Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061019

ARTICLE

PNSS: Unknown Face Presentation Attack Detection with Pseudo Negative Sample Synthesis

Hongyang Wang[1,2], Yichen Shi[3], Jun Feng[1,2,*], Zitong Yu[4] and Zhuofu Tao[5]

[1]School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang, 050043, China
[2]Shijiazhuang Key Laboratory of Artificial Intelligence, Shijiazhuang Tiedao University, Shijiazhuang, 050043, China
[3]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
[4]School of Computer Science and Technology, Great Bay University, Dongguan, 523808, China
[5]School of Electrical and Computer Engineering, University of California, Los Angeles, CA 90095, USA

*Corresponding Author: Jun Feng. Email: fengjun@stdu.edu.cn

Received: 14 November 2024; Accepted: 11 February 2025; Published: 16 April 2025

ABSTRACT: Face Presentation Attack Detection (fPAD) plays a vital role in securing face recognition systems against various presentation attacks. While supervised learning-based methods demonstrate effectiveness, they are prone to overfitting to known attack types and struggle to generalize to novel attack scenarios. Recent studies have explored formulating fPAD as an anomaly detection problem or one-class classification task, enabling the training of generalized models for unknown attack detection. However, conventional anomaly detection approaches encounter difficulties in precisely delineating the boundary between bonafide samples and unknown attacks. To address this challenge, we propose a novel framework focusing on unknown attack detection using exclusively bonafide facial data during training. The core innovation lies in our pseudo-negative sample synthesis (PNSS) strategy, which facilitates learning of compact decision boundaries between bonafide faces and potential attack variations. Specifically, PNSS generates synthetic negative samples within low-likelihood regions of the bonafide feature space to represent diverse unknown attack patterns. To overcome the inherent imbalance between positive and synthetic negative samples during iterative training, we implement a dual-loss mechanism combining focal loss for classification optimization with pairwise confusion loss as a regularizer. This architecture effectively mitigates model bias towards bonafide samples while maintaining discriminative power. Comprehensive evaluations across three benchmark datasets validate the framework's superior performance. Notably, our PNSS achieves 8%–18% average classification error rate (ACER) reduction compared with state-of-the-art one-class fPAD methods in cross-dataset evaluations on Idiap Replay-Attack and MSU-MFSD datasets.

KEYWORDS: Face presentation attack detection; pseudo negative sample; anomaly detection; one-class classification

1 Introduction

Face recognition is widely used due to its low cost, accessibility and specificity [1], but it is vulnerable to presentation attacks (PAs), such as photo prints, video replays, and 3D masks. The growing sophistication of these attacks poses serious challenges to the security and reliability of face recognition systems. Face Presentation Attack Detection (fPAD) is crucial for protecting these systems [2], as it must distinguish between bonafide and attack presentations.

In recent years, researchers have used handcrafted feature-based methods or deep learning techniques to address the attack detection task, which is typically framed as a supervised binary classification problem and thus requires large-scale datasets [3,4]. However, these methods often rely heavily on known attack types during training, leading to poor generalization performance when encountering unknown attacks at the testing stage. To mitigate these challenges, Domain Adaptation (DA) [5] and Domain Generalization (DG) [6] methods have been explored. DA reduces the discrepancy between labeled source domains and unlabeled target domains, while DG learns a shared feature space between known and unknown domains [7,8].

Despite their benefits, DA and DG rely on fully labeled source domains and attempt to cover all attack types, which is impractical. As an alternative, fPAD can be formulated as a one-class classification problem, where anomaly detection is used to identify spoofing as outliers [9,10]. This approach assumes bonafide face features are compact, while attack features vary significantly, allowing models to be trained using only bonafide samples and treating any deviations as attacks. However, the key challenge is accurately defining the decision boundary between bonafide faces and unknown attacks without explicit access to attack samples during the training stage.

Although factors like acquisition and sample variability introduce uncertainty, bonafide face images tend to cluster in feature space. Based on this, we make two assumptions: (1) bonafide face images follow a Gaussian distribution, and (2) there exists a boundary between bonafide and unknown attacks.

To address these challenges, we propose a novel Pseudo Negative Sample Synthesis (PNSS) strategy that generates pseudo-negative samples from the low-likelihood regions of the bonafide feature space. By simulating unknown attack samples, PNSS enables the model to learn a compact and robust decision boundary between bonafide and attack classes without relying on real attack data during training. Experimental results demonstrate that PNSS significantly outperforms existing methods, achieving notable reductions in error rates on public datasets.

The main contributions of this paper are summarized as follows:

- We introduce a novel anomaly detection-based fPAD framework, only using bonafide face images as positive samples.
- We propose a novel pseudo negative sample synthesis strategy that synthesizes negative samples from the low likelihood region of positive feature space.
- Extensive experiments on three datasets demonstrate that our method achieves remarkable improvements over existing one-class fPAD methods.

2 Related Work

2.1 Face Presentation Attack Detection Methods

Face Presentation Attack Detection (fPAD) methods have evolved through two stages: traditional methods and deep learning-based methods. Traditional approaches rely on hand-crafted features, such as LBP [11], SIFT [12], SURF [13], and HOG [14], combined with classifiers like LDA (Linear Discriminant Analysis) and SVM (Support Vector Machine) to distinguish between bonafide and spoof samples.

The rise of deep learning has enabled the use of CNNs (Convolutional Neural Networks) [15,16] and Vision Transformers (ViTs) [17], which improve performance by automatically extracting discriminative features. Auxiliary cues like depth maps [18], frequency domain features [19], reflection maps [20], and rPPG (Remote photoplethysmography) signals [21] further enhance accuracy. To mitigate overfitting and improve generalization, Domain Adaptation (DA) and Domain Generalization (DG) techniques have been proposed [5,6]. However, two-class-based methods heavily depend on known attack types during training, limiting their ability to generalize to unseen scenarios. This reliance on labeled attack data also leads to challenges in handling the diversity and complexity of real-world attacks.

One-class learning approaches aim to overcome these limitations by training exclusively on bonafide samples and treating deviations as anomalies. For instance, Baweja et al. [9] employed Gaussian-distributed pseudo-negatives, but their oversimplified assumptions limited the diversity of generated samples. Similarly, OC-SCMNet [22] refined decision boundaries using static latent cues, which are less effective in dynamic and complex real-world scenarios. Hyp-OC [23] enhanced feature compactness but did not address the need for diverse and representative pseudo-negative samples.

