Learnable Multi-level Frequency Decomposition and Hierarchical Attention Mechanism for Generalized Face Presentation Attack Detection

Meiling Fang[1,2], Naser Damer[1,2], Florian Kirchbuchner[1,2], Arjan Kuijper[1,2]
[1]Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2]Mathematical and Applied Visual Computing, TU Darmstadt, Darmstadt, Germany
Email: meiling.fang@igd.fraunhofer.de

Abstract

With the increased deployment of face recognition systems in our daily lives, face presentation attack detection (PAD) is attracting much attention and playing a key role in securing face recognition systems. Despite the great performance achieved by the hand-crafted and deep-learning-based methods in intra-dataset evaluations, the performance drops when dealing with unseen scenarios. In this work, we propose a dual-stream convolution neural networks (CNNs) framework. One stream adapts four learnable frequency filters to learn features in the frequency domain, which are less influenced by variations in sensors/illuminations. The other stream leverages the RGB images to complement the features of the frequency domain. Moreover, we propose a hierarchical attention module integration to join the information from the two streams at different stages by considering the nature of deep features in different layers of the CNN. The proposed method is evaluated in the intra-dataset and cross-dataset setups, and the results demonstrate that our proposed approach enhances the generalizability in most experimental setups in comparison to state-of-the-art, including the methods designed explicitly for domain adaption/shift problems. We successfully prove the design of our proposed PAD solution in a step-wise ablation study that involves our proposed learnable frequency decomposition, our hierarchical attention module design, and the used loss function. Training codes and pre-trained models are publicly released [1].

1. Introduction

In recent years, face recognition systems have been widely used in our daily lives for person authentication or access control due to their convenience and remarkable accuracy. However, most existing face recognition systems are vulnerable to Presentation Attacks (PAs). Attackers can use different PAs to impersonate someone or obfuscate their identity. PAs such as print, replay, or 3D mask attacks have been shown to be a serious threat to face recognition systems. Therefore, face Presentation Attack Detection (PAD) plays a critical role in the security of face recognition systems. PAD methods can be broadly categorized into ones based on hand-crafted features [19, 20, 2, 17, 7], and ones based on deep-learning [11, 35, 18, 8, 34]. Hand-crafted based methods utilized traditional texture features such as Local Binary Pattern (LBP) and its extended versions [19, 20, 2] that are robust to some variations, e.g., color texture, noise artifacts, in PAs. However, the extracted features may not be discriminative enough between bona fide and attacks. Recent PAD studies [18, 33, 35, 8] are competing to boost the performance using Convolution Neural Networks (CNNs) to facilitate more discriminative feature learning. However, CNN-based methods have been a risk of overfitting and thus affect the performance generalization over variations, such as unseen sensors or varied illumination conditions. Considering the characteristics of the hand-crafted and deep learning-based features, it is worth exploring the integration of both features for more discriminative and generalized PAD decisions.

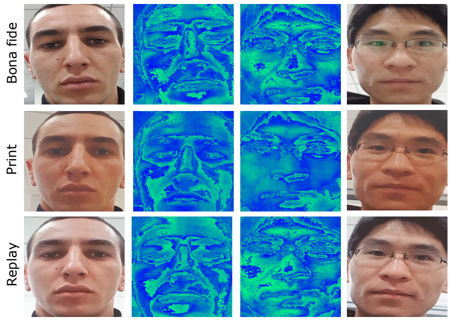


Figure 1: The frequency decomposed image components by DCT and inverse DCT obtained according to Equation 1, for bona fide (top), print PA (middle), and replay PA (bottom). The face images are in OULU-NPU dataset [3]. The lighter blue represents the response to high-frequency. It can be observed that the print attack contains relatively less high-frequency information.

[1] https://github.com/meilfang/LMFD-PAD

In addition to widely used LBP features, several studies [17, 4, 5] attempted to transform images to the frequency domain. Li *et al*. [17] utilized the dissimilarity in Fourier spectra by considering that less high-frequency components exist in attacks compared to bona fide samples. These hand-crafted features are less relevant to the advanced semantic information like identity information but more relevant to the capture conditions, like displayed screen, used photo, or capture sensors. However, most existing hand-crafted features are extracted by static filters, which might limit the representation capacity and make capturing the relevant patterns harder. A recent study [22] proposed the adaptive partition of images in the frequency domain based on a set of learnable frequency filters to detect face forgery clues. In our work, we adapt several learnable filters to capture the PAs cues. Figure 1 presents the Discrete Cosine Transform (DCT) based frequency-aware decomposed images. We can observe that the print attacks have less responses to high frequencies (light blue region) compared to the bona fide and replay attacks. Considering the great progress achieved by the deep learning-based methods, we successfully aim at using CNNs to learn subtle differences between bona fide and attacks on both decomposed components in the frequency domain and RGB images in the spatial domain.

Recently, attention mechanisms were proposed to model the interdependencies between the channel and spatial features on feature maps of CNNs. Woo *et al*. [31] proposed a Convolutional Block Attention Module (CBAM) that can be integrated into any CNN architectures and is end-to-end trainable along with the base CNN. The intermediate feature map is adaptively refined by a combination of channel and spatial wise attention. However, most existing attention-based networks do not consider the nature of features in different layers. The features become more abstract and complex when moving from lower to higher layers in a CNN. The features in the lower layers are relevant to the texture information (e.g., edges), and the features in the higher layers emphasize advanced semantic information. Therefore, simply using a combined channel and spatial attention module may be sub-optimal. In our work, we successfully apply different attention modules according to the nature of the deeply learned features.

