ do
    Extract PAD feature $f_p$ through $D(x_j)$;
    Derive task specific features $t_i \in T$ from $FR(x_j)$;
    Transform $t_i$ to vector $v_i \in V$ as graph vertexes;
    Construct edge matrix $E$;
    Derive vertex features $V^{'}$ with attention weighs;
    Extract transformed features $f_i$ through $G(V^{'},E)$;
    Obtain attention weights $f_w$ from $f_i, i \in [1, n-1]$;
    Calculate face-related feature $f_t$ by $f_w * f_n$;
    Derive hierarchical feature $f_h$ from $f_p$ and $f_t$;
    Predict the PAD result by $C(f_h)$;
    Update $D(\cdot)$, $G(\cdot)$ and $C(\cdot)$by minimizing Eq. 5;
end for
Return $D(\cdot)$, $G(\cdot)$ and $C(\cdot)$;
```

a hierarchical feature f_h is derived by concatenating PAD feature f_p and face-related feature f_t. Through f_h, classifier $C(\cdot)$ can effectively distinguish bona fides with PAs . In the training process, $D(\cdot)$, $G(\cdot)$ and $C(\cdot)$ are trained by a cross entropy as follows:

$$\mathcal{L}_{x_j \in \mathcal{X}_T}(x_j, y_j^{'}) = -\frac{1}{N}\sum_{j=1}^{N}[y_j log(y_j^{'}) + (1 - y_j) log(1 - y_j^{'})] \tag{5}$$

where $(x_j, y_j), j \in [1, N]$ are the paired samples from training set $\mathcal{X}_T$, and $y_j^{'}$ is the prediction result of $C(\cdot)$. For clarity, the proposed method is summarized in Algorithm 1.

4 Experimental Results and Analysis

In this section, we evaluate the performance of the proposed method by carrying experiments on the publicly-available datasets [3,38,4,33], First, the datasets and the corresponding implementation details are introduced. Then, we validate the effectiveness of the proposed method through analyzing the influences of each network component to PAD performance. Finally, to prove the superiority of our method, we compare the PAD performance of the proposed method with the state-of-the-art methods.

4.1 Datasets and Implementation Details

We use four public face anti-spoofing datasets, including OULU-NPU [3] (denoted as O), CASIA-FASD [38] (denoted as C), Idiap Replay-Attack [4] (denoted as I) and MSU-MFSD [33] (denoted as M) to evaluate the effectiveness of our method. Existing methods were evaluated on the protocol [16], denoted as Protocol-I. In this protocol, three of datasets are used as training set and the remaining one is used for test. However, in the reality, there are much more unseen PAs than the known ones in the training set. Using 3/4 datasets to train model is not strict to the real application scenario. Thus, we design a different cross-dataset protocol (Protocol-II) to evaluate the generalization ability of our method. In detail, we only use two datasets from [O, M, C, I] to train model and the remaining two datasets to test. Due to the number of samples varies greatly among each dataset, some data divisions will be unreasonable for model training. To make the number of training set and test set as close as possible, we only divide the datasets into two groups, i.e. [O, M] and [C, I]. To reduce the influence caused by the background, resolution, and illustration, MTCNN algorithm [37] is used for face detection and alignment. All the detected faces are resized to (256, 256). ResNet18 [13] is fine-tuned as the CNN based PA detector. Three network trained through the face related tasks are used to obtain task specific features. ResNet18 trained by face related task [6] and face expression recognition [31] is set as the feature extractor. For face attribute editing task, the trained discriminator of StarGAN [5] is set to extract task specific features.

In the graph embedding, a two-layer GAT with two-head attention mechanisms is adopted. In the training phase, the parameters of networks with face related tasks are fixed and the weight of task specific features are automatically determined. In terms of connection among vertexes, the attention mechanisms can specify the connection strength between them to show the most beneficial vertex. To evaluate the cross-modal adapter of our method, besides GAT, we adopt other deep learning methods, including ResNet18 [13] and transformer [29], as the competing methods.