Our proposed PNSS method directly addresses these limitations by synthesizing pseudo-negatives from low-likelihood regions of the bonafide feature space. This approach avoids the oversimplified Gaussian assumptions of Baweja et al. [9] and the static cues of OC-SCMNet [22]. Unlike Hyp-OC [23], PNSS ensures diversity in pseudo-negative samples, aligning them with real-world attack scenarios. By defining more compact and robust decision boundaries, PNSS significantly improves its ability to detect unknown attacks.

2.2 Anomaly Detection Methods

Anomaly detection aims to identify abnormal samples that deviate significantly from the norm, often in the presence of covariate or semantic shifts. In the context of fPAD, anomaly detection methods are closely related to one-class classification, where models are trained exclusively on bonafide data, and deviations are treated as attacks [24]. Oza et al. [25] utilized an autoencoder-regularized CNN to learn compact feature embeddings for one-class classification, while Perera et al. [26] proposed a dual-minimax probability machine to effectively handle unknown impostors.

In the specific domain of fPAD, Boulkenafet et al. [27] initially compared one-class and two-class classifiers, demonstrating the potential of anomaly detection. Subsequently, Baweja et al. [9] employed Gaussian-distributed pseudo-negative samples to refine decision boundaries, while Huang et al. [22] introduced OC-SCMNet, leveraging latent spoof cues for enhanced classification accuracy. Du et al. [28] proposed Virtual Outlier Synthesis (VOS), a method that dynamically generates synthetic outliers to improve the generalization of the model to diverse types of attacks.

Despite their promising results, these methods have notable weaknesses. Gaussian-based pseudo-negative sampling methods, such as those proposed by Baweja et al. [9], often fail to generate diverse and representative negative samples due to their reliance on simplistic Gaussian distribution assumptions. This results in an inadequate representation of the real-world variability of attacks. While VOS [28] dynamically generates synthetic outliers to improve generalization, it is not specifically tailored for fPAD tasks and may not capture the complexities of attack scenarios encountered in face presentation attacks. These limitations hinder the ability of these methods to generalize effectively to new and unseen attacks, which is a critical challenge in fPAD.

PNSS overcomes these limitations by synthesizing pseudo-negative samples directly from the low-likelihood regions of the bonafide feature space. This approach ensures that the generated samples are relevant to the task and diverse enough to better represent the variety of real-world attacks. By addressing the issues of sample diversity and task-specific relevance, PNSS significantly improves generalization to unknown attack types, outperforming previous methods in real-world scenarios.

3 Proposed Method

Inspired by the existing out-of-distribution detection and fPAD methods [9,28], we propose a deep fPAD method based on anomaly detection with pseudo negative sample synthesis (PNSS). In this section, we will first introduce our network framework in Section 3.1, then introduce the pseudo negative sample synthesis strategy in Section 3.2. The loss function for model training is described in Section 3.3. And we offer PyTorch-like pseudocode of our methods in Section 3.4.

3.1 Overall Framework

In the framework, as shown in Fig. 1, training set only includes bonafide face images $\{x_i\}_{i=1}^N$ ($x_i \in [0, 255]^{3\times H\times W}$) where N is the size of training set and $H \times W$ is the spatial size. We use a traditional CNN as feature extraction backbone ω, and these images are fed into ω and get $\{\omega(x_i)\}_{i=1}^N$ ($\omega(x_i) \in \mathbb{R}^d$), where d is the dimension of the extracted feature. We define all $\omega(x_i)$ as positive set S_{bf} and sample some virtual outliers S_{pn} from the feature space (bf is short for bonafide and pn is short for pseudo negative). Then S_{bf} and S_{pn} are concatenated and fed into the classifier ϕ to train the attack detector.

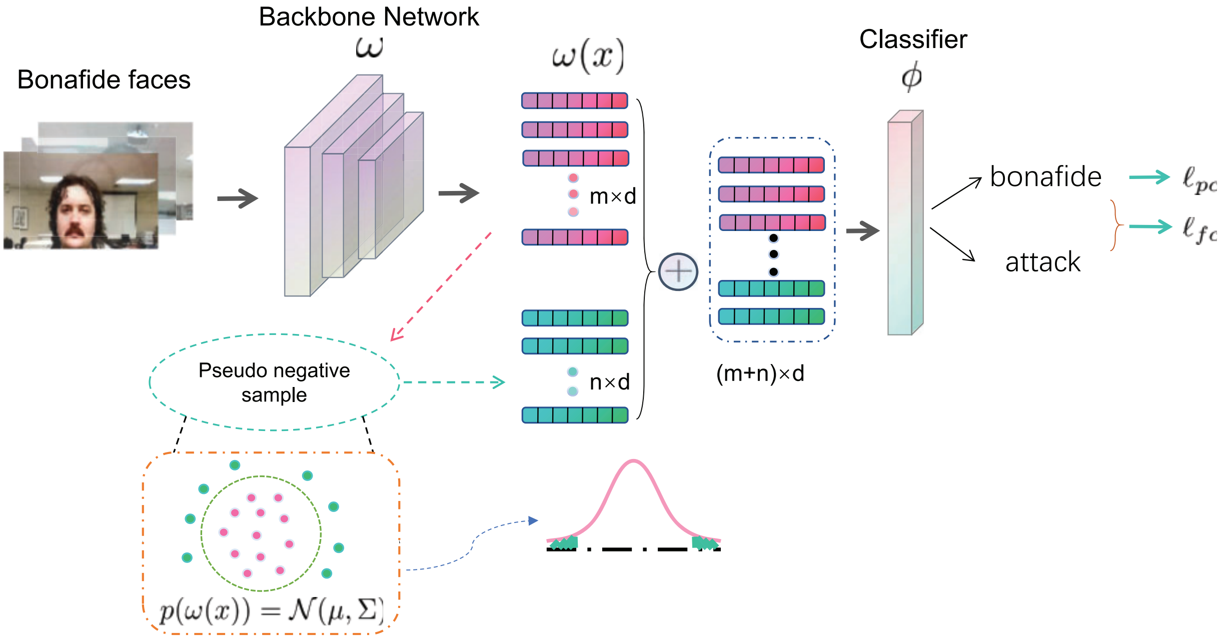

Figure 1: The framework of our proposed method. Input data x only includes bonafide face images in one batch, then a backbone network ω is used to extract image feature $\omega(x)$ as positive samples (in red). We generate pseudo negative samples (in green) from the low likelihood region of a Gaussian Distribution modeled by positive samples. Then concatenate all samples and feed them into Classifier ϕ to train an MLP network optimized by ℓ_{pc} and ℓ_{fl}. Note ℓ_{pc} is only calculated from positive samples

3.2 Pseudo Negative Feature Synthesis

Our goal is to synthesize negative samples S_{pn} from the feature space. Ideally, S_{pn} should help ϕ learn a compact boundary between known bonafide faces and unknown attacks. We assume that the positive samples S_{bf} forms a multivariate Gaussian distribution:

$$p(\omega(x_{bf})) = \mathcal{N}(\mu, \Sigma), \tag{1}$$

where ω is backbone, x_{bf} is the bonafide images, μ is the mean of positive samples and Σ is the covariance matrix.