In this work, we aim to integrate learned features from the frequency domain and the spatial domain for better PAD generalization capability. The main contributions of this work are: 1) We propose a dual-stream PAD solution based on learnable multi-level frequency decomposition (MFD) and our proposed hierarchical attention mechanism (HAM) to capture discriminative and generalize features from both the spatial and frequency domains, namely the LMFD-PAD; 2) An evaluation in both intra-dataset and cross-dataset settings that demonstrates the superiority of our model in cross-dataset PAD when compared to the state-of-the-at, including the PAD methods explicitly targeting domain adaption/shift problem; 3) An ablation study successfully demonstrates the benefits of the proposed LMFD-PAD components, in a step-wise manner, to the cross-dataset PAD performance.

2. Related work

This section reviews the most relevant prior works by focusing on feature-based and deep learning-based face PAD methods, especially those aiming to demonstrate cross-dataset generalizability.

Feature-based methods: Hand-crafted features, such as Local Binary Pattern (LBP) and image distortion, are utilized broadly to detect presentation attacks. For instance, the commonly used LBP projects the faces to a low-dimension representation and has shown good performance on Idiap Replay-Attack dataset [6]. Boulkenafet *et al*. [1] held an IJCB Mobile Face Anti-Spoofing (IJCB-MFAS) competition [1] carried out on the publicly available OULU-NPU dataset [3] in 2017. The goal of the competition was to evaluate the generalizability of PAD algorithms in a mobile environment. The best performing algorithm among all protocols, named GRADIANT, fused color, texture, and motion information from different color spaces. In addition to LBP, transforming face images into the frequency domain was also previously used. Jourabloo *et al*. [15] used Fast Fourier Transform (FFT) to analyze the spoofing noise. They found that low-frequency features are related to the color distortion and replay artifacts, while high-frequency responses were more obvious on print attacks. Recently, Chen *et al*. [5] fused the high and low-frequency features for advanced generalizability of face PAD. In their work, three fixed filters were used to extract the high-frequency information from the input images, and low-frequency features were extracted by Gaussian blur filters. However, the hand-crafted and fixed filters might fail to cover the complete frequency domain, and it is hard to use them to capture features adaptively. Thus, Qian *et al*. [22] proposed a set of learnable frequency filters for face forgery detection. In our work, we adapt three learnable filters as suggested in [22] and add one more general filter to obtain the frequency-aware decomposed image components, which is complemented by RGB images.

Deep learning-based methods: Deep learning-based methods have been pushing the frontier of face PAD research and have shown significant improvement in PAD performance. George *et al*. [11] proposed a PAD based on pixel-wise and binary supervised (DeepPixBis) training. However, the DeepPixBis method did not generalize well on unseen attacks/sensors scenarios. To further improve the intra-dataset performance and increase the generalization capability, some studies use auxiliary information, e.g., depth [35] and Remote Photoplethysmography (rPPG) signals [18], for training supervision. For example, Yu *et al.* [34] proposed Neural Architecture Search based method for face PAD (NAS-FAS) based on their previous work on Center Difference Convolution Network (CDCN) [35]. They obtained significantly improved results in both intra-dataset and cross-dataset experimental settings. However, the expensive computation cost of NAS must be considered, and the higher error rates in the cross-dataset scenarios suggest that the generalizability is still an open problem. Several methods explicitly targeted the domain generalization problem as an inherent domain shift that can be found between different face PAD datasets. Saha *et al.* [23] proposed a class-conditional domain discriminator module to generate discriminative bona fide and attack features to tackle the domain shift problem. Most domain generalization face PAD methods [24, 16, 25, 23] performed experiments on four publicly available dataset: Oulu-NPU [3], CASIA-MFSD [37], Idiap Replay-Attack [6], and MSU-MFSD [29]. We follow this cross-dataset setting to compare our method against those state-of-the-art methods later in this work (as reported in Section 4.2.2).

3. Methodology

In this section, we will provide details of our LMFD PAD solution. We will introduce the multi-level frequency decomposition (MFD), including four learnable frequency filters in Section 3.1. Then we introduce the dual-stream network architecture where using a hierarchical attention mechanism to integrate the features learned in frequency and spatial domain in Section 3.2, and at last present the used loss functions in Section 3.3.

3.1. Multi-level Frequency Decomposition (MFD)

Deep-learning based face PAD methods achieved great progress in intra-dataset evaluations. However, the performance normally drops drastically when testing on unseen datasets [21]. This might be caused by the variations in the attacks and capture environments, such as illuminations and sensors. To address this issue, our proposed LMFD solution decomposes an input face image into different level frequency components. Frequency domain analysis is a classical method in image signal processing and has been widely used for general image classification and texture classification tasks [26, 12]. Moreover, some face PAD methods attempted [17, 4, 5] to transform the images in frequency domain and mine the artifacts cues. The results showed that features in the frequency domain are less sensitive to the variations of the capture environments (e.g., sensors or light conditions). However, most existing frequency-based face PAD methods used filters with fixed weight and maybe sub-optimal for discriminative feature learning.

In our work, we use a set of adaptively learnable frequency filters described in [22] for face forgery detection. First, N manually designed binary base filters $\mathcal{F}_b = \{f_b^i | 1 \leq i \leq N\}$ partition the frequency domain into low, middle, high frequency bands. The goal of the binary base filters is a roughly equal division of spectrum intp N bands from low frequency to high frequency. Then, N learnable filters $\mathcal{F}_l = \{f_l^i | 1 \leq i \leq N\}$ are added to such binary base filters. The benefit of such learnable filters is the adaptive selection of the frequency of interest beyond the fixed base filters. Finally, a decomposed image component C_i of an input image x can be computed following the equation:

$$\mathrm{C}_i = \mathcal{D}^{-1}\{\mathcal{D}(x) \odot [f_b^i + \sigma(f_l^i)]\}, i = \{1, ..., N\}, \quad (1)$$

where $\mathcal{D}$ is DCT, $\mathcal{D}^{-1}$ is inversed DCT, and $\odot$ is the element-wise product. The $\sigma(f) = \frac{1-exp(-f)}{1+exp(-f)}$ is used to normalize the value of f between -1 and $+1$.