In summary, we train PA detector, classifier and cross-modal adapter by Adam with 1e-4 learning rate and 5e-5 weight decay. Batch size for training is 32. To validate the superiority of our FRT-PAD method, the state-of-the-art PAD methods, including DeepPixBiS [10], SSDG-R [16], CDC [36], and IF-OM [19] are conducted in this paper. Following the work of [19], We use Half Total Error Rate (HTER), Area Under Curve (AUC) and Bona Fide Presentation Classification Error Rate (BPCER) when Attack Presentation Classification Error Rate ($APCER_{AP}$) is 1% to evaluate the performance of PAD. This paper adopts the public platform pytorch for all experiments using a work station with CPUs of 2.8GHz, RAM of 512GB and GPUs of NVIDIA Tesla V100.

4.2 Effectiveness Analysis of the Proposed Method

Face-related Features We perform the ablation study to quantify the influence of each component in our model for face PAD. First, to evaluate the effectiveness of face-related features, we test the performance of PAD with or without the face-related features. Table 1 shows the results carried on the cross-dataset protocol. The baseline is set as the ResNet18 model pretrained from ImageNet. We use three face related tasks, including face recognition, face expression recognition, and face attribute editing to extract task specific features. Then, by a step-by-step GAT, we can respectively obtain three different face-related features (*F.R.*, *F.E.*, *F.A.*). Compared with the baseline, all three face-related features can improve the PAD performance. Specifically, *F.A.* feature adapted from face attribute editing task improves the HTER of baseline from 26.90% to 15.08%. This indicates that face-related features are useful to face PAD. As experiments are carried on the cross-dataset protocol, it also indicates that face-related features can improve generalization ability of face PAD.

Table 1. Performance of the Proposed Method with or without Face-related Features Obtained from Three Different Tasks.

	[O.M] to [C,I]			[C,I] to [O,M]		
	HTER(%)↓	AUC(%)↑	BPCER(%)↓	HTER(%)↓	AUC(%)↑	BPCER(%)↓
Baseline	25.65	79.14	95.93	28.14	79.05	81.34
Baseline w/ *F.R.*	18.17	87.37	78.52	16.47	90.68	62.81
Baseline w/ *F.E.*	17.93	85.97	90.90	16.62	91.78	55.66
Baseline w/ *F.A.*	**16.98**	**90.66**	**58.32**	**13.18**	**94.36**	**43.70**

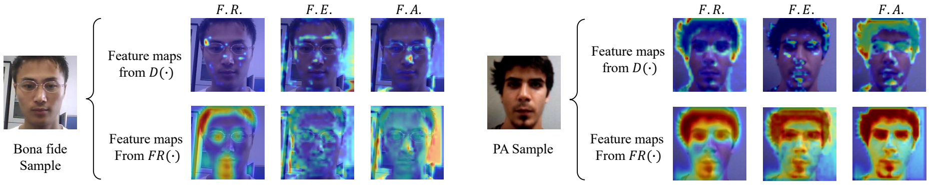

F.R.
F.E.
F.A.
Feature maps from $D(\cdot)$
Feature maps From $FR(\cdot)$
Bona fide Sample
F.R.
F.E.
F.A.
Feature maps from $D(\cdot)$
Feature maps From $FR(\cdot)$
PA Sample

Fig. 4. The visualization on CASIA-FASD using Grad-CAM. The first row for each sample shows the discriminative regions obtained from CNN-based PA detector $D(\cdot)$ and the second row for each sample illustrate the region localization extracted from network of face related tasks $FR(\cdot)$.

