Anomaly Detection-Based Unknown Face Presentation Attack Detection

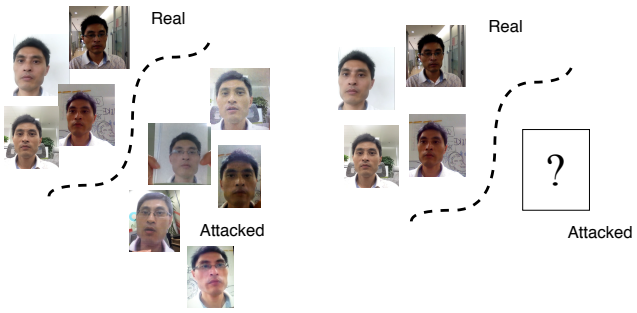
Yashasvi Baweja Poojan Oza Pramuditha Perera Vishal M. Patel
Department of Electrical and Computer Engineering
Johns Hopkins University, 3400 N. Charles St., Baltimore, MD 21218, USA
{ybaweja1,poza2,pperera3,vpatel36}@jhu.edu

Abstract

Anomaly detection-based spoof attack detection is a recent development in face Presentation Attack Detection (fPAD), where a spoof detector is learned using only non-attacked images of users. These detectors are of practical importance as they are shown to generalize well to new attack types. In this paper, we present a deep-learning solution for anomaly detection-based spoof attack detection where both classifier and feature representations are learned together end-to-end. First, we introduce a pseudo-negative class during training in the absence of attacked images. The pseudo-negative class is modeled using a Gaussian distribution whose mean is calculated by a weighted running mean. Secondly, we use pairwise confusion loss to further regularize the training process. The proposed approach benefits from the representation learning power of the CNNs and learns better features for fPAD task as shown in our ablation study. We perform extensive experiments on four publicly available datasets: Replay-Attack, Rose-Youtu, OULU-NPU and Spoof in Wild to show the effectiveness of the proposed approach over the previous methods. Code is available at: https://github.com/yashasvi97/IJCB2020_anomaly

1. Introduction

With the ubiquitous use of mobile phones and laptops, security of digital devices have attracted considerable interest in the research community. Recent advances in deep learning and facial recognition has prompted developers to use face and fingerprint-based authentication mechanisms in digital systems. Biometric-based authentication methods are vulnerable to carefully designed spoof attacks. Face presentation attack is one such spoofing technique against which many face recognition modules fail considerably [39, 38, 22]. In a face presentation attack, attacker(s) tries to fool the biometric system by presenting a face picture (or a video) of the enrolled user with the intention of surpassing authentication. To defend against such attacks, the device needs to learn how to distinguish between attacked images from the bonafide presentation images. This problem is commonly referred as the face Presentation Attack Detection (fPAD).



(a) Two Class fPAD (b) One Class fPAD

Figure 1. Different formulations in fPAD. In (a) bonafide presentation images and attacked images of known attack types are available during training. A decision boundary is learned to distinguish between two types of images. (b) Only bonafide presentation images are available during training and a decision boundary containing all bonafide presentation data is learned.

Conventionally, fPAD has been formulated as a two class problem in the literature [16, 3, 20, 40]. Specifically, a binary classification model is trained using both attacked and bonafide presentation images (or other biometric cues) as shown in Figure 1(a). During inference, a query image is evaluated using the learned classification model. In this scenario, it is assumed that details of the attacking mechanism is known during training. In practice, attackers can potentially develop more sophisticated attacking mechanisms unknown to the device. Therefore, it is important that an fPAD mechanism is robust to attacks that are not seen during training. However, a study carried out by Arashaloo *et al*. [2] showed that binary classification-based fPAD methods do not generalize well to new attack types. Arashaloo *et al*. [2] further showed that models with better generalization can be obtained by training a method only on bonafide pre-

978-1-7281-9186-7/20/$31.00 ©2020 IEEE

sentation data (as done in anomaly detection) where known attack images are only used for model evaluation. To this end, [2] introduced a new branch of research in fPAD called *anomaly detection-based fPAD* where spoof detectors are learned only based on bonafide presentation images of the enrolled users, as illustrated in Figure 1(b).