In our case, N is set to 4 to obtain explicitly divided frequency domain of low, middle, and high-frequency bands and the complementary full frequency band. Three bands are chosen as described in [22]: 1) the low frequency band f_{base}^1 is the first 1/16 of the entire spectrum, 2) the middle frequency band f_{base}^2 is between 1/16 and 1/8 of the entire spectrum, 3) the high frequency band f_{base}^3 is between 1/8 and 7/8 of the entire spectrum. However, the partitioned frequencies may not be sufficient to obtain subtle cues between bona fide and attacks. Therefore, we add one additional learnable filter f_{base}^4 where the frequency band is the entire spectrum. Moreover, we also keep the input RGB image to provide more visual information and complementary to frequency domain information (as shown in Figure 2).

In the experiments, face detection is firstly performed on the input image by MTCNN framework [36]. Then, the detected RGB face image is resized to $224 \times 224 \times 3$ pixels. According to the Equation 1, four obtained components are stacked along the channel axis, i.e, the size of a stacked decomposition is $224 \times 224 \times 12$. Then, we utilize dual-stream (RGB and MFD) networks to extract different features in a face image (see Figure 2). In our work, we use the ResNet-50 [13] as our backbone network.

3.2. Hierarchical Attention Mechanism (HAM)

So far, we use the dual-stream to learn discriminate features in parallel, which may be sub-optimal for a final PAD decision. To enhance that, we propose our hierarchical attention mechanism (HAM) to integrate features from the frequency domain and semantic image domain and to utilize the features from different layers in the dual-stream.

This HAM is inspired by Convolutional Block Attention Mechanism (CBAM) [31], which proposed channel and spatial attention blocks for the general computer vision task, and Attention Pixel-wise Binary Supervision (A-PBS) method [10], which employed and fused spatial attention features from multi-layers for the iris PAD task. The

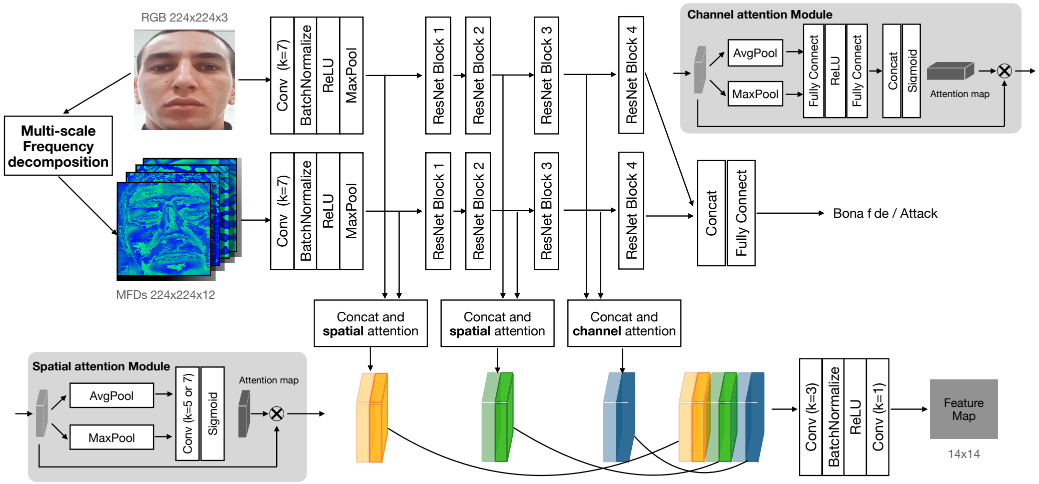


Figure 2: The overall workflow of our proposed LMFD-PAD solution. Note the utilization of our MFD and HAM (three different channel attention components) components.

CBAM [31] consisting of the channel, and distinctive spatial sub-modules can be added into networks according to the custom design needs and showed improvements in classification and detection performance with various neural architectures. A-PBS method [10] adopted only spatial attention module (i.e., no channel attention module) aiming to locate the most informative region in an RGB eye image, where might contribute most to a PAD decision. However, in our MFD stream, we have multi-level frequency features, and the weights of filters are adaptively learning while the model is training. The high-frequency component emphasizes features like edges and texture information, while the low-frequency component is related to the spatial distribution of the color gamut. Therefore, channel attention is additionally applied in our framework.

Figure 2 shows that spatial attention modules are inserted after the first convolution block and the second ResNet block, respectively, while a channel attention module is added following the third ResNet block. The reason for such attention modules arrangement is based on the nature of the features extracted from different layers. The features from lower to higher layers become more abstract and complex. More specifically, the features in the lower layers are related to the appearance and texture cues, and the features in the higher layers might reveal the semantic content information. Consequently, we perform a spatial attention module on a fused feature in lower layers to focus on texture details like the edge. Then, a channel attention module is added after the third ResNet block to learn the advanced semantic features. To be consistent with the observation on the nature of features in different layers, the size of the convolutional kernel is 7 × 7 in the first spatial attention module and 5 × 5 in the second spatial attention module, as the smaller convolutional kernel is more suitable for locating the small-scale texture cues. Finally, the attentive features are fused to preserve richer patterns. Moreover, we use pixel-wise and binary supervision to train the dual-stream networks as suggested in [11] where the intermediate feature map can be considered as the scores generated from the patches in an image and thus improve the performance. On the one hand, the attentive feature maps from different layers are concatenated and fed to the stacked two convolution layers to output a feature map. The size of the output feature map in our case is 14 × 14 for pixel-wise supervision. On the other hand, the features from the last ResNet block in two streams are also concatenated and fed to the fully connected layer for binary supervision.