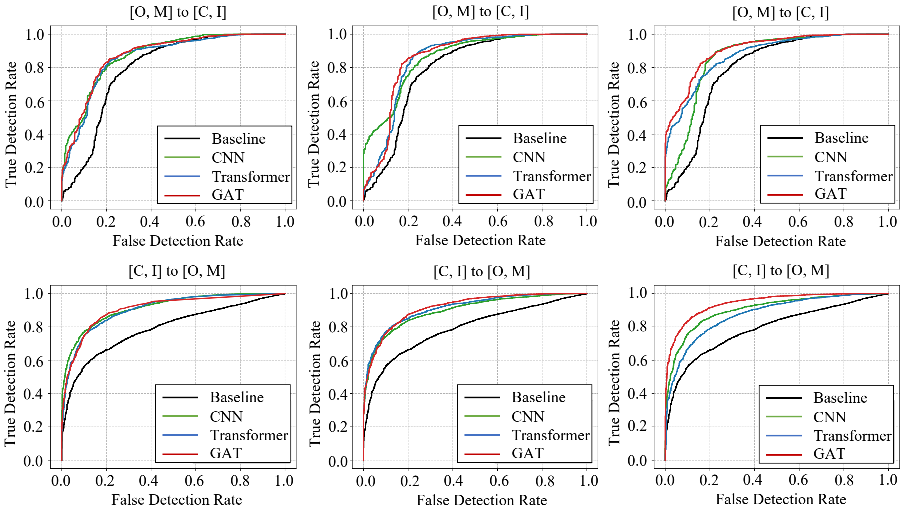
To further verify the effectiveness of face-related features, we adopt Grad-CAM [27] to visualize the discriminative regions from feature maps in our proposed FRT-PAD model. We compare the visualization results with three face-related features. Cross-Modal Adapters are set as the Step-by-Step Graphs. As shown in Fig. 4, when using face-related features, the model can find discriminative features for both bona fide and PA samples. In the visualization region obtained form $FR(\cdot)$, the visualization shows that the hair, eyes, nose and mouth are important to distinguish live faces and spoofs. This further indicates the effectiveness of the face-related features. Typically, comparing with the visualization of $F.R.$ and $F.E.$ face-related features, $F.A.$ features can provide more effective region of face attributes.

Cross-Modal Adapter For each task specific feature v_i, we further compare two other different deep learning models with GAT, i.e. CNN based model and transformer model, to justify the effectiveness of the Cross-Modal Adapter. In CNN based model, we use the same CNN-based block and latent feature attention in Fig. 3 to obtain the face-related feature. In transformer based model, each v_i is transformed to vector adopting CNN-based block in Fig. 3 and encoded with position encoding module in [29]. Then, we adopt six-layer transformer encoders with eight-head-attention to obtain the face-related feature. As CNN model and transformer model are sequential models, we only use the Step-by-Step Graph to ensure the fairness of the comparison. The PAD results in Table 2 show that the performance of the proposed method with different deep learning models in Cross-Modal Adapter and face-related features is better than the baseline.

Corresponding to Table 2, Fig. 5 presents the ROC (Receiver Operating Characteristic) curves of baseline and three face-related features when using different model in Cross-Modal Adapters. It can be seen that, for all face-related features, Cross-Modal Adapter using GAT model (with red lines in Fig. 5) achieves a higher performance than CNN model and transformer model. This indicates that GAT model is more suitable for the Cross-Model Adapter. In particular, face-related feature $F.A.$ re-mapped from GAT can obtain the best results in [C, I] to [O, M] protocol, and achieve an HTER of 13.18% and AUC

Table 2. Performance of the Proposed Method Using Different Models in Cross-Modal Adapter for Three Face-related Features.

Face-related Features	**Cross-Modal Adapters**	[O.M] to [C,I]			[C,I] to [O,M]		
		HTER(%)↓	AUC(%)↑	BPCER(%)↓	HTER(%)↓	AUC(%)↑	BPCER(%)↓
×	×	25.65	79.14	95.93	28.14	79.05	81.34
F.R.	CNN	19.89	87.56	78.71	16.70	91.59	52.60
	Transformer	19.53	86.18	81.69	17.72	90.53	68.14
	GAT	18.17	87.37	78.52	16.47	90.68	62.81
F.E.	CNN	21.58	86.47	65.95	17.94	89.93	57.86
	Transformer	19.23	85.40	91.19	16.76	91.31	56.16
	GAT	17.93	85.97	90.90	16.62	91.78	55.66
F.A.	CNN	18.05	86.34	89.93	16.55	90.50	60.16
	Transformer	20.59	87.72	65.49	20.58	87.57	67.87
	GAT	**16.98**	**90.66**	**58.32**	**13.18**	**94.36**	**43.70**



(a) *F.R.* (b) *F.E.* (c) *F.A.*

Fig. 5. ROC curves for the ablation study when using different models in Cross-Modal Adapter. The experiments are under the cross-dataset setting (Protocol-II) and adopting three different face-related features, which are (a) *F.R.*, (b) *F.E.* and (c) *F.A.*. Baseline (black line) in the figure represents the model without Cross-Modal Adapter.