Existing methods that address fPAD as an anomaly detection problem utilize off-the-shelf one class classification algorithms such as one class SVM (OC-SVM) [35], Support Vector Data Descriptor (SVDD) [37], and one class Gaussian Mixture Models (OC-GMM) [23]. These classifiers are learned on hand-crafted and/or deep features extracted from bonafide presentation images [23, 2]. Fatemifar *et al*. [13] utilized an ensemble of multiple one class classifiers to deal with this problem. Recently, Fatemifar *et al*. [12] proposed to include user id information to improve the fPAD performance. Note that these approaches either use hand-crafted features or extract features from a pre-trained deep network. These features are not learned directly from the training data. Furthermore, none of these methods provide a standalone fPAD model that is end-to-end trainable.

In this paper, we propose an fPAD model that can be trained in an end-to-end fashion. Specifically, we utilize the representation learning power of a Deep Convolutional Neural Networks (CNN) to learn better representation from given bonafide presentation image data and simultaneously learn classification boundary to enclose the learned representations. Since only one class data (bonafide presentation images of the users) is available during training for the given problem, CNN training becomes a challenging task. To overcome this challenge, we propose a novel training strategy for CNNs by introducing pseudo-negative class samples in the feature space which help the network learn a better decision boundary for fPAD. In addition, we use a Pairwise Confusion (PC) loss [11] to further regularize the fPAD network during training.

This paper makes the following contributions:

1. An end-to-end deep learning solution is presented for fPAD based on anomaly detection.
2. We propose training deep CNNs with one class data for fPAD with the help of pseudo-negative sampling in the feature space.
3. We perform extensive experiments on four publicly available datasets: Replay-Attack, Rose-Youtu, OULU-NPU and Spoof in Wild and show that the proposed approach is able to perform better compared to the existing fPAD methods.

2. Related Work

Traditional two class based fPAD. Traditional algorithms in fPAD require a training dataset with labeled bonafide presentation and attacked images [20, 5, 3]. Usually a binary classifier model is trained on this data and in many cases with the help of additional biometric cues [3, 20], and finally the model predicts a test image as bonafide presentation or attacked based on the learned binary classification model. For example, Atoum *et al*. [3] introduced a novel end-to-end fPAD model which fuses scores from two deep CNNs – one from depth-based and another from image patch-based. In some of the recent methods, Liu *et al*. in [20] trained a binary classifier for fPAD using other modalities, like rPPG signal, depth map, optical flow *etc*. George *et al*. [16] used a multi channel information (like depth, Infra-Red, *etc*.) across same feature extraction network and classify the image as bonafide presentation or attacked based on the concatenated output feature.

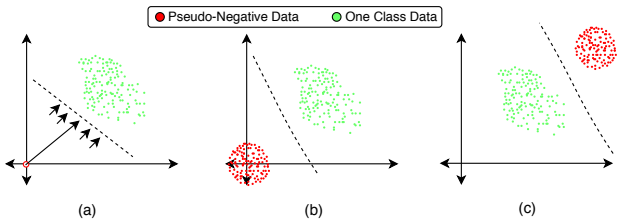


Figure 2. Comparison of one class classification methods. (a) One Class SVM [35]. (b) OC-CNN [25]. (c) Proposed method.

Several other studies use temporal information of videos to detect whether a video is attacked or not. Liu *et al*. [18] use rPPG signal to differentiate between bonafide presentation and 3D mask attacked videos. Liu *et al*. [19] further extended the approach and used Near Infrared (NIR) to work with remote photoplethysmography correspondence features for detecting 3D mask attacked videos. Several works have used texture-based analysis for detecting attacked images. Boulkenafet *et al*. [5] deploy Local Binary Patterns (LBP) [1] features calculated for all RGB dimensions and use the concatenated LBP to classify it as bonafide presentation or attacked image. Boulkenafet *et al*. improved this method later [6] using SIFT [21] and SURF [4] features in both RGB and YCbCr space. In, [14] authors explored the possibility of using Image Quality Measure (IQM) features to perform fPAD.