We compute the empirical mean μ and Σ of positive distribution S_{bf}:

$$\mu = \frac{1}{N_{bf}} \sum_{i=1}^{N_{bf}} \omega(x_i), \tag{2}$$

$$\Sigma = \frac{1}{N_{bf}} \sum_{i=1}^{N_{bf}} (\omega(x_i) - \mu)(\omega(x_i) - \mu)^{\top}, \tag{3}$$

where N_{bf} is the number of bonafide face images in one iteration.

To generate unknown attack features for training stage(with bonafide face features) and get the compact boundary between bonafide face and various unknown presentation attacks, we propose a strategy to synthesize pseudo negative samples using the positive multivariate Gaussian distribution. Specifically, as shown in Fig. 1, we sample the pseudo negative samples from the ε-likelihood region of the following distribution:

$$S_{pn} = \left\{ s_{pn} \middle| \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(s_{pn} - \mu)^{\top}\Sigma^{-1}(s_{pn} - \mu) \right) < \varepsilon \right\}, \tag{4}$$

where S_{pn} represents the sampled negative features of presentation attacks (PAs), and ε is a small threshold that ensures the training of a compact boundary for the positive samples. The ε-likelihood region refers to areas of the feature space where the likelihood of a bonafide face sample occurring is very low. These low-likelihood regions are considered to be representative of potential unknown attacks.

Although ε can theoretically be infinitely small, in practice, we sample T points from the Gaussian distribution and select the k-th minimum likelihood as the pseudo-negative samples. The assumption is that when k is fixed, a larger T results in a smaller effective ε, leading to pseudo-negative samples closer to the low-likelihood regions of the bonafide distribution. This enables the model to establish a more compact decision boundary between bonafide faces and unknown attacks.

3.3 Loss Function

Since in one iteration, the selected k, the number of pseudo negative samples, may not equal the number of positive samples, we use the following focal loss [29] during training to deal with the class imbalance problem:

$$\ell_{fl}(p) = -(1 - p)^{\gamma} \log(p), \tag{5}$$

where p is the estimated probability and γ is focusing parameter.

Following [9], we employ pairwise confusion loss [30] to remove identity-specific information from bonafide images. This loss helps the model generalize better by encouraging the features of different bonafide samples to be more distinct from each other:

$$\ell_{pc} = \sum_{i} \sum_{j \neq i} |\omega_i - \omega_j|_2^2, \tag{6}$$

where ω_i represents the feature of image x_i. This loss term is computed exclusively on bonafide images, without including the pseudo-negative samples. The goal is to ensure that features of bonafide samples become less correlated, thus reducing the chances of the model learning identity-specific features. This helps the model focus on more general characteristics that differentiate bonafide presentations from unknown attacks.

The overall loss function is a linear combination of focal loss and pairwise confusion loss as follows:

$$\ell_{final} = \lambda_1 \ell_{pc} + \lambda_2 \ell_{fl}, \tag{7}$$

where λ_1 and λ_2 are coefficients of ℓ_{pc} and ℓ_{fl}, respectively.

3.4 Pseudocode

Algorithm 1 presents the PyTorch-like pseudocode of PNSS:

Algorithm 1: Pseudocode of PNSS, PyTorch-like

```
# f: backbone
# h: classification mlp
# minibatch bonafide face images x
for x in loader do
    # bf: features of bonafide faces
    # m_vec: mean of bf
    # cor_vec: covariance matrix of bf
    bf = f(x)
    m_vec = mean(bf)
    cor_vec = mm((bf - m_vec).t(), bf - m_vec)/len(bf)
    sampler = MV_Normal(m_vec, cor_vec)
    # get negative samples from the low likeli-hood region of bonafide face distribution
    _buffer = sampler.rsample((sample_num,))
    prob_density = sampler.log_prob(_buffer)
    _, index = tc.topk(-prob_density, k)
    negative = _buffer[index]
    L = lambda_1 * L_pc + lambda_2 * L_fl
    # back-propagate
    L.backward()
    update(f, h)
end for
```

4 Experiment

4.1 Experimental Setting

Datasets

Three public databases are used to test our method, including Oulu-NPU [31], MSU-MFSD [32], and Idiap Replay-Attack [33].

Idiap Replay-Attack: This dataset consists of 50 volunteers and 1300 videos, evenly distributed across two attack types: print attacks using photos on A4 paper and replay attacks displayed on iPhone/iPad screens. Videos were recorded under controlled conditions with variations in lighting and device quality. The main challenge lies in distinguishing real faces from spoofing materials, as replay attacks utilize high-resolution screens and print attacks maintain detailed photo textures, requiring robust feature extraction to handle these subtle differences.

MSU-MFSD: This dataset includes 35 volunteers and 380 videos, recorded using two cameras with different resolutions, covering both low-resolution and high-resolution devices. Spoofing attacks include print photo and screen replay attacks, with balanced distribution across devices. The primary challenge comes from variations in video quality due to camera resolution and attack material fidelity, such as inconsistent print quality and screen refresh rates, testing a model's adaptability to noisy inputs.

Oulu-NPU: This dataset consists of 55 volunteers and nearly 5000 videos, recorded under three environmental conditions: indoor, outdoor, and dark, using six mobile phones with diverse camera specifications. The dataset is balanced across attack types, including high-resolution photo and screen replay attacks, and introduces significant variability in lighting, device quality, and environmental complexity. These factors make it notably more challenging than other datasets, especially for generalization in both in-domain and cross-domain scenarios.

Evaluation metrics To compare our method with previous work, our method is evaluated using the following metrics: Average Classification Error Rate (ACER) [2], which is calculated by Attack Presentation Classification Error Rate (APCER) [2] and Bonafide Presentation Classification Error Rate (BPCER) [2]. They are defined as follows:

$$APCER = \frac{FP}{FP + TN}, \tag{8}$$

$$BPCER = \frac{FN}{FN + TP}, \tag{9}$$

$$ACER = \frac{APCER + BPCER}{2}, \tag{10}$$

where *TP*, *TN*, *FP*, *FN* are True Positive, True Negative, False Positive, and False Negative, respectively.