of 94.36%. These results can verify the contribution of Cross-Modal Adapter to improve the PAD performance and generalization.

Different Backbones To further verify the effectiveness of face related task for PAD, we apply existing PAD methods, including CDC [36], DeepPixBiS [10] and SSDG-R [16] as backbones. We compare the PAD results when above PAD methods with or without face-related features (*F.A.*) from face related tasks on Protocol-II. As shown in Table 3, the proposed face-related features can improve

Table 3. Performance of Other PAD Backbones with or without Face-related Feature from FRT on Protocol-II.

	[O.M] to [C,I]			[C,I] to [O,M]		
	HTER(%)↓	AUC(%)↑	BPCER(%)↓	HTER(%)↓	AUC(%)↑	BPCER(%)↓
CDC [36]	28.94	78.96	86.07	23.30	83.42	74.17
CDC w/FRT	**25.95**	**82.93**	**81.82**	**22.59**	**84.06**	**69.86**
DeepPixBiS [10]	22.93	79.13	100.0	22.45	85.70	75.63
DeepPixBiS w/ FRT	**21.93**	**80.01**	100.0	**19.44**	**88.89**	**65.46**
SSDG-R [16]	20.92	88.07	90.28	22.57	85.61	84.05
SSDG-R w/FRT	**15.63**	**91.77**	**71.46**	**18.68**	**88.51**	**54.92**

performance of all three methods. It indicates that the proposed method can be adopted to existing PAD methods and improve their generalization capability on unseen attacks.

4.3 Comparison with other Methods

To further verify the effectiveness of the proposed method, we compare it with the state-of-the-art methods in two protocols. Table 4 lists the comparison results in Protocol-I. Here, we give our results using the best model (i.e. adopting Step-by-Step Graph as the cross-modal Adapter of the *F.A.* feature). It can be seen that, the proposed FRT-PAD method outperforms the state-of-the-art methods e.g. IF-OM and SSDG-R by the HTER. Specifically, in experiment [I, C, M] to O, our method can outperform both SSDG-R and IF-OM by average 2.60% HTER.

Table 4. Performance Comparison between the Proposed Method and the State-Of-The-Art Methods under the Cross-Dataset Setting. (Protocol-I).

Method	[O, C, I] to M		[O, M, I] to C		[O, C, M] to I		[I, C, M] to O	
	HTER (%)↓	AUC(%)↑	HTER (%)↓	AUC(%)↑	HTER (%)↓	AUC(%)↑	HTER (%)↓	AUC(%)↑
MS-LBP [22]	29.76	78.50	54.28	44.98	50.30	51.64	50.29	49.31
Binary CNN [34]	29.25	82.87	34.88	71.94	34.47	65.88	29.61	77.54
IDA [33]	66.67	27.86	55.17	39.05	28.35	78.25	54.20	44.59
Color Texture [2]	28.09	78.47	30.58	76.89	40.40	62.78	63.59	32.71
LBP-TOP [9]	36.90	70.80	33.52	73.15	29.14	71.69	30.17	77.61
Auxiliary [21]	-	-	28.40	-	27.60	-	-	-
MADDG [28]	17.69	88.06	24.50	84.51	22.19	84.99	27.89	80.02
SSDG-R [16]	7.38	97.17	10.44	95.94	11.71	**96.59**	15.61	91.54
IF-OM [19]	7.14	97.09	15.33	91.41	14.03	94.30	16.68	91.85
Baseline	13.10	92.76	16.44	91.25	24.58	79.50	22.31	85.65
Ours: FRT-PAD	**5.71**	**97.21**	**10.33**	**96.73**	**11.37**	94.79	**13.55**	**94.64**

Moreover, in more challenge Protocol-II (two datasets as training set and other two datasets as test set), we can obtain good results with both adapters (Dense Graph and Step-by-Step Graph) using the *F.A.* feature. As shown in Table 5, the proposed method based on Step-by-Step Graph can outperform other methods by a large margin. Compared with CDC [36], our method can

Table 5. Performance Comparison between the Proposed Method and the State-Of-The-Art Methods under the Cross-Dataset Setting. (Protocol-II).