3.3. Loss function

Binary Cross Entropy (BCE) loss has proved to perform well when used for pixel-wise and binary supervision [11]. Nevertheless, to reduce the sensitivity to outliers in the output feature map, we use the Smooth L1 (SL) function to compute the loss between the output feature map and the ground truth binary mask. For binary supervision, we use the Focal Loss instead of BCE loss because the Focal loss (FL) with a relaxing factor can down-weight easy samples (i.e., samples correctly classified with high confidence) and make the model focus on the hard samples with low classification confidence. The equation for Smooth L1 is shown below as:

$$\mathcal{L}_{SL} = \frac{1}{n}\sum z$$

$$\text{where} \quad z = \begin{cases} \frac{1}{2}\cdot(y-x)^2, & if \quad |y-x| < 1 \\ |y-x| - \frac{1}{2}, & otherwise \end{cases}$$

where n is the number of pixels in the output map (14 in our case). x and y refer to the values in the output feature map and the ground truth label, respectively. The equation of Focal loss is:

$$\mathcal{L}_{FL} = -(1 - p_t)^{\gamma} \log(p_t)$$

$$\text{where} \quad p_t = \begin{cases} p, & \textit{if} \quad y = 1 \\ 1 - p, & \textit{otherwise} \end{cases}$$

where p is the predicted probability when the ground truth label y is 1 (bona fide in our case) and γ is a tunable focusing parameter (γ is 2 in our experiments). The overall loss function is given as:

$$\mathcal{L}_{overall} = \lambda_1 \cdot \mathcal{L}_{SL} + \lambda_2 \cdot \mathcal{L}_{FL} \tag{2}$$

For exploring the effect of loss functions, we also report the results of BCE loss as used in [11] as an ablation study (as shown in Table 3).

4. Experiments

4.1. Experimental setting

Datasets: Our method is evaluated on four publicly available face PAD datasets: Oulu-NPU [3], CASIA-MFSD [37], Idiap Replay-Attack [6], and MSU-MFSD [29] under different scenarios. Oulu-NPU [3] dataset consists of 55 subjects and 5940 videos recorded by six mobile phones. Four protocols are provided to evaluate the generalizability of algorithms. Protocol-1 studies the impact of illumination variations, while Protocol-2 evaluates different attacks created by various instruments. Protocol-3 examines the effect of different capture cameras, and Protocol-4 explores all the challenges above by leave-one-out cross-validation. CASIA-MFSD [37] includes 50 subjects and 600 videos captured by three different quality cameras. This dataset contains three attack types: warped photo attack, cut photo attack, and video replay attack. Idiap Replay-Attack [6] contains 50 subjects and 300 videos captured by different sensors and different illumination conditions. Moreover, two types of attacks are included in this dataset: print and replay attacks. MSU-MFSD [29] contains 35 subjects and 440 videos captured by two different resolutions of cameras. This dataset also includes two types of attacks, printed photo attacks and replay attacks. The videos in datasets are recorded under different environments with variant cameras and subjects, suitable for cross-dataset domain generalization protocol. Moreover, the subjects in the training set and test set are disjoint in intra-dataset settings.

Implementation details: The proposed dual-stream networks are based on ResNet-50 [13] with pre-trained weights on the ImageNet dataset [9]. The data in all PAD datasets are videos, thus, we sample 10 frames in the average time interval of each video to train and test our method. For each frame, the face is detected and cropped by the MTCNN method [36] and resized to $224 \times 224 \times 3$ pixels. In the training phase, the SGD optimizer is used with an initial learning rate of 0.001, the momentum of 0.9, and a weight decay of 0.0001. Then, the exponential learning rate scheduler is used with a multiplicative factor of the learning rate decay value (γ) of 0.995. The ratio of bona fide and attack data is close to 1:1 by simply duplicating the needed images to reduce the effect of biased data. Several data augmentation techniques are used for better generalization ability, including horizontal flip, rotation, cutout, RGB channel shift, and color jitter. To further reduce overfitting, the early stopping technique is utilized with the maximum epochs of 100 and the patience epochs of 15. The batch size in the training phase is 32. In our experiments, the λ_1 in overall loss function 2 is set manually to 1 at the beginning of the training and changed to 100 after five training epochs, while λ_2 is set to 1 in the whole training phase. In the testing phase, a final PAD decision score of a video is a fused score (mean-rule fusion) of all frames.

Evaluation metrics: We follow the sub-protocols and metrics as defined in the competition [1] which was performed on the OULU-NPU [3] dataset for a fair comparison. The Attack Presentation Classification Error Rate (APCER) [14] is computed separately for each presentation attack instrument (PAI), e.g., print or replay following the equation:

$$ACPER_{PAI} = \frac{1}{N_{PAI}} \sum_{i=1}^{N_{PAI}} (1 - p_i) \tag{3}$$

where N_{PAI} is the number of attack samples for a given PAI, p_i is the predicted binary label of the i^{th} presentation (0 for bona fide and 1 for attack). Then, following the OULU-NPU protocol [3], APCER$_{wc}$ is the highest APCER is selected to report the overall performance, i.e., the worst case among all the presentation instruments. The equation is APCER$_{wc}$ = $\max$ (APCER$_{PAI}$) among all PAIs. Bona Fide Presentation Classification Error Rate (BPCER) [14] is the proportion of incorrectly classified bona fide samples. Average Classification Error Rate (ACER) is the mean of APCER$_{wc}$ and BPCER. Moreover, to report the cross-dataset results and to be consistent with previous works [2, 18, 34, 16, 25], we report Half Total Error Rate (HTER) and Area Under the receiver operating Curve (AUC) are used for the cross-dataset domain generalization protocol on OULU-NPU [3] , CASIA-MFSD [37], Idiap Replay-Attack [6] and MSU-MFSD [29] datasets. The HTER is half of the sum of the APCER and BPCER.

4.2. Comparison with the State-of-the-Art Methods

4.2.1 Intra-dataset results on OULU-NPU

An IJCB-MFAS competition [1] was carried out on the publicly available OULU-NPU dataset. To assess the generalizability of the face PAD methods, four protocols are provided consisting of cross-environment, cross-PAIs, cross-sensors, cross-all scenarios. For a fair comparison, we strictly follow the definition and evaluation metric of those protocols.