Implementation Details

We use MTCNN [34] to detect faces, which are then cropped and resized to 256 × 256 for uniform input size. Data augmentation includes random horizontal flipping and normalization with mean and standard deviation derived from the VGGFace dataset.

We adopt VGGFace [35], a pre-trained VGG16-based model, as the feature extractor. Specifically, the fc6 layer output is used as the face feature representation. The models are implemented in PyTorch [36] and trained on an NVIDIA A100 GPU using mixed precision for efficiency. The optimizer is Adam with a learning rate of 1×10^{-4} and a weight decay of 5×10^{-4}. In Eq. (5), the focal loss parameter γ is set to 2. In Eq. (7), λ_1 and λ_2 are set to 3 and 1, respectively, balancing the contributions of different loss terms.

The training uses a batch size of 80 for training and 160 for testing. The model is trained for 40 epochs with an α value of 0.8. White noise with a standard deviation of 1.0 is applied during training to enhance robustness.

4.2 Experimental Comparison

4.2.1 Comparison with Other Methods

Six baseline methods are chosen for comparison:

(1) **OC-CNN [10]:** One-Class CNN synthesizes samples from a Gaussian distribution centered at the origin as pseudo negative samples and trains a classifier with bonafide face samples, as shown in Fig. 2a. The distribution of OC-CNN is $N(0, I \cdot \sigma^2)$.

(2) **Adaptive Mean Estimation** [9]**:** Baweja et al. assume that the attacked images are very similar to bonafide images used during training. So images from PAs and bonafide faces are very close to each other in the feature space. So that we proposed the pseudo negative features sampling strategy from a shifted Gaussian distribution with adaptive mean estimation. Concretely, in an iteration, the mean μ^* of the Gaussian distribution is $\alpha\mu_{old} + (1-\alpha)\mu_{new}$, where μ_{old} is adaptive mean of previous batch and μ_{new} is empirical mean of current batch. The distribution of Adaptive Mean Estimation is $N(\mu_{new}, I \cdot \sigma^2)$.

(3) **OC-SVM** [37]**:** One-Class Support Vector Machine (OC-SVM) classifies normal samples and potential abnormal samples by creating hyperplane in high-dimensional space. OC-SVM is implemented using the Sklearn library with default parameters.

(4) **SVDD** [38]**:** Support Vector Data Description (SVDD) creates a hypersphere to wrap most normal samples to achieve anomaly detection. SVDD is implemented using the LibSVM library.

(5) **MD** [39]**:** Mahalanobis distance (MD) can be used for anomaly detection. Calculate the boundary threshold from the data center with normal data, and then judge whether the position of a point from the center of the data set exceeds the threshold. If it exceeds the threshold, it is determined as an abnormal point. MD is implemented using the Sklearn library.

(6) **OC-GMM** [40]**:** The Gaussian mixture model can be regarded as a model composed of k single Gaussian models, and the k sub models are the hidden variables of the mixture model.

(7) **OC-SCMNet** [22]**:** One-Class Spoof Cue Map estimation Network (OC-SCMNet) uses a novel approach that leverages Spoof Cue Maps (SCMs) and a memory bank to generate and explore latent spoof features. This method aims to improve the decision boundary between live and spoof samples by guiding the feature learning process using pseudo spoof cue maps, which are synthesized from the latent feature space.

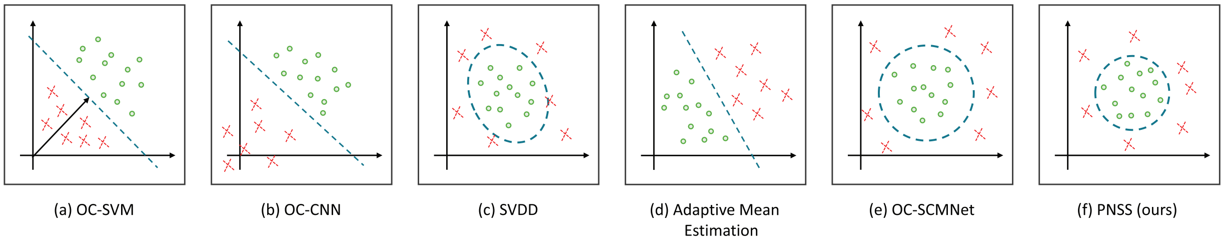
(a) OC-SVM (b) OC-CNN (c) SVDD (d) Adaptive Mean Estimation (e) OC-SCMNet (f) PNSS (ours)

Figure 2: Graphical illustration of the popular statistical one-class methods. The green circle represents the target data, the red dotted cross represents the unknown data, and the blue green dotted line represents the decision boundary of each method

Among the seven baseline methods, OC-CNN, Adaptive Mean Estimation, and OC-SCMNet are more related to our method. The main difference for us is that our method generates pseudo samples from the low likelihood region of the original bonafide face distribution, while comparative method use a new distribution to model unknown attacks.

To provide a comprehensive evaluation of the proposed PNSS method, we summarize the experimental results across intra-testing and cross-testing scenarios in Table 1. This table highlights the average ACER performance of different methods under both testing conditions, providing a overview before delving into the detailed results in the following subsections.

Table 1: Summary of average ACER results for intra-testing and cross-testing

Method	Intra-testing AVG (%)	Inter-testing AVG (%)	Overall AVG (%)
OC-CNN [10]	40.34	32.06	36.20
Adaptive mean estimation [9]	26.39	33.69	30.04
OC-SCMNet [20]	21.47	34.66	28.07
PNSS (Ours)	**16.47**	**26.84**	**21.65**

Note: Bold values denote the best performance.

The results in Table 1 demonstrate the effectiveness of the proposed PNSS method in addressing domain generalization challenges in face presentation attack detection. PNSS achieves the lowest ACER scores in both intra-testing (16.47%) and cross-testing (26.84%), resulting in an overall ACER of 21.65%. This superior performance highlights its ability to generate diverse and representative pseudo-negative samples, which enhance robustness to domain shifts and improve generalization to unseen attack scenarios. These results validate PNSS as a reliable approach for practical applications across diverse environments.

In the following subsections, we provide detailed results for each testing scenario. Section 4.2.2 presents the intra-testing results on three datasets, while Section 4.2.3 discusses the inter-testing performance across six domain pairs.