	[O, M] to [C,I]			[C, I] to [O, M]		
	HTER(%)↓	AUC(%)↑	BPCER(%)↓	HTER(%)↓	AUC(%)↑	BPCER(%)↓
CDC [36]	28.94	78.96	86.07	23.30	83.42	74.17
DeepPixBiS [10]	22.93	79.13	100.0	22.45	85.70	75.63
SSDG-R [16]	20.92	88.07	90.28	22.57	85.61	84.05
IF-OM [19]	18.96	89.48	69.52	18.60	89.76	69.70
Baseline	25.65	79.14	95.93	28.14	79.05	81.34
Ours: FRT-PAD w/ Dense Graph	18.78	87.99	87.93	16.30	92.32	52.73
Ours: FRT-PAD w/ Step-by-Step Graph	**16.98**	**90.66**	**58.32**	**13.18**	**94.36**	**43.70**

improve the HTER of PAD from 23.30% to 13.18% and AUC from 83.42% to 94.36%. These results indicate that the proposed method can generalize better than other PAD methods in both protocols, which further prove the superiority of our method.

5 Conclusion

Existing face presentation attack detection methods cannot generalize well to unseen PAs, due to the highly dependence on the limited datasets. In this paper, to improve generalization ability of face PAD, we proposed a face PAD mechanism using feature-level prior knowledge from face related task in a common face system. By designing a Cross-Modal Adapter, features from other face related tasks can re-map to more effective features for PAD. Experimental results have shown the effectiveness of the proposed method. Compared with the state-of-the-art methods in existing dataset partition (i.e. Protocol-I), we can improve HTER to 5.71%. Furthermore, when the dataset partition becomes more challenging (i.e. Protocol-II where more PAs are unseen to the model), our method largely improve the HTER to 13.18%, which demonstrates the strong generalization ability of our method to handle unpredictable PAs.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (Grant 62076163 and Grant 91959108), and the Shenzhen Fundamental Research Fund (Grant JCYJ20190808163401646). Raghavendra Ramachandra is supported by SALT project funded by the Research Council of Norway.

References

1. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face antispoofing using speeded-up robust features and fisher vector encoding. IEEE Signal Processing Letters **24**(2), 141–145 (2016)
2. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face spoofing detection using colour texture analysis. IEEE Transactions on Information Forensics and Security **11**(8), 1818–1830 (2016)

3. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: Oulu-npu: A mobile face presentation attack database with real-world variations. In: 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017). pp. 612–618. IEEE (2017)
4. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: 2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG). pp. 1–7. IEEE (2012)
5. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR. pp. 4690–4699 (2019)
7. Drozdowski, P., Grobarek, S., Schurse, J., Rathgeb, C., Stockhardt, F., Busch, C.: Makeup presentation attack potential revisited: Skills pay the bills. In: 2021 IEEE International Workshop on Biometrics and Forensics (IWBF). pp. 1–6. IEEE (2021)
8. de Freitas Pereira, T., Anjos, A., De Martino, J.M., Marcel, S.: Can face anti-spoofing countermeasures work in a real world scenario? In: 2013 international conference on biometrics (ICB). pp. 1–8. IEEE (2013)
9. de Freitas Pereira, T., Komulainen, J., Anjos, A., De Martino, J.M., Hadid, A., Pietikäinen, M., Marcel, S.: Face liveness detection using dynamic texture. EURASIP Journal on Image and Video Processing **2014**(1), 1–15 (2014)
10. George, A., Marcel, S.: Deep pixel-wise binary supervision for face presentation attack detection. In: 2019 International Conference on Biometrics (ICB). pp. 1–8. IEEE (2019)
11. George, A., Marcel, S.: Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. IEEE Transactions on Information Forensics and Security **16**, 361–375 (2020)