In this study, we compare our LMFD-PAD method with the best performing method in IJCB-MFAS competition [1], GRADIANT. Moreover, we also compare with several recently PAD methods: Auxiliary [18], FAS-TD [28], STASN [33], DeepPixBis [11], CDCN++ [35], SSR-FCN [8], NAS-FAS [34] proposed from 2018 to 2021. The results are reported in Table 1. [2] The LMFD-PAD achieved ACER values of each protocol are 1.5%, 2.0%, 3.4%, and 3.3%, respectively. It can be observed that our method obtain competitive results in comparison to state-of-the-art methods. For example, the lowest ACER in the most challenging Protocol-4 is 2.9% achieved by NAS-FAS [34], while our LMFD-PAD ACER value is 3.3%. This result indicates that our model generalizes well on the cross-test scenarios. Considering that we employ pixel-wise supervision, we can group those PAD methods into three groups based on supervision manner for further comparison. GRADIANT [1] and STASN [33] was trained only by binary supervision. DeepPixBis [11], SSR-FCN [8] and our method utilized the pixel-wise and binary supervision. The left four PAD approaches used depth or/and rPPG supervision. It can be found in Table 1 that our method possesses improved performance compared to pixel-wise and binary supervised models in most cases but scored below the depth/rPPG supervised networks in some cases. This might drive an extension of our work by generating depth or/and rPPG information to improve the intra-dataset performance. In this case, however, the trade-off between computational resource/time and performance needs to be considered.

Prot.	Method	APCER$_{wc}$(%)	BPCER(%)	ACER(%)
1	GRADIANT [1]	1.3	12.5	6.9
	Auxiliary [18]	1.6	1.6	1.6
	FAS-TD [28]	2.5	0.0	1.3
	STASN [33]	1.2	2.5	1.9
	DeepPixBis [11]	0.8	0.0	0.4
	CDCN++ [35]	0.4	0.0	**0.2**
	SSR-FCN [8]	1.5	7.7	4.6
	NAS-FAS [34]	0.4	0.0	**0.2**
	LMFD-PAD (ours)	1.4	1.6	1.5
2	GRADIANT [1]	3.1	1.9	2.5
	Auxiliary [18]	2.7	2.7	2.7
	FAS-TD [28]	1.7	2.0	1.9
	STASN [33]	4.2	0.3	2.2
	DeepPixBis [11]	11.4	0.6	6.0
	CDCN++ [35]	1.8	0.8	1.3
	SSR-FCN [8]	3.1	3.7	3.4
	NAS-FAS [34]	1.5	0.8	**1.2**
	LMFD-PAD (ours)	3.1	0.8	2.0
3	GRADIANT [1]	2.6 ± 3.9	5.0 ± 5.3	3.8 ± 2.4
	Auxiliary [18]	2.7 ± 1.3	3.1 ± 1.7	2.9 ± 1.5
	FAS-TD [28]	5.9 ± 1.9	5.9 ± 3.0	5.9 ± 1.0
	STASN [33]	4.7 ± 3.9	0.9 ± 1.2	2.8 ± 1.6
	DeepPixBis [11]	11.7 ± 19.6	10.6 ± 14.1	11.1 ± 9.4
	CDCN++ [35]	1.7 ± 1.5	2.0 ± 1.2	1.8 ± 0.7
	SSR-FCN [8]	2.9 ± 2.1	2.7 ± 3.2	2.8 ± 2.2
	NAS-FAS [34]	2.1 ± 1.3	1.4 ± 1.1	**1.7 ± 0.6**
	LMFD-PAD (ours)	3.5 ± 3.2	3.3 ± 3.2	3.4 ± 3.1
4	GRADIANT [1]	5.0 ± 4.5	15.0 ± 7.1	10.0 ± 5.0
	Auxiliary [18]	9.3 ± 5.6	10.4 ± 6.0	9.5 ± 6.0
	FAS-TD [28]	14.2 ± 8.7	4.2 ± 3.8	9.2 ± 3.4
	STASN [33]	6.7 ± 10.6	8.3 ± 8.4	7.5 ± 4.7
	DeepPixBis [11]	36.7 ± 29.7	13.3 ± 14.1	25.0 ± 12.7
	CDCN++ [35]	4.2 ± 3.4	5.8 ± 4.9	5.0 ± 2.9
	SSR-FCN [8]	8.3 ± 6.8	13.3 ± 8.7	10.8 ± 5.1
	NAS-FAS [34]	4.2 ± 5.3	1.7 ± 2.6	**2.9 ± 2.8**
	LMFD-PAD (ours)	4.5 ± 5.3	2.5 ± 4.1	3.3 ± 3.1

Table 1: The results of the intra-dataset evaluation under the four protocols of the OULU-NPU dataset [3]. The bold numbers refer to the lowest ACER in each protocol. Note that our LMFD-PAD achieves competitive performance overall and performs better than most methods that do not use auxiliary information (depth or rPPG) as detailed in Section 4.2.1.

4.2.2 Cross-dataset results

In the cross-dataset scenario, four publicly available face PAD datasets: Oulu-NPU [3] (O for short), CASIA-MFSD [37] (C for short), Idiap Replay-Attack [6] (I for shot), and MSU-MFSD [29] (M for short) are used. Three datasets are randomly selected for training and the remained one is used for testing. Specifically, following previous works targeting the domain adaption and generalization capability of face PAD [16, 24, 25, 23], four settings are performed: O&C&I → M, O&M&I → C, O&C&M → I and I&C&M → O.