4.2.2 Intra-Testing

In intra-testing, we follow the protocol used in [9], where the models are trained using only bonafide presentation image data and evaluated on the test set with both bonafide images and attacked images. The identities of training set and test set are non-overlapping. As shown in Table 2, the proposed PNSS achieves competitive performance in three datasets. In Idiap Replay-Attack and MSU-MFSD, our method achieves the best results and reduces about 8%–18% ACER over other methods.

Table 2: Intra-testing results (ACER(%)) of one-class fPAD methods on three benchmark datasets

Method	Idiap Replay-Attack	MSU-MFSD	Oulu-NPU
OC-SVM [37]	31.14	–	47.56
SVDD [38]	32.96	–	47.52
MD [39]	31.75	–	45.19
OC-GMM [40]	30.10	–	46.96
OC-CNN [10]	35.99	39.22	45.80
Adaptive mean estimation [9]	20.74	28.20	30.24
OC-SCMNet [22]	14.70	27.22	**23.49**
PNSS (Ours)	**6.67**	**8.93**	33.80

Note: Bold values denote the best performance, while underlined values indicate the second-best.

The models most similar to our method are OC-CNN [10], Adaptive Mean Estimation [9], and OC-SCMNet [22]. As shown in Fig. 2, OC-CNN synthesizes samples from a Gaussian distribution centered at the origin as pseudo negative samples, while Adaptive Mean Estimation synthesizes samples from a shifted distribution among batches. OC-SCMNet utilizes Spoof Cue Maps (SCMs) and a memory bank to generate potential spoof features, thereby expanding the decision boundary. However, these methods generate pseudo negative samples or features with limited diversity, which may not effectively represent all attack types.

Our method does not estimate the parameters of the pseudo negative distribution, and uses the low likelihood samples from bonafide face feature space as pseudo negative samples to obtain a more compact boundary. As shown in PyTorch-like pseudocode of Section 3.4, the clear differences between the three methods are the strategy of negative sample synthesis. Fig. 3 gives the visualization of the feature distributions on Idiap Replay-Attack dataset via t-SNE where Fig. 3a and b are distribution representation of PNSS and Adaptive Mean Estimation [9], respectively. The green points represent bonafide face image data, red points represent PAs data and blue crosses represent the pseudo negative samples. Some blue crosses overlapped with red points in Fig. 3a, which are noises for model training due to their wrong label. Only few blue crosses overlap with red points in Fig. 3b, demonstrating the pseudo negative samples are consistent with the PAs in practical. The results show that the negative sample synthesis strategy of PNSS is more accurate and effective compared with Adaptive Mean Estimation [9].

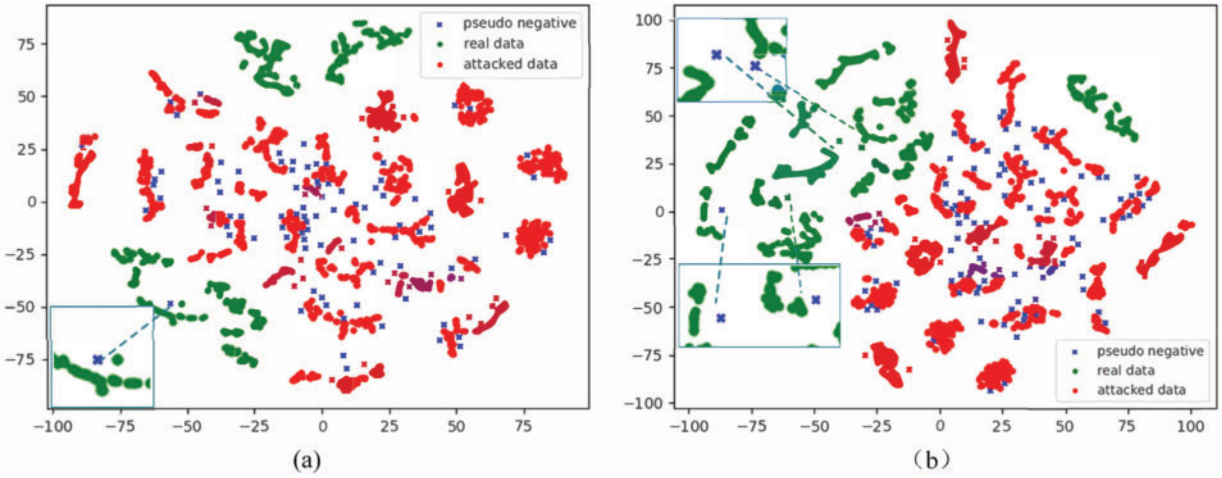

Figure 3: Visualization of the feature distributions on Idiap Replay-Attack dataset via t-SNE. The green points represent bonafide face image data, red points represent attack presentation data and blue crosses represent the pseudo negative samples used for estimating attack in bonafide world. (a) and (b) are visualization for PNSS and Adaptive Mean Estimation [9]

To evaluate PNSS, we visualize feature distributions on the MSU-MFSD and OULU-NPU datasets using t-SNE. In Fig. 4, pseudo-negative samples effectively separate bonafide and attack samples in the controlled MSU-MFSD dataset. In Fig. 5, despite the higher variability in OULU-NPU, pseudo-negative samples maintain clear boundaries, demonstrating PNSS's robustness and adaptability across diverse domains.

4.2.3 Inter-Testing

Inter-testing is a challenging scenario since the model faces distribution shifts caused by various factors between source domains and target domains. To evaluate the generalization capability of our method, we conduct experiments on Idiap ReplayAttack (denoted as R), MSU-MFSD (denoted as M) and OuluNPU (denoted as O). A-B means that train on dataset A and teston dataset B. As shown in Table 3, the proposed PNSS method achieves the lowest ACER in the M-O, O-M, and O-R scenarios, highlighting its strong generalization capability across diverse datasets. Although PNSS does not attain the best performance in the M-R and R-M scenarios—where OC-CNN and OC-SCMNet exhibit slightly better results—PNSS still secures the second-best outcomes, as indicated by the underlined values in the table. This consistent performance across multiple scenarios underscores the robustness of PNSS. The success of PNSS can be attributed to its effective use of a compact boundary between positive and negative samples, enabling it to generalize well even in cross-dataset testing where the training and testing sets differ significantly.

