

Spoof Face Detection Via Semi-Supervised Adversarial Training

Chengwei Chen^a, Wang Yuan^a, Xuequan Lu^{b, **}, Lizhuang Ma^a

^a*East China Normal University, North Zhongshan Road Campus: 3663 N, Shanghai and 200062, China*

^b*Deakin University, 75 Piggons Rd, Waurn Ponds VIC 3216, Australia*

ABSTRACT

Face spoofing causes severe security threats in face recognition systems. Previous anti-spoofing works focused on supervised techniques, typically with either binary or auxiliary supervision. Most of them suffer from limited robustness and generalization, especially in the cross-dataset setting. In this paper, we propose a semi-supervised adversarial learning framework for spoof face detection, which largely relaxes the supervision condition. To capture the underlying structure of live faces data in latent representation space, we propose to train the live face data only, with a convolutional Encoder-Decoder network acting as a Generator. Meanwhile, we add a second convolutional network serving as a Discriminator. The generator and discriminator are trained by competing with each other while collaborating to understand the underlying concept in the normal class(live faces). Since the spoof face detection is video based (i.e., temporal information), we intuitively take the optical flow maps converted from consecutive video frames as input. Our approach is free of the spoof faces, thus being robust and general to different types of spoof, even unknown spoof. Extensive experiments on intra- and cross-dataset tests show that our semi-supervised method achieves better or comparable results to state-of-the-art supervised techniques.

© 2020

1. Introduction

Biometrics plays a key part in authentication and security applications. Access control using face, fingerprint or iris has been existed for quite a while in our daily life. Face recognition, one of the prevalent biometric applications, has achieved noticeable successes (Galbally et al., 2014). Face data has been a promising data type, due to its convenience, universality and acceptability for users. However, traditional face recognition systems can be easily fooled with common attacks like printed facial photographs. To obtain access to systems, criminals are already using some techniques to accurately simulate the biometric characteristics of valid users, such as faces. This process is known as face spoofing attack, which poses a great threat to face recognition systems (Patel et al., 2016b; Ratha et al., 2001). Presentation attacks (abbreviated as PA), including printed paper face, replaying a video and wearing a mask, are one of the most prevalent face spoofs. It has been demonstrated that traditional face recognition systems could be vulnerable to PA

(Chetty and Wagner, 2006; Frischholz and Dieckmann, 2000; Frischholz and Werner, 2003). Therefore, it is necessary to design robust countermeasure techniques to deal with the weakness in traditional face recognition application and prevent such frauds. As a result, various face anti-spoofing techniques have been proposed to detect spoof and live faces, before the face recognition stage. The main challenge of face anti-spoofing is how to achieve robustness and generalization to different kinds of PA.

Many previous strategies have been proposed to deal with the spoof face detection task. Spatial image information plays a critical role in face recognition system. Each facial region in our face includes different visual patterns and rich and discriminative information. These information could help to distinguish some faces from others. Therefore, some strategies are proposed to find different spoofing cues from different facial regions by using handcrafted features, such as LBP (de Freitas Pereira et al., 2012) and HOG (Yang et al., 2013). It is hard to obtain robust texture features, due to the cost of handcrafted features and a lack of an explicit correlation between pixel intensities and different types of attacks. With the recent development of deep learning, face spoofing detection based

**Corresponding author

e-mail: xuequan.lu@deakin.edu.au (Xuequan Lu)

on high-level learning features achieve more promising performance. However, these CNN-based methods (Li et al., 2016; Patel et al., 2016a) are adopted in spoof face detection with a softmax loss based on binary supervision. These supervised methods have the risk of overfitting on the training data and obtain low performance in the cross-dataset setting. In addition, temporal information is also a critical part in spoof face detection. For example, a liveness detection method (Bao et al., 2009) is proposed for spoofing face detection with using optical flow. It attempts to find the differences in motion patterns. That model attempts to learn the concept of optical flow generated by 3D objects and 2D planes. The motion of an optical flow field consists of four basic movements: translation, rotation, moving, and swing. Previous motion based methods (Jee et al., 2006; Sun et al., 2007; Kollreider et al., 2005) usually need to learn or obtain some explicit features using complicated modules such as modeling the motion. Based on these features which focus on representing specific characteristics, the trained model can make the real face images and spoof face images more separable. However, because of the specificity, these methods are hard to be generalized to other spoofing types.

Previous face anti-spoofing works focused on supervised methods, with the utilization of hand-crafted or learned features. Most approaches typically depend on binary or auxiliary supervision. Nevertheless, many previous works suffer from the following major limitations partially or wholly: (1) fully supervised setting—the utilization of both live and spoof face data (with labels), (2) the assumption of binary classification, and (3) the impracticality to take all types of spoof (maybe unknown spoof) into account. Furthermore, collecting spoof face data for training purpose is costly and time-consuming. Also, binary supervision could be insufficient to learn a good model and make desired predictions in cross dataset scenario. As a result, those face anti-spoofing techniques have limited robustness and generalization to various types of spoofing.

Motivated by the above limitations and analysis, we propose a novel adversarial network for anti-spoofing under the semi-supervised setting. We propose to train the live face data only, with a convolutional Encoder-Decoder network acting as a Generator. Besides, a second convolutional network is regarded as a Discriminator. The generator attempts to reconstruct the original input sample to fool the discriminator, while the discriminator tries to distinguish original images from generated images. In the process of training, both sub-networks compete with each other to achieve high-quality reconstructions for live faces data only.

While testing, the learned model has a lower reconstruction error of live face data than spoof face data. This is mainly because we train on live face data only, the model captures the real characteristic of live faces samples and the learned model can better describe the characteristics of live faces than those of spoof faces. We naturally take the optical flow maps converted from consecutive video frames as input, as the task of spoof face detection is video-based and involves temporal information. The semi-supervised setting significantly reduces the efforts in collecting spoof face data, thus making our method more robust and general to different types of face spoofing. As

such, the proposed approach is practical in the real world. We validate our method on challenging datasets. We also compare our semi-supervised method with state-of-the-art supervised anti-spoofing techniques, showing that our method produces better or comparable results to those approaches.

In summary, the main contributions of this paper are:

- a novel semi-supervised approach training on live face data only for spoof face detection.
- we propose a framework trained by generator and discriminator adversarially while collaborating to understand the real underlying concept in the normal class and classifying the testing samples by pixel-wise reconstruction error.
- we design a domain adaption algorithm which tries to learn some transfer components across domains in a Reproducing Kernel Hilbert Space (RKHS) using Maximum Mean Discrepancy (MMD).
- validation on challenging datasets, and extensive comparisons (intra- and cross-dataset testing) with current supervised anti-spoofing techniques.

The rest the paper is organized as follows. Section 2 reviews the relevant research. We elaborate our approach in Section 3. Section 4 gives various experimental results, and Section 5 concludes our work.

2. Related Work

The previous face anti-spoofing methods (Boulkenafet et al., 2017; de Freitas Pereira et al., 2012, 2013; Komulainen et al., 2013a; Määttä et al., 2011; Mirjalili and Ross, 2017; Patel et al., 2016b; Yang et al., 2013) can be generally divided into four categories: feature based methods, temporal information based methods, Hybrid methods as well as approaches based on other cues. Tab. 1 compares the characteristics of these previous spoof detection methods, including LBP (Määttä et al., 2011), DoG-SL (Peixoto et al., 2011), Color-texture (Boulkenafet et al., 2015), Optical flow field (Bao et al., 2009), Liveness optical flow (Smiatacz, 2012), Structure-tensor(Kollreider et al., 2005), Spatial-temporal domain (Sun et al., 2018), Patch-based CNN (Atoum et al., 2017), VGG (Li et al., 2016), Auxiliary (Liu et al., 2018) and De-Spoof (Jourabloo et al., 2018).

2.1. Feature-based Methods

Most early face anti-spoofing works used handcrafted features of texture information for binary classification (e.g., SVM). They expected that differing feature descriptors such as LBP (de Freitas Pereira et al., 2012, 2013; Määttä et al., 2011), HOG (Komulainen et al., 2013a; Yang et al., 2013), DoG-SL (Komulainen et al., 2013a; Yang et al., 2013), SIFT (Patel et al., 2016b) and SURF (Chingovska et al., 2012) could be computed for live and spoof faces. Nonetheless, many feature descriptors are largely affected by illumination, imagery and other factors. Such feature-based methods often have poor generalization in cross-dataset testing (Liu et al., 2018).

Table 1. Characteristics of different face spoof detection methods.