In our work, we compare our LMFD-PAD model against eight state-of-the-art face PAD methods including depth/rPPG supervision based Auxiliary [18] and NAS-FAS [34] which outperformed in intra-testing on OULU-NPU dataset [3]. In addition, we also compare our method with four state-of-the-art domain generalization face PAD methods: MMD-AAE [16], MADDG [24], RFMetaFAS [25], and CCDD [23], which explicitly target the domain shift problem. The results are reported in Table 2 where the last four methods are face methods addressing domain shift problems. Our proposed LMFD-PAD method achieves significantly improved performance in three experiment settings. For example, the HTER value of our model is 10.48% in O&C&I → M setting and 12.50% in O&M&I → C and 12.41% in I&C&M → O, while the second-ranking results in those settings are 13.89%, 15.21%, and 13.16%, respectively. Although our LMFD-PAD method is not explicitly designed for the domain shift problem, our method obtains better performance than domain generalization face PAD methods in most cases. The cross-dataset results are consistent with the result in the most challenging intra-dataset Protocol-4 of OULU-NPU dataset [3]. We conclude that our method is able to learn more generalized features, which perform well on unseen domains. However, it is still unclear which part of our model benefits the improved results. This question will be answered in the following section by

[2] The results of state-of-the-art solutions listed in Table 1 and 2 are those reported in their paper.

Method	O&C&I → M		O&M&I → C		O&C&M → I		I&C&M → O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MS LBP [19]	29.76	78.50	54.28	44.98	50.30	51.64	50.29	49.31
Binary CNN [32]	29.25	82.87	34.88	71.94	34.47	65.88	29.61	77.54
IDA [30]	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
LBPTOP [20]	36.90	70.80	42.60	61.05	49.45	49.54	53.15	44.09
Auxiliary(Depth Only) [18]	22.72	85.88	33.52	73.15	29.14	71.69	30.17	77.61
Auxiliary(All) [18]	-	-	28.40	-	27.60	-	-	-
NAS-FAS [34]	16.85	90.42	15.21	92.64	**11.63**	**96.98**	13.16	94.18
MMD-AAE [16]	27.08	83.19	44.59	58.29	31.58	75.18	40.98	63.08
MADDG [24]	17.69	88.06	24.50	84.51	22.19	84.99	27.98	80.02
RFMetaFAS [25]	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16
CCDD [23]	15.42	91.13	17.42	90.12	15.87	91.72	14.72	93.08
LMFD-PAD (ours)	**10.48**	**94.55**	**12.50**	**94.17**	18.49	84.72	**12.41**	**94.95**

Table 2: The results of the cross-dataset evaluation under different experimental settings on four face PAD datasets. In each setting, three datasets are used for training, and one remaining dataset is used for testing. Our LMFD-PAD method is compared with state-of-the-art face PAD methods reporting on this protocol. Not that four of the state-of-the-art methods MMD-AAE, MADDG, RFMetaFAS, and CCDD are explicitly designed to target the domain shift problem. The bold numbers indicate the lowest HTER and highest AUC in each setting.

RGB	MFD	HAM	BCE	FL+SL	O&C&I → M		O&M&I → C		O&C&M → I		I&C&M → O	
					HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
√			√		17.14	90.47	22.12	82.10	24.62	82.28	19.47	88.16
√	√		√		15.47	93.17	17.21	87.50	23.51	83.25	17.26	90.41
√	√	√	√		11.19	93.39	16.83	90.62	21.42	83.92	22.27	85.98
√	√	√		√	**10.48**	**94.55**	**12.50**	**94.17**	**18.49**	**84.72**	**12.41**	**94.95**

Table 3: The results of the ablation study on model inputs, components, and loss functions. The ablation study is performed on cross-dataset experimental settings to uncover the components generalizability benefits. One can note that in most experiments, each of the proposed components contributes positively to the cross-dataset PAD performance.

exploring the effect of the MFD, HAM parts, and loss function in an ablation study.

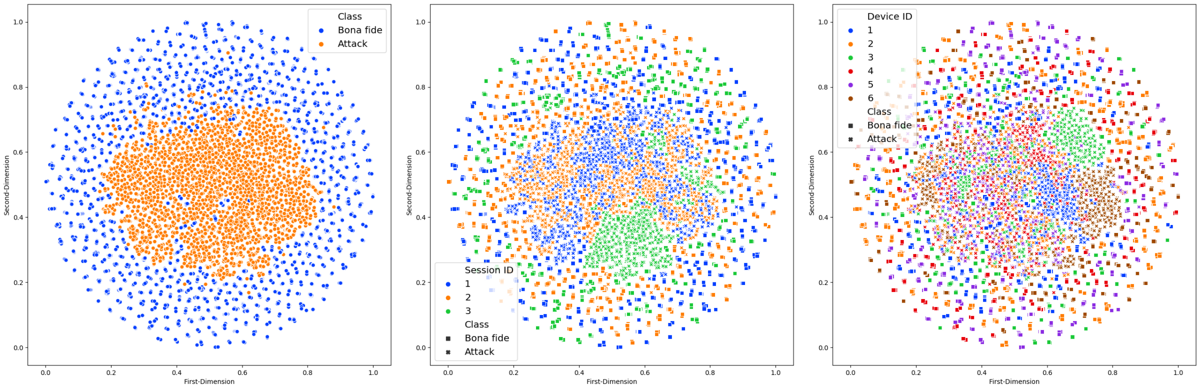
4.3. Ablation study on model components

So far, the results in Table 1 and Table 2 are obtained by our *full model* including the MFD, HAM and a combined loss function of Focal loss and Smooth L1 loss (Equation 2). However, the detailed effect of each part is unknown. Therefore, we present an ablation study on model components, and the results are summarized in Table 3. This aims at understanding the generalization benefits of each of the proposed components. The training hyper-parameters are the same for all combinations in Table 3 (training details are described in Section 4.1). Since we assume that our method is able to learn discriminative and generalize features, the ablation study is demonstrated under the cross-dataset experimental setups on four datasets.