12. George, A., Marcel, S.: Cross modal focal loss for rgbd face anti-spoofing. In: CVPR. pp. 7882–7891 (2021)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
14. Heusch, G., George, A., Geissbühler, D., Mostaani, Z., Marcel, S.: Deep models and shortwave infrared information to detect face presentation attacks. IEEE Transactions on Biometrics, Behavior, and Identity Science **2**(4), 399–409 (2020)
15. Jia, S., Guo, G., Xu, Z.: A survey on 3d mask presentation attack detection and countermeasures. Pattern Recognition **98**, 107032 (2020)
16. Jia, Y., Zhang, J., Shan, S., Chen, X.: Single-side domain generalization for face anti-spoofing. In: CVPR. pp. 8484–8493 (2020)
17. Karasugi, I.P.A., Williem: Face mask invariant end-to-end face recognition. In: ECCV Workshops. pp. 261–276 (2020)
18. Liu, F., Liu, H., Zhang, W., Liu, G., Shen, L.: One-class fingerprint presentation attack detection using auto-encoder network. IEEE Transactions on Image Processing **30**, 2394–2407 (2021)
19. Liu, H., Kong, Z., Ramachandra, R., Liu, F., Shen, L., Busch, C.: Taming self-supervised learning for presentation attack detection: In-image de-folding and out-of-image de-mixing. arXiv preprint arXiv:2109.04100 (2021)
20. Liu, H., Zhang, W., Liu, F., Wu, H., Shen, L.: Fingerprint presentation attack detector using global-local model. IEEE Transactions on Cybernetics (2021)

21. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: CVPR. pp. 389–398 (2018)
22. Määttä, J., Hadid, A., Pietikäinen, M.: Face spoofing detection from single images using micro-texture analysis. In: 2011 international joint conference on Biometrics (IJCB). pp. 1–7. IEEE (2011)
23. Nguyen, S.M., Tran, L.D., Arai, M.: Attended-auxiliary supervision representation for face anti-spoofing. In: ACCV (2020)
24. Patel, K., Han, H., Jain, A.K.: Secure face unlock: Spoof detection on smartphones. IEEE transactions on information forensics and security **11**(10), 2268–2283 (2016)
25. Raghavendra, R., Raja, K.B., Busch, C.: Presentation attack detection for face recognition using light field camera. TIP **24**(3), 1060–1075 (2015)
26. Ramachandra, R., Busch, C.: Presentation attack detection methods for face recognition systems: A comprehensive survey. ACM Computing Surveys (CSUR) **50**(1), 1–37 (2017)
27. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017)
28. Shao, R., Lan, X., Li, J., Yuen, P.C.: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: CVPR. pp. 10023–10031 (2019)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
30. Wang, G., Han, H., Shan, S., Chen, X.: Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In: CVPR. pp. 6678–6687 (2020)
31. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y.: Suppressing uncertainties for large-scale facial expression recognition. In: CVPR. pp. 6897–6906 (2020)
32. Wang, Y., Song, X., Xu, T., Feng, Z., Wu, X.J.: From rgb to depth: Domain transfer network for face anti-spoofing. IEEE Transactions on Information Forensics and Security **16**, 4280–4290 (2021)
33. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Transactions on Information Forensics and Security **10**(4), 746–761 (2015)
34. Yang, J., Lei, Z., Li, S.Z.: Learn convolutional neural network for face anti-spoofing. arXiv preprint arXiv:1408.5601 (2014)
35. Yu, J., Hao, X., Xie, H., Yu, Y.: Fair face recognition using data balancing, enhancement and fusion. In: ECCV Workshops. pp. 492–505 (2020)
36. Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F., Zhao, G.: Searching central difference convolutional networks for face anti-spoofing. In: CVPR. pp. 5295–5305 (2020)
37. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
38. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: 2012 5th IAPR international conference on Biometrics (ICB). pp. 26–31. IEEE (2012)