Methods	Analysis type	Strategy	Datasets	Algorithm type
LBP	Texture analysis	Micro-texture analysis via LBP with SVM as a classifier	NUAA Photograph Imposter Database	Supervised
DoG-SL	Texture analysis	Applying an adaptive histogram equalisation to the images	Yale Face Database and NUAA Photograph Imposter Database	Supervised
Color-texture	Texture analysis	Computing a half of Face with another half that is divided in two ways: horizontally and vertically	CASIA Face Anti-Spoofing and the Replay-Attack databases	Supervised
Optical flow field	Motion analysis	Analyzing the optical flow field to detect real face	-	Supervised
Structure tensor	Motion analysis	Face motion estimation based on the structure tensor and a few frames	XM2VTS database	Supervised
Spatial-temporal Domain	Motion analysis + Texture analysis	A two-stream structure (spatial, temporal)	Replay-Attack, CASIA and 3DMAD	Supervised
Patch-based CNN	Texture analysis +cue analysis	Extracting the local features and holistic depth maps from the face images	CASIA-FASD, MSU-USSA, and Replay-Attack	Supervised
VGG	Texture analysis	Extracting the deep partial features from the convolutional neural network (CNN)	Replay-Attack and CASIA	Supervised
Liveness optical flow	Motion analysis	Applying the Support Vector Machine to distinguish between the motion information of real faces and photographs	Regensburg university dataset	Supervised
Auxiliary	Texture analysis + cues analysis	Fusing the estimated depth and rPPG to distinguish live v.s. spoof faces	CASIA-MFSD and Replay-Attack	Supervised
De-Spoof	Texture analysis	A CNN architecture with proper constraints and supervisions	Oulu-NPU, CASIA-MFSD and Replay-Attack	Supervised
Our method	Motion analysis	A semi-supervised adversarial learning framework	Nuua, CASIA-MFSD and Replay-Attack	Semi-Supervised

CNN is good at extracting and learning deep features. (Yang et al., 2014) treated CNN as a classifier for face anti-spoofing, and used different spatial scales of live and spoof face images for training. Xu et al. (2015) proposed a LSTM-CNN architecture to predict the frames of videos. Most previous CNN techniques for face anti-spoofing utilized a binary classification to predict live or spoof faces (Feng et al., 2016; Li et al., 2016; Patel et al., 2016a; Yang et al., 2014). However, both live and spoof face data have to be considered in the training procedure. In worse cases, the test face data does not involve cues like printed page edges or digital replay devices while the trained model might use such cues to detect spoof faces. As a result, the classification ability for live and spoof faces is limited. Also, it is difficult to explain the final results.

2.2. Methods Based on Temporal Information

As with other video-based tasks like activity recognition, temporal information is also useful in face anti-spoofing. Some researchers paid attention to the movement of key parts in a face, for example, eye-blinking and lip movements. The temporal information based methods are usually vulnerable to the replay attack (i.e., replaying video with a digital device). Gan et al. (2017) proposed a 3D convolutional network to classify live and spoof faces, by supervisedly learning temporal features with a stacked structure. Unfortunately, it relies on a large amount of data and could perform poorly on small datasets. Xu et al. (2015) introduced a new structure by combining LSTM units with CNN for binary classification. Feng et al. (2016) presented a CNN by taking both optical flow features and shearlet features as input. These methods took advantage of temporal information to distinguish between live and spoof faces.

2.3. Hybrid Methods

Hybrid techniques combining features and temporal information have also been proposed for spoof face detection. Schwartz et al. (2011) used multiple low-level features to create one high dimensional vector with the size of more than one million. They further adopted the partial least squares approach on this vector to distinct between live and spoof faces. Komulainen et al. (2013b) introduced the combination of computationally inexpensive linear classifiers for robust face anti-spoofing. They used the fusion of motion information and features. Both methods depend on the multi-block local binary pattern and motion estimation from input videos.

2.4. Methods Based on Other Cues

There have been considerable amount of works using other cues derived from the original video frames (Komulainen et al., 2013a; George et al., 2019). For example, rPPG signal, IR image (Zhang et al., 2011), depth image (Wang et al., 2013) and voice (Chetty, 2010) are some common cues. Nevertheless, such cues have their own limitations. Taking rPPG-based methods as an instance, researchers often need to extract the rPPG signals from a long video, to achieve decent predictions (Liu et al., 2018). As a matter of fact, it is unfeasible for a face anti-spoofing system to detect spoof faces through analyzing a long video (e.g., 50 seconds).

2.5. Anomaly detection

Anomaly detection is a classical problem in computer vision. When samples are deviating from the expected behavior defined by “normal” samples of a training dataset, these samples are classified as the abnormal class.

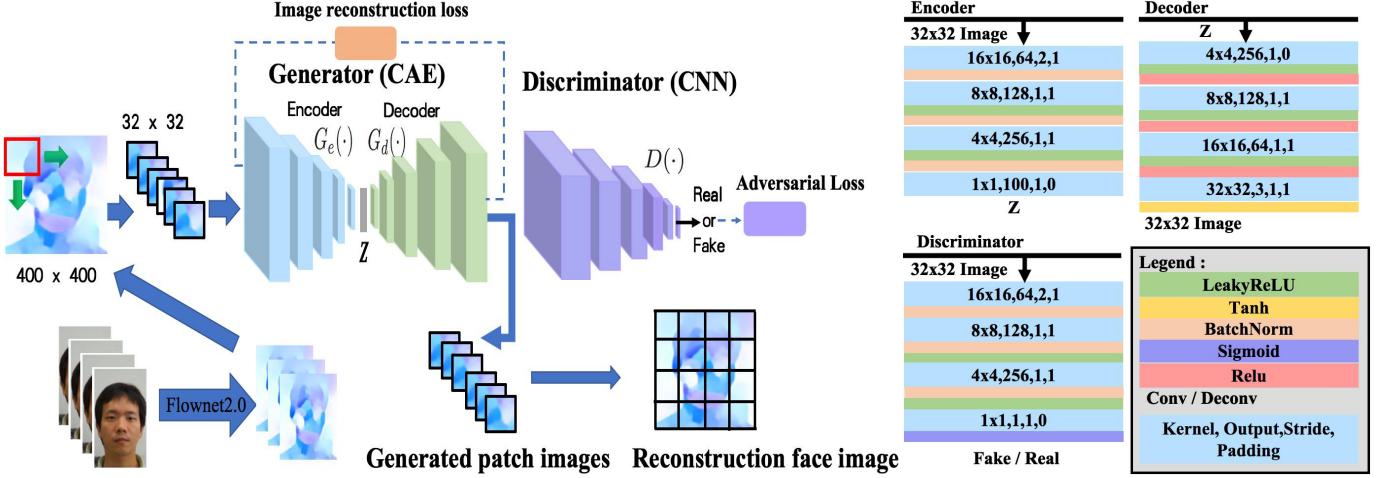


Fig. 1. Our framework consists of a generator and a discriminator. The generator and discriminator are trained by competing and collaborating with each other to understand the underlying structure in the live faces data. The architecture layers of each component are described on the right.

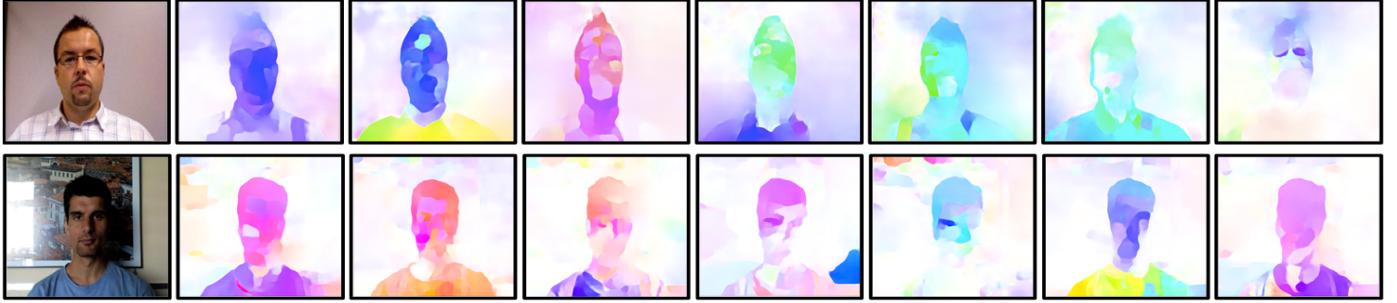


Fig. 2. Two live faces with optical flow data visualization in each row. The first image is one of the frames in each live face video, followed by seven optical flow maps which are generated from its follow-up frames.

Recently, deep learning based autoencoders are used to learn the pattern of normal behaviors and exploit the reconstruction loss to detect anomalies. For example, Baur et al. (2018) tackles the problem by learning a mapping to a lower dimensional representation, where the real distribution is modeled. The decoder upscales the latent feature vector to reconstruct the image. In recent research, a lot of abnormal detection methods (Zenati et al., 2018; Xia et al., 2019) based on the Generative Adversarial Networks (GANs) are proposed. For instance, Xia et al. (2019) proposed latent spatial features based on generative adversarial networks for face anti-spoofing with an additional feature classifier. The input of this framework extracts the appearance information from the original face image with different sizes. In our work, instead we use the motion information from the original face images. According to the ablation study (Section 4.2.2), the performance of using motion information is better than using appearance information. Moreover, each size of the input corresponds to one GAN model, which induces significantly higher costs. In addition, their framework only reports a high performance in the intra-dataset setting. It disregards the generalization issue by excluding the cross-dataset setting, which is critical for spoofing detection.

3. Proposed Approach

In this section, we present how to learn the intrinsic structure of live faces by using the proposed adversarial training framework. We start by describing the details of the overview network architecture, then depict each term in loss function, and finally give the description of the testing method.