Impact of MFD: To explore the effect of the learnable frequency decomposition, we train a one-stream network using only RGB face images as input and a dual-stream network consisting of RGB and MFD, both solutions are trained by minimizing the BCE loss. The results in Table 3 shows the improvement by the additional MFD component (the HTER is decreased from 17.14% to 15.47% in the O&C&I → M setting). A consistent performance enhancement is seen under all the experimental setups in Table 3.

Impact of HAM: In contrast to learning the features in the image and the frequency domains in parallel and fusing such features just before the classification layer, we add the HAM component to fuse such features earlier followed by different attention blocks according to the levels of layers, as described in Section 3.2. The corresponding results are reported in the second row and third row of Table 3 where it is noticeable that the addition of the proposed HAM did enhance the performance in most experimental settings.

Impact of loss function: In our LMFD-PAD solution, we use the Focal loss to supervise the binary label prediction and the Smooth L1 to supervise the feature map label prediction instead of the commonly used BCE loss. To explore the effect of such modification, we compare it to using the BCE loss for pixel-wise and binary supervision. The weights of both BCE losses is set to 0.5 as used in [11]. As presented in Table 3, the loss combination used in our LMFD-PAD solution strongly enhances the PAD performance across all the cross-dataset experimental settings.



(a) Bona fide and attack.

(b) Three different capture scenarios.

(c) Six different capture devices.

Figure 3: t-SNE visualization of a cross-dataset setting I&C&M → O using our LMFD-PAD embeddings, where the test set is OULU-NPU dataset consisting of three capture environments with different illumination conditions and six mobile devices. The first t-SNE plot represents the two classes: bona fide (blue) and attack (orange). The second and third t-SNE plot indicates three capture scenarios and six capture devices, respectively. In Figures 3b and 3c, each color corresponds to an environment or device, the signs ■ and **x** refers to bona fide and attack, respectively. It is noted that the embeddings from the LMFD-PAD still find a common pattern between the attacks captured under different settings.

We conclude that our LMFD-PAD full model boosts the performance generalizability further by adding each of the MFD, HAM, and a combined loss function.

4.4. Visualization and analysis

In our assumption, the MFD module is able to learn rich generalizable features that adapt well to unseen datasets, especially for unseen sensors or illumination. To further verify this assumption, we use t-SNE [27] plots to visualize deep features in the cross-dataset case I&C&M → O. This setting is chosen because the unseen test set is OULU-NPU dataset [3] consisting of more variation of environment and capture devices and thus it is better for visualization. The deep features are extracted from the last convolution layer before the classification layer, and then the Principal Component Analysis (PCA) is used to reduce the dimensionality of features to 128-D to reduce the computational cost of the t-SNE. Such features are then projected to 2-D features by t-SNE. Figure 3 depicts t-SNE plots on two classes (bona fide and attack), three capture environments, and six capture devices from left to right. As seen in Figure 3a, bona fides and attacks can be considered as coarsely non-linearly separable. This indicates that our model learns discriminative and generalizes features between bona fides and attacks. In Figure 3b, blue, orange, and green represent three environments of various illuminations. It can be seen that different environments are more obviously clustered in the attack category, while they are clustered more randomly in the bona fide category. A similar observation can be found on different mobile devices in Figure 3c. These findings suggest that our model is able to mine the general attack artifacts patterns across data capture variety and thus generalizability on unseen datasets is less effect by different sensors or illuminations. This confirms the achieved cross-dataset results in Section 4.2.2.

5. Conclusion

In this work, we proposed a learnable multi-level frequency decomposition based face PAD method, LMFD-PAD, targeting the generalizability of PAD performance. We employed a dual-stream network architecture. The first stream learns discriminative features in the frequency domain by using learnable frequency filters to obtain frequency decomposed image components, while the other stream uses RGB face images as input to learn features in the spatial domain. Moreover, we proposed the hierarchical attention mechanism to fuse features from both domains at different stages of the network. A spatial attention module is added at the lower layers of the CNN to capture the texture features, and the channel attention module is added at the higher layers of CNN to obtain advanced semantic information. The experiments are demonstrated under intra-dataset and cross-dataset settings. Our LMFD-PAD method achieved comparable results in intra-dataset scenarios. Moreover, in most cross-dataset cases, our proposed solution outperforms state-of-the-art face PAD methods, including the methods addressing the domain adaption/shift and generalization capability problem. The proposed components of our LMFD-PAD solution are additionally proved in a step-wise ablation study.

Acknowledgements: This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Zinelabdine Boulkenafet, Jukka Komulainen, Zahid Akhtar, Azeddine Benlamoudi, Djamel Samai, Salah Eddine Bekhouche, Abdelkrim Ouafi, Fadi Dornaika, Abdelmalik Taleb-Ahmed, Le Qin, Fei Peng, L. B. Zhang, Min Long, Shruti Bhilare, Vivek Kanhangad, Artur Costa-Pazo, Esteban Vázquez-Fernández, Daniel Pérez-Cabo, J. J. Moreira-Perez, Daniel González-Jiménez, A. Mohammadi, Sushil Bhattacharjee, Sébastien Marcel, S. Volkova, Y. Tang, N. Abe, L. Li, X. Feng, Z. Xia, X. Jiang, S. Liu, R. Shao, Pong C. Yuen, Waldir R. de Almeida, Fernanda A. Andaló, Rafael Padilha, Gabriel Bertocco, William Dias, Jacques Wainer, Ricardo da Silva Torres, Anderson Rocha, Marcus A. Angeloni, Guilherme Folego, Alan Godoy, and Abdenour Hadid. A competition on generalized software-based face presentation attack detection in mobile scenarios. In *IJCB*, pages 688–696. IEEE, 2017.

[2] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face spoofing detection using colour texture analysis. *IEEE Trans. Inf. Forensics Secur.*, 11(8):1818–1830, 2016.