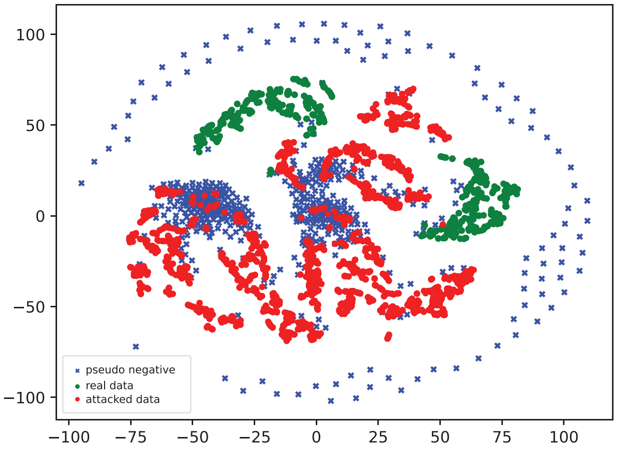

Figure 4: Visualization of feature distributions on MSU-MFSD dataset

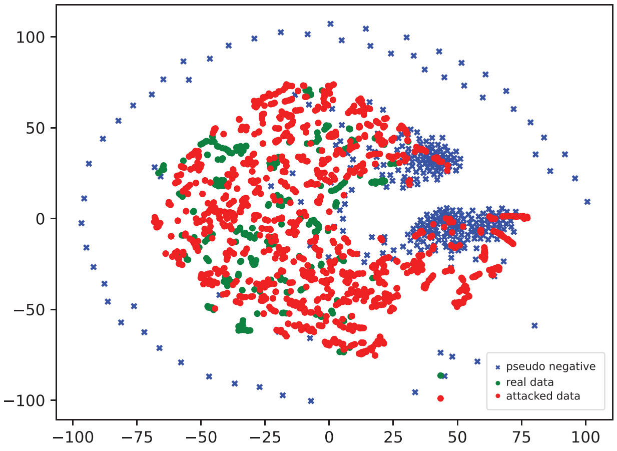

Figure 5: Visualization of feature distributions on OULU-NPU dataset

Table 3: Cross-testing results (ACER(%)) of one-class fPAD methods on three benchmark datasets

Method	**M-R**	**R-M**	**M-O**	**O-M**	**R-O**	**O-R**
OC-CNN [10]	**18.16**	38.23	35.99	34.01	38.76	35.21
Adaptive mean estimation [9]	39.38	39.82	33.99	30.40	**34.09**	32.83
OC-SCMNet [22]	44.05	**14.70**	41.71	33.58	40.66	32.92
PNSS (Ours)	21.87	15.31	**31.05**	**26.62**	38.04	**23.19**

Note: Bold values denote the best performance, while underlined values indicate the second-best.

As shown in Table 3, the proposed PNSS achieves progress in both M-R and R-M scenarios compared with other two methods. And we get the best performance on R-M, MO, O-M, and O-R settings. We believe this is because of the great representation ability of CNN and the compact boundary between positive samples and negative samples.

4.3 Ablation Study

In this subsection, we will exploit the effect of the *combined loss* in Eq. (7). Then we study the effect of the number of pseudo negative samples. Finally, we investigate the effect of the ratio β of pseudo negative samples against all samples from the distribution of bonafide face features.

Effect of the combined loss function. To evaluate the effect of the *combined loss*, we conducted experiments on the Idiap Replay-Attack dataset, as shown in Fig. 6. We fix the number of positive samples and negative samples to 32 and 16 in one batch. The results indicate that class imbalance significantly affects model accuracy. Using *focal loss* achieves an ACER of 9.97%, which outperforms *cross-entropy*. Adding the *pairwise confusion loss* further reduces ACER by approximately 3% for both losses. We also tested different values of the focusing parameter γ in *focal loss* (Fig. 7), with $\gamma = 2$ yielding the best performance, improving results over *cross-entropy* with $\gamma = 1.5$. Additionally, an ablation study on the coefficients λ_1 and λ_2 in the loss function (Table 4) shows that their combination significantly impacts ACER. Setting $\lambda_1 = 3$ and $\lambda_2 = 1$ achieves the best performance with an ACER of 6.67%.

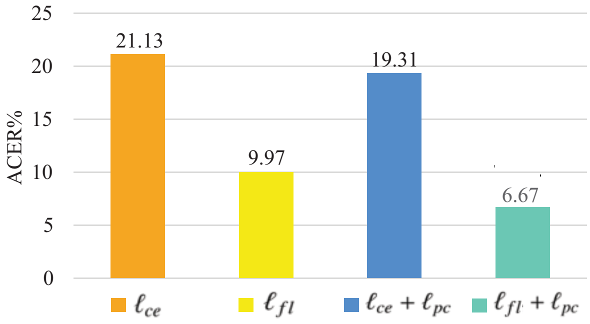

Figure 6: The impact of each item in combined loss

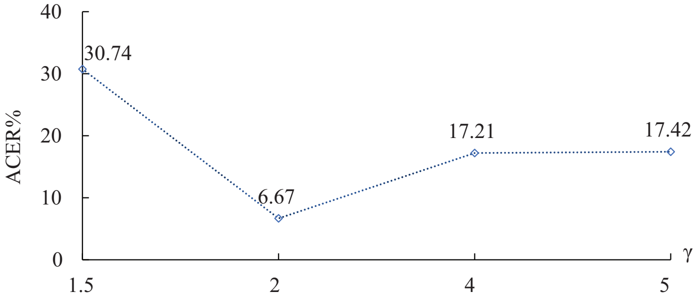

Figure 7: The impact of ratio γ

Effect of pseudo negative samples number. To investigate the effect of the number of pseudo negative samples, we do experiments on Idiap Replay-Attack, where varying the number in $\{2, 4, 6, 8, 16\}$ and fix the ratio β to 1:2000. As shown in Fig. 8, increasing the number of samples can significantly improve the ability of the model.

Table 4: Ablation study results for loss function coefficients λ_1 and λ_2. The metric ACER (%) and total loss ℓ_{final} are reported

Coefficient of ℓ_{pc} (λ_1)	Coefficient of ℓ_{fl} (λ_2)	Total loss (ℓ_{final})	ACER (%)
$\lambda_1 = 1$	$\lambda_2 = 1$	$\ell_{final} = \ell_{pc} + \ell_{fl}$	9.12
$\lambda_1= 1$	$\lambda_2= 2$	$\ell_{final} = \ell_{pc} + 2 \cdot \ell_{fl}$	7.12
$\lambda_1 = 1$	$\lambda_2 = 3$	$\ell_{final} = \ell_{pc} + 3 \cdot \ell_{fl}$	14.53
$\lambda_1 = 1$	$\lambda_2 = 4$	$\ell_{final} = \ell_{pc} + 4 \cdot \ell_{fl}$	11.51
$\lambda_1 = 2$	$\lambda_2 = 1$	$\ell_{final} = 2 \cdot \ell_{pc} + \ell_{fl}$	12.02
$\lambda_1 = 3$	$\lambda_2 = 1$	$\ell_{final} = 3 \cdot \ell_{pc} + \ell_{fl}$	**6.67**
$\lambda_1 = 4$	$\lambda_2 = 1$	$\ell_{final} = 4 \cdot \ell_{pc} + \ell_{fl}$	7.63

Note: Bold values denote the best performance.