3.1. Network Architecture

Our method consists of a data preprocessing step and a GAN-style architecture. The preprocessing step is to convert consecutive video frames into optical flow maps. Fig. 2,3,4 shows the visualization of optical flow map. The GAN-style architecture, inspired by the anomaly detection (Sabokrou et al., 2018), comprises of two components: the generation network and the discrimination network. Fig. 1 shows the overview of our framework.

Due to the outstanding performance of CNN (Krizhevsky et al., 2012; Lawrence et al., 1997; Kalchbrenner et al., 2014), we take a convolutional autoencoder as the Generator. The main idea is that we only consider the live face data for training. The learned model is therefore not good at depicting the characteristics of spoof face data, leading to high reconstruction errors. The reason why we employ Convolutional AutoEncoder (CAE) in the proposed framework can be concluded as

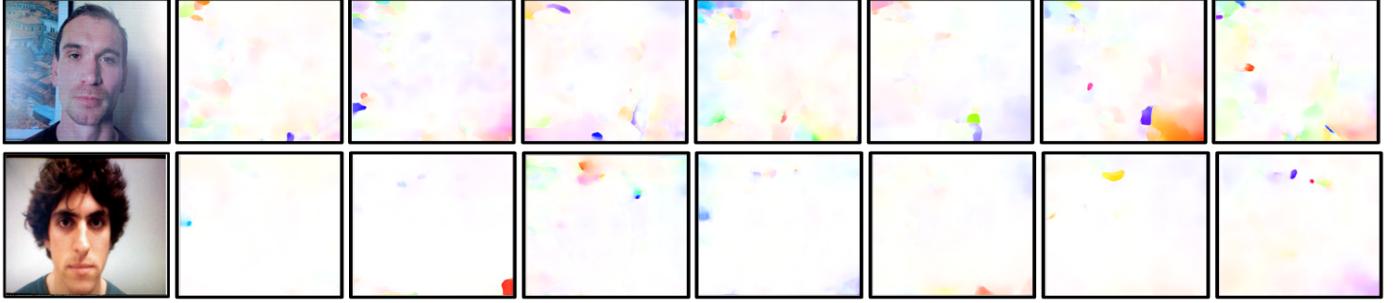


Fig. 3. Two fixed spoofing faces with optical flow data visualization in each row. The first image is one of the frames in each spoofing face video by holding the client biometry, followed by seven optical flow maps which are generated from the follow-up frames.



Fig. 4. Two hand spoofing faces with optical flow data visualization in each row. The first image is one of the frames in each spoofing face video from the device held by the attacker’s hands, followed by seven optical flow maps which are generated from the follow-up frames.

follows: (1) Conventional Autoencoders (AEs) often ignore the structure of 2D images, and interpret the input as a single latent vector. (2) The network is constrained by the number of input images. The redundant parameters in AEs force each feature to be global by spanning the entire visual field. (3) The Convolutional AutoEncoder (CAE) can learn the optimal filters to minimize the reconstruction error. In fact, Convolutional Neural Networks are usually referred to supervised learning algorithms. CAE, instead, is trained only to learn filters to extract features that can be used to reconstruct the input.

To prevent being fooled by the generator, the discriminator learns the core characteristics in the original data during the period of training. The discriminator also assists the generator to get robust and stable parameters in the process of training. This part of parameters would increase the reconstruction gap between live faces and fake faces in the process of testing.

3.2. Overall Loss Function

To train our model, we define a loss function in Eq. (1) including two components, the adversarial loss and the pixel-wise image reconstruction loss.

$$\mathcal{L} = w_i \mathcal{L}_{irec} + w_a \mathcal{L}_{adv}, \quad (1)$$

where w_i and w_a are the weighting parameters balancing the impact of individual item to the overall object function.

Adversarial loss. The Generative Adversarial Network (GAN) (Creswell et al., 2018) originates from a game between two players. One player is called the generator $G(x)$. The generator creates samples that are intended to come from the same distribution as the training data. The other player is called the

discriminator $D(x)$. The discriminator would make a decision whether the samples are generated by the generator or taken from the training data. The generator attends to fool the discriminator by reconstructing fake samples similar to the true training data. This adversarial game between the generator and discriminator can be formulated as:

$$\begin{aligned} \mathcal{L}_{adv} = \min_G \max_D & \left(E_{x \sim p_x} [\log(D(\mathbf{x}))] \right. \\ & \left. + E_{\mathbf{x} \sim p_x} [\log(1 - D(G(\mathbf{x})))] \right). \end{aligned} \quad (2)$$

Image reconstruction loss: While the discriminator tries to differentiate between realistic images and generated images, and the generator trying to fool the discriminator. However, the generator is not optimized towards learning the real concept from input data only by adversarial loss. Some prior works have proposed that the distance between input images and generated images should be considered. Isola et al. (2017) shows that the use of L1 yields less blurry results than L2. Therefore, we use L1 loss function to penalize the generator by minimizing the distance between original input \mathbf{x} and generated images $G(\mathbf{x})$ as follows.

$$\mathcal{L}_{irec} = \mathbb{E}_{x \sim p_x} \|x - G(x)\|_1 \quad (3)$$

3.3. Data Preprocessing

We extract frames from each video with 30 frames per second. FlowNet2.0 (Ilg et al., 2017) is then employed to estimate the optical flow between frames, due to its effectiveness. Optical flow is the pattern of apparent motion, which is calculated based on two adjacent images. It defines both horizontal

and vertical displacements for each pixel, and reflects motion about objects and scene. The pre-trained FlowNet model estimates the optical flow between each pair of two adjacent frames and outputs the optical flow files. The horizontal and vertical components are included in optical flow files. The color-coding scheme (López, 2017) allows us to visualize the horizontal and vertical displacements in one image, as illustrated in Fig. 5. Colors can be assigned to each pixel. We utilize the color coding scheme to convert these optical flow files into images where the displacement vector is color.

The output flow maps are also RGB images with colors indicating the flow signal. The patches are generated from each flow map by a sliding window. The size of this window is set to 32×32 . Fig. 2, 3 and 4 visualize optical flows of live faces, spoofing faces by holding the client biometry (i.e., fixed spoofing) and spoofing faces from the device held by the attacker’s hands (i.e., hand spoofing), respectively. It shows that the optical flows of live faces are more clear than the spoof faces. Hand spoofing leads to considerable amount of noise on the flow maps. This is because that the movement of spoofing faces and digital device screens is consistent. For fixed spoofing faces, printed faces are fixed in front of the detection systems. This type involves few noise and nearly no optical flows.

3.4. Testing method

To demonstrate the effectiveness of the proposed framework, we conduct two intra-testing experiments and one cross-testing experiment. For intra-testing experiments, the model is trained in the training dataset accordingly, as with the state-of-the-art methods (Yu and Jia, 2017). The testing dataset in the same domain is used to evaluate the performance of each method. Different from intra-testing experiments, cross-database experiments with different domains are more challenging. Domain adaptation (Finkel and Manning, 2009) is a field associated with machine learning and transfer learning. The aim of the domain adaptation problem is to train a well performing model from the source data distribution. The trained model could still perform well on a different (but related) target data distribution. As such, we attempt to extend the domain adaptation in our study. *To our knowledge*, we are the first to investigate the domain adaptation issue in the face anti-spoofing area.

In the cross-database situation, the labels of all target samples are unknown during training. Compared with the intra-database

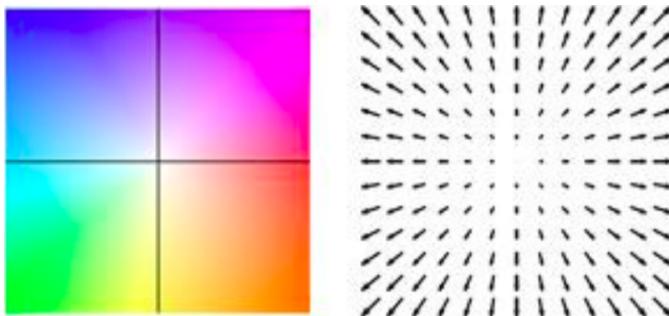


Fig. 5. Visualization of horizontal and vertical displacements in one RGB image.

setting, it is more ubiquitous in real-world applications. Due to the unavailability of labels in the target domain, one commonly used strategy is to learn domain-invariant representations via minimizing the domain distribution discrepancy. In our cross-database scenario, the model is trained on dataset A and tested on dataset B. There exists some difference between the source domain and target domain, for example, image quality, reflection and environment. One intuitive solution is to consider mapping the reconstruction data to a high (possibly infinite) dimensional space and computing the sample means in this space using high-order statistics (up to infinity). As a result, we could achieve a better discrimination threshold for live and spoofing faces. By contrast, directly training a classifier on the source data and using the threshold set in the source data often leads to certain “overfitting” to the source distribution and reduced performance while testing on the target domain.