[3] Zinelabdine Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. In *FG*, pages 612–618. IEEE Computer Society, 2017.

[4] Baoliang Chen, Wenhan Yang, Haoliang Li, Shiqi Wang, and Sam Kwong. Camera invariant feature learning for generalized face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.*, 16:2477–2492, 2021.

[5] Baoliang Chen, Wenhan Yang, and Shiqi Wang. Face anti-spoofing by fusing high and low frequency features for advanced generalization capability. In *MIPR*, pages 199–204. IEEE, 2020.

[6] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, volume P-196 of *LNI*, pages 1–7. GI, 2012.

[7] Naser Damer and Kristiyan Dimitrov. Practical view on face presentation attack detection. In Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016.

[8] Debayan Deb and Anil K. Jain. Look locally infer globally: A generalizable face anti-spoofing approach. *IEEE Trans. Inf. Forensics Secur.*, 16:1143–1157, 2021.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009.

[10] Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. Iris presentation attack detection by attention-based and deep pixel-wise binary supervision network. In *IJCB*, pages 1–8. IEEE, 2021.

[11] Anjith George and Sébastien Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *ICB*, pages 1–8. IEEE, 2019.

[12] George M. Haley and B. S. Manjunath. Rotation-invariant texture classification using a complete space-frequency model. *IEEE Trans. Image Process.*, 8(2):255–269, 1999.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.

[14] International Organization for Standardization. ISO/IEC DIS 30107-3:2016: Information Technology – Biometric presentation attack detection – P. 3: Testing and reporting, 2017.

[15] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *ECCV (13)*, volume 11217 of *Lecture Notes in Computer Science*, pages 297–315. Springer, 2018.

[16] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409. IEEE Computer Society, 2018.

[17] Jiangwei Li, Yunhong Wang, Tieniu Tan, and Anil K. Jain. Live face detection based on the analysis for fourier spectra. In *Biometric Technology for Human Identification*, volume 5404, pages 296–303, 2004.

[18] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, pages 389–398. IEEE Computer Society, 2018.

[19] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *IJCB*, pages 1–7. IEEE Computer Society, 2011.

[20] Tiago Freitas Pereira, Jukka Komulainen, André Anjos, José Mario De Martino, Abdenour Hadid, Matti Pietikäinen, and Sébastien Marcel. Face liveness detection using dynamic texture. *EURASIP J. Image Video Process.*, 2014:2, 2014.

[21] Sandip Purnapatra, Nic Smalt, Keivan Bahmani, Priyanka Das, David Yambay, Amir Mohammadi, Anjith George, Thirimachos Bourlai, Sébastien Marcel, Stephanie Schuckers, Meiling Fang, Naser Damer, Fadi Boutros, Arjan Kuijper, Alperen Kantarci, Basar Demir, Zafer Yildiz, Zabi Ghafoory, Hasan Dertli, Hazim Kemal Ekenel, Son Vu, Vassilis Christophides, Dashuang Liang, Guanghao Zhang, Zhanlong Hao, Junfu Liu, Yufeng Jin, Samo Liu, Samuel Huang, Salieri Kuei, Jag Mohan Singh, and Raghavendra Ramachandra. Face liveness detection competition (livdet-face) - 2021. In *IJCB*, pages 1–10. IEEE, 2021.

[22] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV (12)*, volume 12357 of *Lecture Notes in Computer Science*, pages 86–103. Springer, 2020.

[23] Suman Saha, Wenhao Xu, Menelaos Kanakis, Stamatios Georgoulis, Yuhua Chen, Danda Pani Paudel, and Luc Van Gool. Domain agnostic feature learning for image and video based face anti-spoofing. In *CVPR Workshops*, pages 3490–3499. IEEE, 2020.

[24] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, pages 10023–10031. Computer Vision Foundation / IEEE, 2019.

[25] Rui Shao, Xiangyuan Lan, and Pong C. Yuen. Regularized fine-grained meta face anti-spoofing. In *AAAI*, pages 11974–11981. AAAI Press, 2020.

[26] José Augusto Stuchi, Marcus A. Angeloni, Rodrigo F. Pereira, Levy Boccato, Guilherme Folego, Paulo Victor de Souza Prado, and Romis Ribeiro Faissol Attux. Improving image classification with frequency domain layers for feature extraction. In *MLSP*, pages 1–6. IEEE, 2017.

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[28] Zezheng Wang, Chenxu Zhao, Yunxiao Qin, Qiusheng Zhou, and Zhen Lei. Exploiting temporal and depth information for multi-frame face anti-spoofing. *CoRR*, abs/1811.05118, 2018.

[29] Di Wen, Hu Han, and Anil K. Jain. Face spoof detection with image distortion analysis. *IEEE Trans. Inf. Forensics Secur.*, 10(4):746–761, 2015.

[30] Di Wen, Hu Han, and Anil K. Jain. Face spoof detection with image distortion analysis. *IEEE Trans. Inf. Forensics Secur.*, 10(4):746–761, 2015.

[31] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In *Computer Vision - 15 th ECCV 2018, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 3–19. Springer, 2018.

[32] Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *ECCV (3)*, volume 8691 of *Lecture Notes in Computer Science*, pages 628–643. Springer, 2014.

[33] Xiao Yang, Wenhan Luo, Linchao Bao, Yuan Gao, Dihong Gong, Shibao Zheng, Zhifeng Li, and Wei Liu. Face anti-spoofing: Model matters, so does data. In *CVPR*, pages 3507–3516. Computer Vision Foundation / IEEE, 2019.

[34] 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 Trans. Pattern Anal. Mach. Intell.*, 43(9):3005–3023, 2021.

[35] 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*, pages 5294–5304. IEEE, 2020.

[36] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503, 2016.

[37] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z. Li. A face antispoofing database with diverse attacks. In *ICB*, pages 26–31. IEEE, 2012.