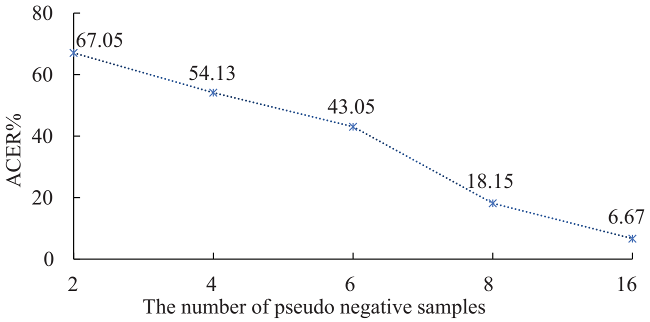

Figure 8: The impact of the number of pseudo negative samples

Effect of $\boldsymbol{\beta}$. Since the threshold ε can be infinitesimally small, to study the effect of ε, we set the number of pseudo negative samples as 16 and vary β in $\{1{:}200, 1{:}500, 1{:}1000, 1{:}2000\}$. A smaller β corresponds a smaller ε. As shown in Fig. 9, with the ratio β going down, a more compact boundary is learned to distinguish bonafide from PAs.

5 Discussion

The proposed PNSS method effectively detects unknown presentation attacks by leveraging pseudo-negative samples, achieving superior generalization to unseen attacks without relying on labeled attack data. Its dynamic synthesis of pseudo-negative samples enhances robustness across domains, addressing key limitations of traditional methods.

However, challenges remain in real-world scenarios, where factors like varying lighting, diverse devices, and complex backgrounds may affect performance. Sophisticated attack techniques, such as 3D masks or video replays, also require further exploration. Furthermore, existing datasets may not fully represent real-world complexities, highlighting the need for more diverse and representative benchmarks.

Despite these challenges, PNSS provides a solid foundation for addressing unknown attacks, with potential for further improvements in robustness and real-world applicability.

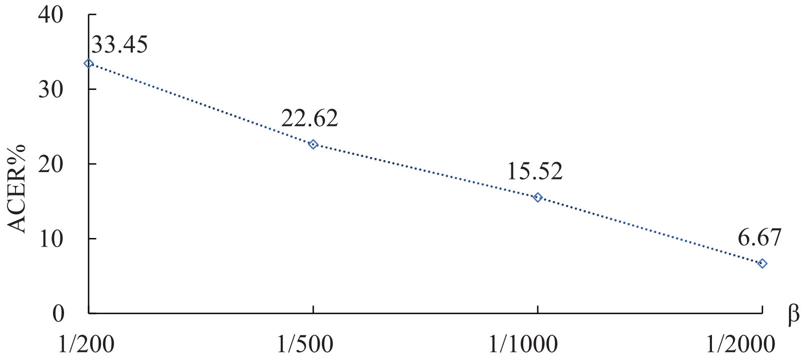

Figure 9: The impact of ratio β

6 Conclusion

This paper presents a pseudo-negative sample synthesis (PNSS) framework for one-class face presentation attack detection. Our approach addresses the boundary definition challenge through two key contributions: (1) Generation of representative negative samples from low-likelihood regions in the bonafide feature space, and (2) a dual-loss training paradigm combining *focal loss* for class imbalance with *pairwise confusion loss* regularization to prevent model bias. Extensive evaluations on Idiap Replay-Attack, MSU-MFSD, and other benchmarks demonstrate superior cross-dataset generalization, achieving 8%–18% average classification error rate (ACER) reductions compared to state-of-the-art methods.

Future investigations should focus on three directions: lightweight architecture integration for edge deployment, multi-modal feature fusion against advanced attacks, and adaptive sample synthesis for evolving threats. These developments will facilitate practical implementation in resource-constrained scenarios while maintaining detection robustness.

Acknowledgement: We thank the EIT High Performance Computing Platform for providing computational resources for this project.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grants 61972267, and 61772070, and in part by the Natural Science Foundation of Hebei Province under Grant F2024210005.

Author Contributions: Study conception and design: Hongyang Wang, Yichen Shi, Jun Feng, Zitong Yu; data collection: Hongyang Wang, Yichen Shi; analysis and interpretation of results: Yichen Shi, Zhuofu Tao; draft manuscript preparation: Hongyang Wang, Yichen Shi, Jun Feng, Zitong Yu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Guo J, Zhu X, Zhao C, Cao D, Lei Z, Li SZ. Learning meta face recognition in unseen domains. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2020; Seattle, WA, USA: IEEE. p. 6162–71.