We consider a source domain $\mathcal{D}_s = \{\mathbf{x}_i^s, y_i^s\}_{i=1,\dots,n_s}$ and a target domain $\mathcal{D}_t = \{\mathbf{x}_i^t, y_i^t\}_{i=1,\dots,n_t}$. Here, $\mathbf{x}_i^s \in \mathbb{R}^{N_s}$, $\mathbf{x}_i^t \in \mathbb{R}^{N_t}$ are the reconstruction errors for each frame in the source domain and the target domain, respectively. $y_i^s \in C$, $y_i^t \in C$ are corresponding labels, where the target labels $\{y_i^t\}_{i=1,\dots,n_t}$ are not available for training. For domain adaption, we assume that the source and target domains are associated with the same label space, while \mathcal{D}_s and \mathcal{D}_t are drawn from distributions \mathbb{P}_s and \mathbb{P}_t which are assumed to be different. That is, the source and target distribution have different joint distributions of data X and labels Y : $\mathbb{P}_s(X, Y) \neq \mathbb{P}_t(X, Y)$.

Maximum Mean Discrepancy (MMD) Yan et al. (2017) is an effective non-parametric metric for comparing the distance between two distributions. Given two distributions s and t , by mapping the data to a reproduced kernel Hilbert space (RKHS) using function $\phi(\cdot)$, the MMD between s and t is defined as,

$$\text{MMD}(s, t) = \sup_{\|\phi\|_{\mathcal{H}} \leq 1} \|E_{\mathbf{x}^s \sim s}[\phi(\mathbf{x}^s)] - E_{\mathbf{x}^t \sim t}[\phi(\mathbf{x}^t)]\|_{\mathcal{H}}, \quad (4)$$

where $E_{\mathbf{x}^s \sim s}[\cdot]$ denotes the expectation with regard to the distribution s , and $\|\phi\|_{\mathcal{H}} \leq 1$ defines a set of functions in the unit ball of a RKHS. Based on the statistical tests defined by MMD, we have $\text{MMD}(s, t) = 0 \iff s = t$. Denote by $\mathcal{D}_s = \{\mathbf{x}_i^s\}_{i=1}^M$ and $\mathcal{D}_t = \{\mathbf{x}_i^t\}_{i=1}^N$, two sets of samples drawn i.i.d. from the distributions s and t respectively, the empirical estimation of MMD can be given by:

$$\text{MMD}(\mathcal{D}_s, \mathcal{D}_t) = \left\| \frac{1}{M} \sum_{i=1}^M \phi(\mathbf{x}_i^s) - \frac{1}{N} \sum_{j=1}^N \phi(\mathbf{x}_j^t) \right\|_{\mathcal{H}}, \quad (5)$$

where $\phi(\cdot)$ denotes the feature map associated with the kernel map $k(\mathbf{x}^s, \mathbf{x}^t) = \langle \phi(\mathbf{x}^s), \phi(\mathbf{x}^t) \rangle$, which is usually defined as the convex combination of several basis kernels.

With the help of MMD, the statistical test method works in the following way. Based on the samples of two distributions, one distribution is the reference distribution formed by training live face samples, and another distribution is obtained in the same way from test samples. By finding the continuous function ϕ in the sample space, the mean value of the samples from different distributions on function ϕ is obtained. Dividing the

two mean values yields an average difference between the two distributions. Finally, MMD is taken as the measurement to determine the category of the test videos. If the value of MMD is smaller than the predefined threshold T, the test samples distribution is considered to be the close to the live face reference distribution; otherwise they are spoof videos. The final testing scheme is summarized in Algorithm 1.

Algorithm 1 Spoofing Detection in video V

```

input: A video  $V = [F_1, F_2, \dots, F_N]$ , trained models: Generator, Auxiliary encoder and decision threshold T
output: report if (ScoreVideo > T) 'spoof' else 'non-spoof'
1: function DISTRIBUTION( $V$ )
2:   FrameArray = []
3:   for  $k \leftarrow 1$  to  $N$  do
4:     ScoreFrame =  $\|G_e(F_k) - G_{e'}(F'_k)\|_1$ 
5:     FrameArray  $\leftarrow$  FrameArray + ScoreFrame
6:   end for
7:   return FrameArray
8: end function
9: ReferDis = Distribution(TrainPosVideo)
10: TestDis = Distribution(TestVideo)
11: ScoreVideo = MMD(ReferDis, TestDis)

```

4. Experimental Results

In this section, we firstly present the experimental setting, including datasets, and more implementation details. Then, for ablation study, two experiments are conducted to analyze the proposed method in detail. Finally, We evaluate proposed method and the state-of-the-art techniques on both intra/inter-dataset settings.

4.1. Experimental Setup and Datasets

The proposed method is mainly implemented in the Tensorflow framework (Abadi et al., 2015). The experiments are carried out on a PC with a NVIDIA-1080 graphics card and a multi-core 2.1 GHz CPU. A good face anti-spoofing system must be robust to different types of attacks. We evaluate our method and the state-of-the-art techniques on three publicly available face spoofing detection databases: (i) NUAA Imposter Database (Tan et al., 2010a), (ii) Replay-Attack (Chin-govska et al., 2012) dataset, and (iii) CASIA MFSD (Zhang et al., 2012) dataset. These structures are kept fixed for all databases, and learning rate is set to 0.02.



Fig. 6. Some samples from the NUAA dataset. The first row and second row show five live samples and five spoofing samples, respectively.

According to the work (Akçay et al., 2018), CIFAR10 and MNIST datasets are used to construct the experiment to illustrate the superiority of our approach over the state-of-the-art

one-class classifiers. One of the classes is regarded as normal class, while the rest ones belong to the abnormal class. In particular, we respectively get ten sets for MNIST and CIFAR10, and then detect the outlier anomalies by only training the model on the normal class data in ablation study.

The NUAA dataset is widely used for the evaluation of face liveness detection. This dataset consists of 15 different subjects captured in different places and illumination conditions, involving 12,614 real and photographed face images. Each subject was asked to look at the webcam frontally with a neutral expression and without noticeable movements such as eyeblink or head movement. For training data, it contains 1,743 real faces and 1,748 photographed faces. For testing, it includes 3,362 real faces and 5,761 photographed faces. Fig. 6 shows some samples from the NUAA dataset.



Fig. 7. Some samples from the Replay-Attack dataset. It includes some live faces, some spoofing faces by holding the client biometry (i.e., fixed spoofing) and spoofing faces from the device held by the attacker's hands (i.e., hand spoofing).

The Replay-Attack dataset is also a widely used and publicly available database. It has 360 videos (60 real faces videos and 300 spoof faces) as the training data. About validation data, it has the same number of videos as training videos. The validation data will be fully used, to calibrate the threshold to distinguish between real and spoof faces (explained in Section 3.4). The resolution of the Replay-Attack data is 320×240 . The dataset considers different lighting conditions used in spoofing attacks. It consists of 80 videos of real faces and 400 videos of fake faces as the testing data. The fake faces are obtained by using the attackers' bare hands or fixed support. Fig. 7 shows some samples from the Replay-Attack dataset.

The CASIA (Zhang et al., 2012) dataset involves 50 subjects, and each subject has 12 videos (3 real faces and 9 fake faces). The dataset is divided into the training set (20 subjects, 240 videos) and the test set (30 subjects, 360 videos). Compared with the Replay-Attack dataset, there is no validation data in this dataset. The CASIA dataset is more difficult in spoof face detection, in terms of image quality, resolution and video length. It consists of print and replay attacks using corresponding photos and replayed videos. Some of the print attack photos are manually cropped around the eyes to deter eye-blinking based techniques.

There exist a lot of face anti-spoofing approaches, and many CNN-based supervised works have achieved promising results in the intra-database setting. However, the high intra-database prediction accuracy does not guarantee a decent performance in the inter-database setting which is more common in real world. In fact, the cross-database performance better reflects the actual

Table 2. Abnormal detection results for MNIST/CIFAR10 datasets using Protocol 2. (Plane and Car classes are annotated as Airplane and Automobile in CIFAR10).

	0	1	2	3	4	5	6	7	8	9	MEAN
OCSVM ('01)	0.988	0.999	0.902	0.950	0.955	0.968	0.978	0.965	0.853	0.955	0.9513
KDE ('06)	0.885	0.996	0.710	0.693	0.844	0.776	0.861	0.884	0.669	0.825	0.8143
DAE ('06)	0.894	0.999	0.792	0.851	0.888	0.819	0.944	0.922	0.740	0.917	0.8766
VAE ('13)	0.997	0.999	0.936	0.959	0.973	0.964	0.993	0.976	0.923	0.976	0.9696
Pix CNN ('16)	0.531	0.995	0.476	0.517	0.739	0.542	0.592	0.789	0.340	0.662	0.6183
AND ('19)	0.984	0.995	0.947	0.952	0.960	0.971	0.991	0.970	0.922	0.979	0.9671
DSVDD ('18)	0.980	0.997	0.917	0.919	0.949	0.885	0.983	0.946	0.939	0.965	0.9480
Autoencoder	0.992	1.0	0.876	0.937	0.949	0.968	0.984	0.959	0.843	0.959	0.9467
Proposed method	0.996	0.999	0.987	0.986	0.977	0.991	0.998	0.987	0.986	0.987	0.9898
	PLANE	CAR	BIRD	CAT	DEER	DOG	FROG	HORSE	SHIP	TRUCK	MEAN
OCSVM ('01)	0.630	0.440	0.649	0.487	0.735	0.500	0.725	0.533	0.649	0.508	0.5856
KDE ('06)	0.658	0.520	0.657	0.497	0.727	0.496	0.758	0.564	0.680	0.540	0.6097
DAE ('06)	0.411	0.478	0.616	0.562	0.728	0.513	0.688	0.497	0.487	0.378	0.5358
VAE ('13)	0.700	0.386	0.679	0.535	0.748	0.523	0.687	0.493	0.696	0.386	0.5833
Pix CNN ('16)	0.788	0.428	0.617	0.574	0.511	0.571	0.422	0.454	0.715	0.426	0.5506
AND ('19)	0.717	0.494	0.662	0.527	0.736	0.504	0.726	0.560	0.680	0.566	0.6172
DSVDD ('18)	0.617	0.659	0.508	0.591	0.609	0.657	0.677	0.673	0.759	0.731	0.6481
Autoencoder	0.735	0.585	0.752	0.703	0.375	0.687	0.594	0.397	0.781	0.500	0.6109
Proposed method	0.996	0.648	0.752	0.770	0.934	0.695	0.958	0.623	0.976	0.587	0.7930

capability of a system in real-world applications. Therefore, a good cross-database performance provides strong evidence that: i) features are generally invariant to different scenarios (i.e., camera and illuminations), ii) a spoof classifier trained in one scenario is generalizable to other scenarios, and iii) data captured in one scenario can be useful for developing effective spoof detectors in other scenarios. As such, to demonstrate the effectiveness of the proposed framework, we conduct two intra-database experiments and two cross-database experiments.

4.2. Ablation study

4.2.1. With/without the discriminator

To show the effectiveness of adversarial learning, it is necessary to conduct the experiment with or without the discriminator part. In the first scenario, we only use the common convolutional autoencoder with a simple image reconstruction error. During the inference, the reconstruction error of the test sample is regarded as the abnormality score. In the second scenario, all components are used in the proposed framework with the discriminator. Besides the image reconstruction error, the adversarial learning loss is also considered. The generator tries to generate a high quality image to fool the discriminator. The discriminator attempts to distinguish the generated image from a realistic image. During the training process, the discriminator helps the generator to capture the underlying concept of normal samples. The test sample is detected in the same way as the first scenario (i.e., image reconstruction error).

Since we formulate the spoof face detection task as abnormal detection task, it is necessary to explore the effectiveness of the proposed method in both tasks. In the abnormal de-

tection task, we conduct an experiment to demonstrate the superiority of our method over state-of-the-art one-class classifiers on MNIST and CIFAR10 datasets, shown in Tab. 2. For both MNIST and CIFAR10, we select one class as the normal class at each time, while leaving the rest to be the abnormal classes, leading to ten sets for abnormal detection. Normal data and abnormal data are to imitate live faces and spoof faces, respectively. Our method typically achieves improvements compared with other methods, including OCSVM (Schölkopf et al., 2001), KDE (Bishop, 2006), DAE (Hadsell et al., 2006), VAE (Kingma and Welling, 2013), Pix CNN (Kalchbrenner et al., 2016), AND (Abati et al., 2019) and DSVDD (Ruff et al., 2018). In addition, it is clear that the proposed method with the discriminator achieves higher performance than the autoencoder without the support from discriminator in both datasets. In our spoofing face detection task, the cross-dataset experiment is also conducted by using the two scenarios (with and without the discriminator) described above. As shown in Tab. 3, the discriminator helps the generator (Autoencoder) to capture the concept of live faces. The trained generator is used to detect the spoof faces directly, and we obtain a better performance with the discriminator than the autoencoder strategy without the discriminator.

4.2.2. Impact on performance with optical flow

Table 4. Classification performance of the proposed approach in different types (motion or appearance) of input, in terms of HTER (%). The algorithm is trained using the CASIA-MFSD dataset and tested on the Replay-Attack dataset, and vice versa.

Methods	Train	Test	Train	Test	Average
	CASIA MFSD	Replay Attack	Replay Attack	CASIA MFSD	
Appearance information			30.8%		49.7%
Motion information			15.6%		44.1%

Table 3. Classification performance of autoencoder (without discriminator) and the proposed method (with discriminator), in terms of HTER (%). They are trained using the CASIA-MFSD dataset and tested on the Replay-Attack dataset, and vice versa.

Methods	Train	Test	Train	Test	Average
	CASIA MFSD	Replay Attack	Replay Attack	CASIA MFSD	
Autoencoder	25.6%		47.3%		36.5%
Proposed method	15.6%		44.1%		29.8%

To explore the influence of optical flow information, we consider to use appearance information and motion information in the experiments, respectively. In the appearance information situation, we only use the original frame from video, which is taken as the input of the proposed method. In the motion information situation, only the optical information is obtained from

Table 5. Performance comparison using AUC on the NUAA dataset

Methods	Accuracy
Ours(semi-supervised)	99.3%
ADKMM ('17)	99.3%
ND-CNN ('17)	99.3%
DS-LSP ('15)	98.5%
CDD ('13)	97.7%
DoG-SL ('11)	94.5%
M-LBP ('11)	92.7%
DoG-LRBLR ('10)	87.5%
DoG-F ('04)	84.5%
DoG-M ('12)	81.8%

original videos. A performance comparison between these two cases is presented in Tab. 4. Compared with appearance information, the motion information could better assist in distinguishing the spoofing faces from live faces.

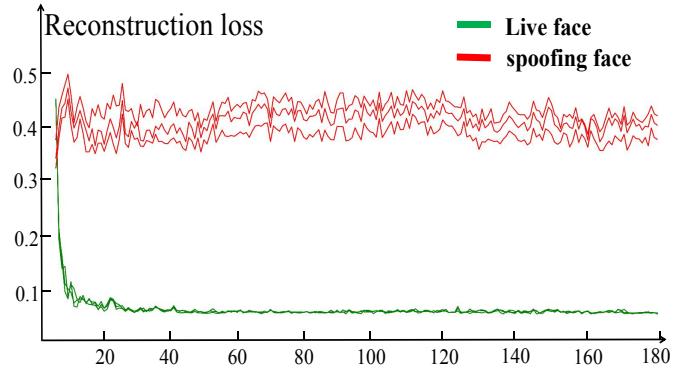
4.3. Intra NUAA Database Experiment

We evaluate the performance of the proposed method and state-of-the-art techniques on the NUAA dataset, in an intra-database sense. The competitors include DoG and high frequency based (DoG-F) (Li et al., 2004), multiple difference of Gaussian (DoG-M) (Zhang et al., 2012), DoG-sparse logistic (DoG-SL) (Peixoto et al., 2011), diffused speed-local speed pattern (DS-LSP) (Kim et al., 2015), multiple local binary pattern (M-LBP) (Määttä et al., 2011), DoG-sparse low-rank bilinear logistic regression (DoG-LRBLR) (Tan et al., 2010b), DoG-sparse logistic (DoG-SL) (Peixoto et al., 2011), component-dependent descriptor (CDD) (Yang et al., 2013), ADKMM (Yu and Jia, 2017) and the nonlinear diffusion based convolution neural network (ND-CNN) (Alotaibi and Mahmood, 2017).

To evaluate the reconstruction performance for each epoch, we choose three face spoofing samples and three live samples from the train set randomly. Once the network are trained in each epoch, the trained model would output the reconstructed images of these live or spoofing samples. Fig. 8 reveals that the gap between the reconstruction losses of spoofing faces and those of live faces are increased until they become stable after a few epochs. This indicates that the proposed approach can quickly distinguish spoof faces from live samples, without requiring spoof face data for training. Tab. 5 shows the accuracies for all methods. Our semi-supervised approach achieves the best performance, which is the same as the supervised AD-KMM (Yu and Jia, 2017) and supervised ND-CNN (Alotaibi and Mahmood, 2017).