2. Yu Z, Qin Y, Li X, Zhao C, Lei Z, Zhao G. Deep learning for face anti-spoofing: a survey. arXiv:2106.14948. 2021.
3. Liu A, Tan Z, Wan J, Escalera S, Guo G, Li SZ. CASIA-SURF CeFA: a benchmark for multi-modal cross-ethnicity face anti-spoofing. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV); 2021; Piscataway: IEEE. p. 1179–87.
4. Zhang Y, Yin Z, Li Y, Yin G, Yan J, Shao J, et al. Celeba-spoof: large-scale face anti-spoofing dataset with rich annotations. In: European Conference on Computer Vision (ECCV); 2020; Cham: Springer.
5. Jia Y, Zhang J, Shan S, Chen X. Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing. Pattern Recognitn. 2021;115:107888. doi:10.1016/j.patcog.2021.107888.
6. Liu M, Mu J, Yu Z, Ruan K, Shu B, Yang J. Adversarial learning and decomposition-based domain generalization for face anti-spoofing. Pattern Recognit Lett. 2022;155(1):171–7. doi:10.1016/j.patrec.2021.10.014.
7. Csurka G. Domain adaptation for visual applications: a comprehensive survey. arXiv:1702.05374. 2017.
8. Zhou K, Liu Z, Qiao Y, Xiang T, Loy CC. Domain generalization: a survey. arXiv:2103.02503. 2022.
9. Baweja Y, Oza P, Perera P, Patel VM. Anomaly detection-based unknown face presentation attack detection. In: IEEE International Joint Conference on Biometrics (IJCB); 2020 Sep 28–Oct 1; Houston, TX, USA: IEEE. p. 1–9.
10. Oza P, Patel VM. One-class convolutional neural network. IEEE Signal Process Lett. 2019;26(2):277–81. doi:10.1109/LSP.2018.2889273.
11. de Freitas Pereira T, Anjos A, Martino JMD, Marcel S. *LBP-TOP* based countermeasure against face spoofing attacks. In: Park JI, Kim J, editors. Computer Vision—ACCV 2012 Workshops. Vol. 7728. Berlin/Heidelberg, Germany: Springer; 2012.
12. Patel K, Han H, Jain AK. Secure face unlock: spoof detection on smartphones. IEEE Transact Inform Foren Secur. 2016;11(10):2268–83. doi:10.1109/TIFS.2016.2578288.
13. Boukanafet Z, Komulainen J, Hadid A. Face antispoofing using speeded-up robust features and fisher vector encoding. IEEE Signal Process Lett. 2017;24(2):141–5.
14. Yang J, Lei Z, Liao S, Li SZ. Face liveness detection with component dependent descriptor. In: 2013 International Conference on Biometrics (ICB); 2013 Jun 4–7; Madrid, Spain: IEEE. p. 1–6.
15. Yu Z, Zhao C, Wang Z, Qin Y, Su Z, Li X, et al. Searching central difference convolutional networks for face anti-spoofing. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA: IEEE; 2020. p. 5294–304.
16. Yu Z, Wan J, Qin Y, Li X, Li SZ, Zhao G. NAS-FAS: static-dynamic central difference network search for face anti-spoofing. IEEE Transact Pattern Anal Mach Intell. 2021;43(9):3005–23. doi:10.1109/TPAMI.2020.3036338.
17. Ming Z, Yu Z, Al-Ghadi M, Visani M, MuzzamilLuqman M, Burie JC. Vitranspad: video transformer using convolution and self-attention for face presentation attack detection. arXiv:2203.01562. 2022.
18. Liu Y, Jourabloo A, Liu X. Learning deep models for face antispoofing: binary or auxiliary supervision. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE; 2018. p. 389–98.
19. Chen B, Yang W, Wang S. Face anti-spoofing by fusing high and low frequency features for advanced generalization capability. In: 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR); 2020 Aug 6–8; Shenzhen, China: IEEE; 2020. p. 199–204.
20. Yu Z, Li X, Niu X, Shi J, Zhao G. Face antispoofing with human material perception. In: European Conference on Computer Vision (ECCV); 2020; Cham: Springer.
21. Yu Z, Peng W, Li X, Hong X, Zhao G. Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea: IEEE; 2019. p. 151–60.
22. Huang PK, Chiang CH, Chen TH, Chong JX, Liu TL, Hsu CT. One-class face anti-spoofing via spoof cue map-guided feature learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024; Piscataway: IEEE. p. 25–9.
23. Narayan K, Patel VM. Hyp-OC: hyperbolic one class classification for face anti-spoofing. arXiv:2404.14406. 2024.

24. Tsai CC, Wu TH, Lai SH. Multi-scale patch-based representation learning for image anomaly detection and segmentation. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2022 Jan 3–8; Waikoloa, HI, USA: IEEE; 2022. p. 3065–73.
25. Oza P, Patel VM. Active authentication using an autoencoder regularized cnn-based one-class classifier. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019); 2019 May 14–18; Lille, France: IEEE; 2019. p. 1–8.
26. Perera P, Patel VM. Dual-minimax probability machines for one-class mobile active authentication. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS); 2018 Oct 22–25; Redondo Beach, CA, USA: IEEE; 2018. p. 1–8.
27. Arashloo SR, Kittler J. An anomaly detection approach to face spoofing detection: a new formulation and evaluation protocol. In: 2017 IEEE International Joint Conference on Biometrics (IJCB); 2017 Oct 1–4; Denver, CO, USA: IEEE; 2017. p. 80–9.
28. Du X, Wang Z, Cai M, Li Y. Vos: learning what you don't know by virtual outlier synthesis. In: The Tenth International Conference on Learning Representations (ICLR); 2022; San Diego: OpenReview. p. 25–9.
29. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy: IEEE; 2017. p. 2999–3007.
30. Dubey A, Gupta O, Guo P, Raskar R, Farrell R, Naik N. Pairwise confusion for fine-grained visual classification. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018; Cham: Springer. p. 70–86.
31. Bouklenafet Z, Komulainen J, Li L, Feng X, Hadid A. OULU-NPU: a mobile face presentation attack database with real-world variations. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017); 2017 May 30–Jun 3; Washington, DC, USA: IEEE; 2017. p. 612–8.
32. Wen D, Han H, Jain AK. Face spoof detection with image distortion analysis. IEEE Transact Inform Foren Secur. 2015;10(4):746–61. doi:10.1109/TIFS.2015.2400395.
33. Chingovska I, Anjos A, Marcel S. On the effectiveness of local binary patterns in face antispoofing. In: International Conference on Biometrics (ICB); 2012; Piscataway: IEEE.
34. Zhang K, Zhang Z, Li Z, Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett. 2016;23(10):1499–503. doi:10.1109/LSP.2016.2603342.
35. Parkhi OM, Vedaldi A, Zisserman A. Deep face recognition. In: Proceedings of the British Machine Vision Conference 2015; 2015; Swansea: British Machine Vision Association. p. 1–12.
36. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al. Automatic differentiation in pytorch. 2017 [cited 2024 Dec 30]. Available from: https://openreview.net/forum?id=BJJsrmfCZ.
37. Schölkopf B, Williamson RC, Smola A, Shawe-Taylor J, Platt J. Support vector method for novelty detection. In: Advances in Neural Information Processing Systems (NIPS); 1999; Cambridge: MIT Press.
38. Tax DMJ, Duin RPW. Support vector data description. Mach Learn (ML). 2004;54(1):45–66. doi:10.1023/B:MACH.0000008084.60811.49.
39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
40. Fatemifar S, Awais M, Arashloo SR, Kittler J. Combining multiple one-class classifiers for anomaly based face spoofing attack detection. In: 2019 International Conference on Biometrics (ICB); 2019 Jun 4–7; Crete, Greece: IEEE. p. 1–7.