4.4. Intra Replay-Attack Dataset Experiment

We also compared our method with state-of-the-art techniques on the Replay-Attack dataset, in the intra-database setting. Competitors includes $LBP_{3 \times 3}^{u2} + x^2$ (Chingovska et al., 2012), $LBP_{3 \times 3}^{u2} + LDA$ (Chingovska et al., 2012), $LBP_{3 \times 3}^{u2} + SVM$ (Chingovska et al., 2012), LBP + SVM (Määttä et al., 2011), DS-LBP (Kim et al., 2015), ND-CNN (Alotaibi and Mahmood, 2017), VGG (Li et al., 2016), Color-texture (Boulkenafet et al., 2015), Fisher-vector-encoding (Boulkenafet et al., 2016), Depth-based-CNNs (Patch-based CNN , Depth -based CNN and Patch and depth CNN) (Atoum et al., 2017), D-K (Yu and Jia, 2017), DTCNN (Tu et al., 2019a), Hand-crafted + CNN (Rehman et al., 2020) and Generalized deep feature (Li et al.,

**Fig. 8. Reconstruction performance with different numbers of iterations (epoch) on the NUAA dataset.****Table 6. Performance comparison using HTER measure on the Replay-Attack dataset (intra-database setting).**

Methods	test
$LBP_{3 \times 3}^{u2} + x^2$ ('12)	34.0%
$LBP_{3 \times 3}^{u2} + LDA$ ('12)	17.2%
$LBP_{3 \times 3}^{u2} + SVM$ ('12)	15.16%
LBP + SVM ('11)	13.9%
DS-LBP ('15)	12.5%
Color-texture ('15)	2.9%
VGG ('16)	4.3%
D-K ('16)	4.3%
Fisher-vector-encoding ('16)	2.0%
ND-CNN ('17)	10.0%
Patch-based CNN ('17)	1.2%
Depth -based CNN ('17)	0.7%
Patch and depth CNN ('17)	0.7%
Generalized deep feature ('18)	1.2%
DTCNN ('19)	20.0%
Hand-crafted + CNN ('20)	2.3%
Ours	12.3%
Ours with motion judgment	3.5%

2018a). Previous spoofing face detectors have achieved outstanding performance in the intra-dataset setting by supervised learning with both positive and negative labels. To certain degree, these supervised methods have the risk of overfitting on the training data and obtain poor generalization in cross-dataset setting. In addition, in the real world, it is impossible for us to collect and cover all kinds of spoof faces. Some types of spoofing faces are even unknown. Based on these challenges, we formulate the spoofing faces detection task as an abnormal detection task by only training the normal samples (live faces), which obtain a comparable performance in intra dataset setting with strong generalization.

The proposed method obtains the best performance with only 40 epochs. Fig. 10 shows the reconstruction loss for live and spoof faces. Besides the training data and testing data, this dataset also provides the development data to evaluate the performance. we calculate the half total error rate (*HTER*) (Bengio and Mariéthoz, 2004) to measure the performance. The *HTER* is half of the sum of the false rejection rate (*FRR*) and false acceptance rate (*FAR*). The half total error rate (*HTER*) would be also used in the metric of cross-database experiments.

$$HTER = \frac{FRR + FAR}{2} \quad (6)$$

The way to perform the attacks can be divided into two sub-

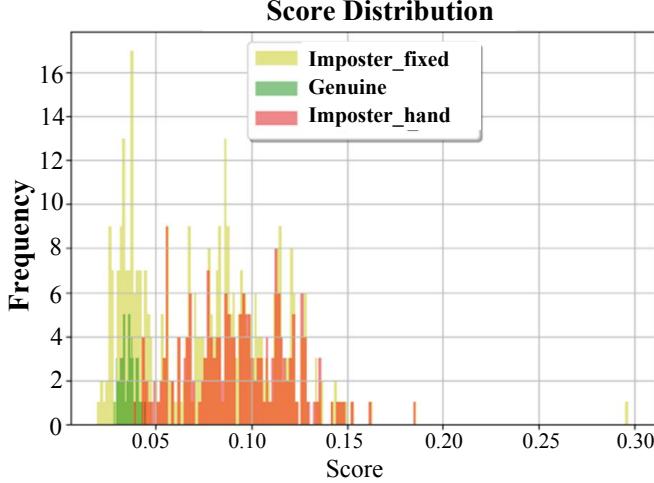


Fig. 9. Three distributions of reconstruction error scores. It consists of live faces, spoofing faces by holding the client biometry (i.e., fixed spoofing) and spoofing faces from the device held by the attacker’s hands (i.e., hand spoofing).

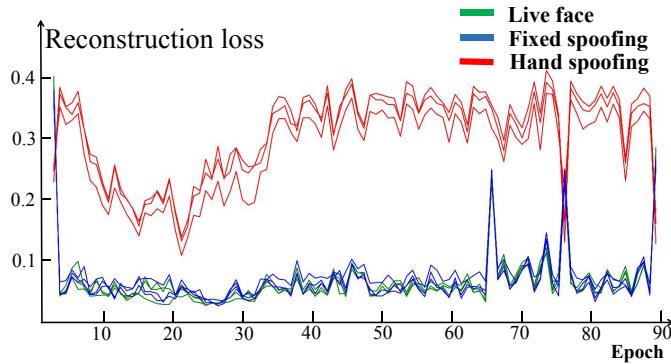


Fig. 10. Three kinds of reconstruction performance with different iteration numbers (epoch) on the intra Replay-Attack database experiment.

sets: the first subset is composed of videos generated using a stand to hold the client biometry (“fixed”). For the second subset, the attackers hold the devices with their own hands. We choose three samples from each set (fixed spoofing, hand spoofing and live validation). Fig. 10 illuminates that the reconstruction loss of live faces decreases sharply at the beginning of training. It tends to be stable with minor changes after that, until the 66-th epoch where both the reconstruction losses of live and spoof faces start to oscillate.

It is noteworthy that the changes in live reconstruction error and fixed spoof faces reconstruction error are consistent with increasing iterations. From Fig. 3, there is no temporal information in the way of using a stand to hold the client biometry (“fixed”) except some noise in the optical flow maps. Thus, the performance of reconstruction loss for fixed spoofing faces would be great and even better than live faces. Fig. 9 also further verifies what we have found in Fig. 3.

A performance comparison with previous methods is shown in Tab. 6. On the test set of the Replay-Attack dataset, the *HTER* of our method is 0.123, and we achieve a comparable *HTER* to other methods (worse than the best). This re-

sult attributes to two factors. Firstly, the binary classification methods, using both positive and negative data with labels, often achieve excellent performance in the intra-database setting (i.e., train and test within the same dataset). Some of the compared methods even use depth information or other extra cues for spoof faces detection. Secondly, as we explained before, most fixed spoof faces are mistakenly identified as live faces based on low reconstruction loss due to no motion cues (e.g., Fig. 3).

To tackle this issue, we design a new module which is responsible for detecting the presence of motion information. This is achieved through calculating the average pixel difference between pairs of optical flow maps of random frame samples. If there is no noticeable difference between each pair of optical flow maps, it means no motion information. With the support of this motion detection module, we could obtain more information for spoof faces before the real face spoofing detection. The *HTER* of our method can be declined from 0.123 to 0.035, with the aid of such motion judgment.

4.5. Cross-database Experiments

The cross-database performance is evaluated by training the proposed method on the CASIA-MFSD dataset and testing it on the Replay-Attack dataset, and vice versa.

Table 7. Classification performance in terms of *HTER* (%). The models are trained using the CASIA-MFSD dataset and tested on the Replay-Attack dataset, and vice versa. 1: supervised method. 2: semi-supervised method.

Methods	Train	Test	Train	Test	Average
	CASIA MFSD	Replay Attack	Replay Attack	CASIA MFSD	
1-LBP ('13)	47.0%		39.6%		43.3%
1-LBP-TOP ('13)	49.7%		60.6%		55.2%
1-Motion ('13)	50.2%		47.9%		49.1%
1-CNN ('14)	48.5%		45.5%		47.0%
1-Color LBP ('15)	37.9%		35.4%		36.7%
1-Color Tex ('16)	30.3%		37.7%		34.0%
1-Auxiliary('18)	27.6%		28.4%		28.0%
1-De-Spoof('18)	28.5%		41.1%		34.8%
1-DA ('18)	27.4%		36.0%		31.7%
1-Dynamic texture ('18)	22.2%		35.0%		28.6%
1-OF Domain ('18)	30.1%		36.8%		33.5%
1-GFA-CNN ('19)	21.4%		34.3%		28.0%
1-ADA ('19)	17.5%		41.6%		29.6%
2-Proposed method	15.6%		44.1%		29.8%

We first consider training on the training set of the CASIA-MFSD database and testing on the testing set of the Replay-Attack database. The quantitative results shown in Tab. 7 confirm that the proposed method achieves the best performance (*HTER* = 0.156) on the Replay-Attack test set which includes different types of spoofing attacks. The competitors consist of LBP (de Freitas Pereira et al., 2013), LBP-TOP (de Freitas Pereira et al., 2013), Motion (de Freitas Pereira et al., 2013), CNN (Yang et al., 2014), Color LBP (Boulkenafet et al., 2015), Color Tex (Boulkenafet et al., 2016), Auxiliary (Liu et al., 2018), De-Spoof (Jourabloo et al., 2018), DA (Li et al., 2018b), Dynamic texture (Shao et al., 2018), OF Domain (Sun et al., 2018), ADA (Wang et al., 2019) and GFA-CNN (Tu et al., 2019b). In Fig. 14, it is obvious that the trained model has strong generalization ability to make live faces and fake faces obtained by using the attackers’ bare hands separable. How-



Fig. 11. First row: original optical flow images of live faces. Second row: generated optical flow images of live faces. Third row: corresponding maps which display the differences between the original images (the first row) and generated images (the second row) from model by red points.

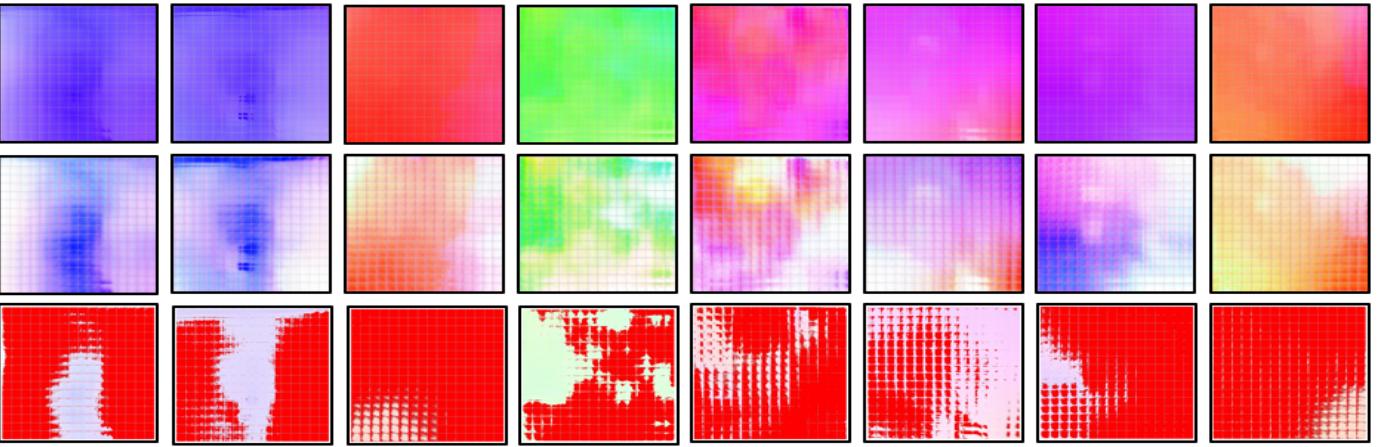


Fig. 12. First row: original optical flow images of spoofing faces. Second row: generated optical flow images of spoofing faces. Third row: corresponding maps which display the differences between original image (the first row) and generated images (the second row) by red point.

ever, some fake faces obtained by fixed support has the same distribution as live faces.

We then conduct the opposite experiment: training on the training set of the Replay-Attack dataset and testing on the testing set of the CASIA-MFSD dataset. Our method achieves a competitive performance ($HTER = 0.441$) for the cross testing on the testing set of the CASIA-MFSD dataset. From Tab. 7, we can see that the $HTER$ of our method is better than most binary supervision methods (Yang et al., 2014; Jourabloo et al., 2018; Wu et al., 2016; Boulkenafet et al., 2016, 2015). This demonstrates that the proposed approach can better identify the differences between live and fake faces.

As with all previous works (Wu et al. 2016; Jourabloo, Liu, and Liu 2018; Boulkenafet, Komulainen, and Hadid 2016), we observe that the models trained on CASIA-MFSD enables to generalize better than the model trained on the Replay Attack Database. We speculate as follows (1) It is probably because the resolution of the CASIA-MFSD data is significantly higher than that in the Replay-Attack dataset. The model trained with high resolution could generalize better than the model trained with low resolution. (2) Compared with Replay-Attack, the CASIA-

MFSD contains more variations in collected database, For example, imaging quality, the distance between camera and face, background and attack types. Hence, the model optimized for Replay-Attack databases faces more challenge in the new acquisition conditions. This is one limitation of the our method and previous works, and worthy further research. As with previous works (Wu et al., 2016; Jourabloo et al., 2018; Boulkenafet et al., 2016), the two above cross-database experiments do not have the same $HTER$. We speculate that it is probably because the resolution of the CASIA-MFSD data is significantly higher than that in the Replay-Attack dataset. In Fig. 11 and 12, the first row and second row show the optical flow maps and the reconstructed optical flow maps, respectively. The third row visualizes the differences between the corresponding optical maps and reconstructed maps (red color). These visualization results clearly demonstrate that the reconstruction errors from live faces are lower than that of spoofing faces.

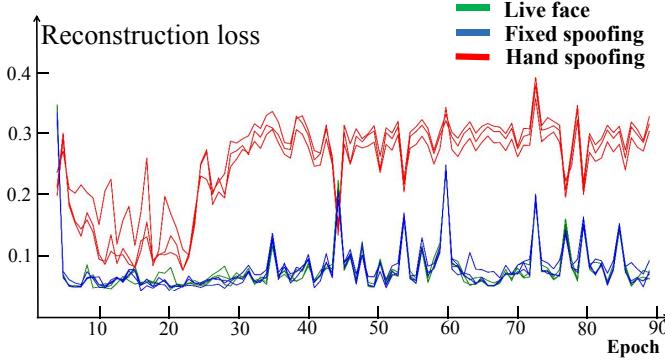


Fig. 13. Three kinds of reconstruction performance with different iteration numbers (epoch) on cross-database experiment with training on the training set of the CASIA-MFSD database and testing on the testing set of the Replay-Attack database.

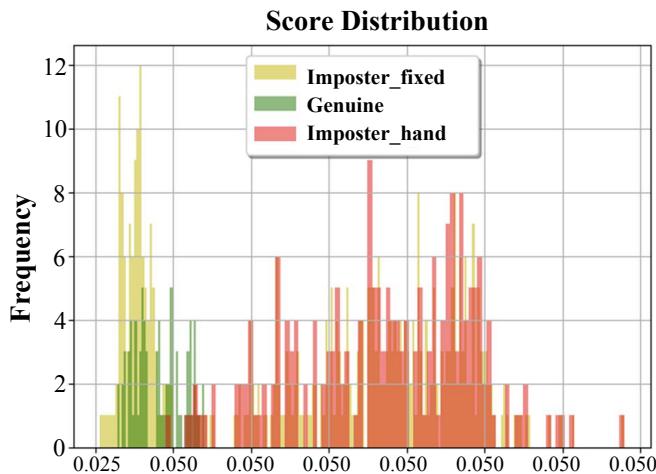


Fig. 14. Three distributions of reconstruction error scores in experiment with training on the training set of the CASIA-MFSD database and testing on the testing set of the Replay-Attack database.

5. Conclusion

We have presented an adversarial framework for the detection of spoofing faces. Given an input face image, the trained model can automatically determine if it is a live or spoof face. Current face anti-spoofing techniques have to utilize both spoof data and live data for training, which can hardly cover every type of spoof faces. By contrast, our approach does not need spoof data for training, and is thus semi-supervised and robust to different types of spoof faces. Both the intra-/cross-database experiments show that our method achieves better or comparable results to state-of-the-art techniques. We believe our research will arouse some new insights in this field.

References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. URL: <http://tensorflow.org/>. software available from tensorflow.org.
- Abati, D., Porrello, A., Calderara, S., Cucchiara, R., 2019. Latent space autoregression for novelty detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 481–490.
- Akcay, S., Atapour-Abarghouei, A., Breckon, T.P., 2018. Ganomaly: Semi-supervised anomaly detection via adversarial training, in: Asian Conference on Computer Vision, Springer, pp. 622–637.
- Alotaibi, A., Mahmood, A., 2017. Deep face liveness detection based on non-linear diffusion using convolution neural network. Signal, Image and Video Processing 11, 713–720.
- Atoum, Y., Liu, Y., Jourabloo, A., Liu, X., 2017. Face anti-spoofing using patch and depth-based cnns, in: 2017 IEEE International Joint Conference on Biometrics (IJCB), IEEE, pp. 319–328.
- Bao, W., Li, H., Li, N., Jiang, W., 2009. A liveness detection method for face recognition based on optical flow field, in: 2009 International Conference on Image Analysis and Signal Processing, IEEE, pp. 233–236.
- Baur, C., Wiestler, B., Albarqouni, S., Navab, N., 2018. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images, in: International MICCAI Brainlesion Workshop, Springer, pp. 161–169.
- Bengio, S., Mariéthoz, J., 2004. A statistical significance test for person authentication, in: Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop.
- Bishop, C.M., 2006. Pattern recognition and machine learning. springer.
- Boulkenafet, Z., Komulainen, J., Hadid, A., 2015. Face anti-spoofing based on color texture analysis, in: 2015 IEEE international conference on image processing (ICIP), IEEE, pp. 2636–2640.
- Boulkenafet, Z., Komulainen, J., Hadid, A., 2016. Face antispoofing using speeded-up robust features and fisher vector encoding. IEEE Signal Processing Letters 24, 141–145.
- Boulkenafet, Z., Komulainen, J., Hadid, A., 2017. Face antispoofing using speeded-up robust features and fisher vector encoding. IEEE Signal Processing Letters 24, 141–145.
- Chetty, G., 2010. Biometric liveness checking using multimodal fuzzy fusion, in: International Conference on Fuzzy Systems, IEEE, pp. 1–8.
- Chetty, G., Wagner, M., 2006. Multi-level liveness verification for face-voice biometric authentication, in: 2006 Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference, IEEE, pp. 1–6.
- Chingovska, I., Anjos, A., Marcel, S., 2012. 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), IEEE, pp. 1–7.
- Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A., 2018. Generative adversarial networks: An overview. IEEE Signal Processing Magazine 35, 53–65.
- Feng, L., Po, L.M., Li, Y., Xu, X., Yuan, F., Cheung, T.C.H., Cheung, K.W., 2016. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. Journal of Visual Communication and Image Representation 38, 451–460.
- Finkel, J.R., Manning, C.D., 2009. Hierarchical bayesian domain adaptation, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 602–610.
- de Freitas Pereira, T., Anjos, A., De Martino, J.M., Marcel, S., 2012. Lbp-top based countermeasure against face spoofing attacks, in: Asian Conference on Computer Vision, Springer, pp. 121–132.
- de Freitas Pereira, T., Anjos, A., De Martino, J.M., Marcel, S., 2013. Can face anti-spoofing countermeasures work in a real world scenario?, in: 2013 international conference on biometrics (ICB), IEEE, pp. 1–8.
- Frischholz, R.W., Dieckmann, U., 2000. Biold: a multimodal biometric identification system. Computer 33, 64–68.
- Frischholz, R.W., Werner, A., 2003. Avoiding replay-attacks in a face recognition system using head-pose estimation, in: Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on, IEEE, pp. 234–235.
- Galbally, J., Marcel, S., Fierrez, J., 2014. Biometric antispoofing methods: A survey in face recognition. IEEE Access 2, 1530–1552.
- Gan, J., Li, S., Zhai, Y., Liu, C., 2017. 3d convolutional neural network based on face anti-spoofing, in: 2017 2nd international conference on multimedia and image processing (ICMIP), IEEE, pp. 1–5.
- George, A., Mostaani, Z., Geissenbuhler, D., Nikisins, O., Anjos, A., Marcel,

- S., 2019. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security* 15, 42–55.
- Hadsell, R., Chopra, S., LeCun, Y., 2006. Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE. pp. 1735–1742.
- Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T., 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 1647–1655.
- Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.
- Jee, H.K., Jung, S.U., Yoo, J.H., 2006. Liveness detection for embedded face recognition system. *International Journal of Biological and Medical Sciences* 1, 235–238.
- Jourabloo, A., Liu, Y., Liu, X., 2018. Face de-spoofing: Anti-spoofing via noise modeling, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 290–306.
- Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al., 2016. Conditional image generation with pixelcnn decoders. *Advances in Neural Information Processing Systems*.
- Kalchbrenner, N., Grefenstette, E., Blunsom, P., 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188* .
- Kim, W., Suh, S., Han, J.J., 2015. Face liveness detection from a single image via diffusion speed model. *IEEE transactions on Image processing* 24, 2456–2465.
- Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .
- Kollreider, K., Fronthaler, H., Bigun, J., 2005. Evaluating liveness by face images and the structure tensor, in: Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05), IEEE. pp. 75–80.
- Komulainen, J., Hadid, A., Pietikäinen, M., 2013a. Context based face anti-spoofing, in: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), IEEE. pp. 1–8.
- Komulainen, J., Hadid, A., Pietikäinen, M., Anjos, A., Marcel, S., 2013b. Complementary countermeasures for detecting scenic face spoofing attacks, in: 2013 International conference on biometrics (ICB), IEEE. pp. 1–7.
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.
- Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D., 1997. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks* 8, 98–113.
- Li, H., He, P., Wang, S., Rocha, A., Jiang, X., Kot, A.C., 2018a. Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security* 13, 2639–2652.
- Li, H., Li, W., Cao, H., Wang, S., Huang, F., Kot, A.C., 2018b. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security* 13, 1794–1809.
- Li, J., Wang, Y., Tan, T., Jain, A.K., 2004. Live face detection based on the analysis of fourier spectra, in: Biometric Technology for Human Identification, International Society for Optics and Photonics. pp. 296–304.
- Li, L., Feng, X., Boulkenafet, Z., Xia, Z., Li, M., Hadid, A., 2016. An original face anti-spoofing approach using partial convolutional neural network, in: 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE. pp. 1–6.
- Liu, Y., Jourabloo, A., Liu, X., 2018. Learning deep models for face anti-spoofing: Binary or auxiliary supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 389–398.
- López, N.M.M., 2017. Accurate estimation of object motion in image sequences. Ph.D. thesis. Universidad de Las Palmas de Gran Canaria.
- Määttä, J., Hadid, A., Pietikäinen, M., 2011. Face spoofing detection from single images using micro-texture analysis, in: 2011 international joint conference on Biometrics (IJCB), IEEE. pp. 1–7.
- Mirjalili, V., Ross, A., 2017. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender, in: 2017 IEEE International joint conference on biometrics (IJCB), IEEE. pp. 564–573.
- Patel, K., Han, H., Jain, A.K., 2016a. Cross-database face antispoofing with robust feature representation, in: Chinese Conference on Biometric Recognition, Springer. pp. 611–619.
- Patel, K., Han, H., Jain, A.K., 2016b. Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security* 11, 2268–2283.
- Peixoto, B., Michelassi, C., Rocha, A., 2011. Face liveness detection under bad illumination conditions, in: 2011 18th IEEE International Conference on Image Processing, IEEE. pp. 3557–3560.
- Ratha, N.K., Connell, J.H., Bolle, R.M., 2001. An analysis of minutiae matching strength, in: International Conference on Audio-and Video-Based Biometric Person Authentication, Springer. pp. 223–228.
- Rehman, Y.A.U., Po, L.M., Komulainen, J., 2020. Enhancing deep discriminative feature maps via perturbation for face presentation attack detection. *Image and Vision Computing* 94, 103858.
- Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M., 2018. Deep one-class classification, in: International conference on machine learning, pp. 4393–4402.
- Sabokrou, M., Khalooei, M., Fathy, M., Adeli, E., 2018. Adversarially learned one-class classifier for novelty detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3379–3388.
- Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. *Neural computation* 13, 1443–1471.
- Schwartz, W.R., Rocha, A., Pedrini, H., 2011. Face spoofing detection through partial least squares and low-level descriptors, in: 2011 International Joint Conference on Biometrics (IJCB), IEEE. pp. 1–8.
- Shao, R., Lan, X., Yuen, P.C., 2018. Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing. *IEEE Transactions on Information Forensics and Security* 14, 923–938.
- Smiatacz, M., 2012. Liveness measurements using optical flow for biometric person authentication. *Metrology and Measurement Systems* 19, 257–268.
- Sun, L., Pan, G., Wu, Z., Lao, S., 2007. Blinking-based live face detection using conditional random fields, in: International Conference on Biometrics, Springer. pp. 252–260.
- Sun, Z., Sun, L., Li, Q., 2018. Investigation in spatial-temporal domain for face spoof detection, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1538–1542.
- Tan, X., Li, Y., Liu, J., Jiang, L., 2010a. Face liveness detection from a single image with sparse low rank bilinear discriminative model, in: European Conference on Computer Vision, Springer. pp. 504–517.
- Tan, X., Li, Y., Liu, J., Jiang, L., 2010b. Face liveness detection from a single image with sparse low rank bilinear discriminative model, in: ECCV.
- Tu, X., Zhang, H., Xie, M., Luo, Y., Zhang, Y., Ma, Z., 2019a. Deep transfer across domains for face antispoofering. *Journal of Electronic Imaging* 28, 043001.
- Tu, X., Zhao, J., Xie, M., Du, G., Zhang, H., Li, J., Ma, Z., Feng, J., 2019b. Learning generalizable and identity-discriminative representations for face anti-spoofing. *arXiv preprint arXiv:1901.05602* .
- Wang, G., Han, H., Shan, S., Chen, X., 2019. Improving cross-database face presentation attack detection via adversarial domain adaptation, in: International Conference on Biometrics (ICB).
- Wang, T., Yang, J., Lei, Z., Liao, S., Li, S.Z., 2013. Face liveness detection using 3d structure recovered from a single camera, in: 2013 International Conference on Biometrics (ICB), IEEE. pp. 1–6.
- Wu, B.F., Chu, Y.W., Huang, P.W., Chung, M.L., Lin, T.M., 2016. A motion robust remote-ppg approach to drivers health state monitoring, in: Asian Conference on Computer Vision, Springer. pp. 463–476.
- Xia, J., Tang, Y., Jia, X., Shen, L., Lai, Z., 2019. Latent spatial features based on generative adversarial networks for face anti-spoofing, in: Chinese Conference on Biometric Recognition, Springer. pp. 240–249.
- Xu, Z., Li, S., Deng, W., 2015. Learning temporal features using lstm-cnn architecture for face anti-spoofing, in: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), IEEE. pp. 141–145.
- Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., Zuo, W., 2017. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2272–2281.
- Yang, J., Lei, Z., Li, S.Z., 2014. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601* .
- Yang, J., Lei, Z., Liao, S., Li, S.Z., 2013. Face liveness detection with component dependent descriptor, in: 2013 International Conference on Biometrics (ICB), IEEE. pp. 1–6.
- Yu, C., Jia, Y., 2017. Anisotropic diffusion-based kernel matrix model for face liveness detection. *arXiv preprint arXiv:1707.02692* .

- Zenati, H., Romain, M., Foo, C.S., Lecouat, B., Chandrasekhar, V., 2018. Adversarially learned anomaly detection, in: 2018 IEEE International Conference on Data Mining (ICDM), IEEE. pp. 727–736.
- Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z., 2012. A face antispoofing database with diverse attacks, in: 2012 5th IAPR international conference on Biometrics (ICB), IEEE. pp. 26–31.
- Zhang, Z., Yi, D., Lei, Z., Li, S.Z., et al., 2011. Face liveness detection by learning multispectral reflectance distributions., in: FG, pp. 436–441.