Anomaly Detection-based fPAD. Arashloo *et al*. [2] argued that such two-class approaches can be biased towards attacked training images and might limit the generalization ability of the model to novel attacks. Hence, [2] provided a novel formulation for fPAD based on anomaly detection where models are learned using only bonafide presentation images. Following this formulation, Arashloo *et al*. [2] proposed a one class SVM [35] based method for fPAD. Nikisins *et al*. [23] proposed another approach by modeling Image Quality Measure (IQM) features [15] extracted from the bonafide presentation images with one class GMM and showed superiority over the one class SVM classifier. Fatemifar *et al*. [12] showed that when client ID informa-

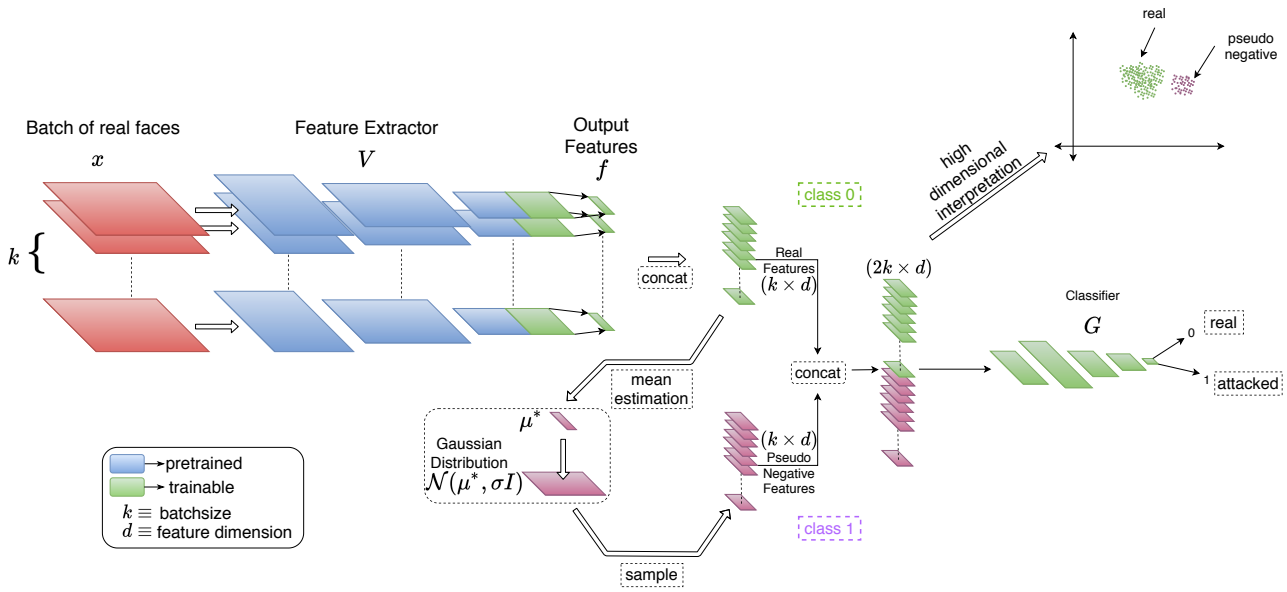


Figure 3. Training algorithm of the proposed method (best viewed in color). A batch of bonafide presentation images (red) is fed into the face extractor V, which in turn produces a batch of bonafide presentation features (green) of dimension $(k \times d)$. A pseudo-negative feature batch (purple) $k \times d$ is sampled from a Gaussian distribution whose mean is calculated considering a running mean of bonafide presentation features. The concatenated batch of bonafide presentation and pseudo-negative features is fed into the classifier. The output of the classifier G produces the posterior probability of bonafide presentation and attacked classes.

tion is available, it can be leveraged to further boost the fPAD performance. Specifically, they train client-specific PAD models that are based on one class SVM, one class GMM and Mahalanobis distance. Fatemifar *et al*. [13] explored the use of features from specific face regions such as eyes, mouth, nose etc. which might be helpful in improving the fPAD performance. Another approach by Fatemifar *et al*. [13] utilized ensembles of one class classifiers. All of these methods use hand-crafted features or features from a pre-trained deep network and learn an off-the-shelf one class classifier. None of these methods have utilized deep networks in the training process to leverage their representation learning power.

One Class Methods Oza *et al*. [25] introduce new one class algorithm where a pseudo negative class centered at origin is used to train the one class classifier. Authors extended the same one class approach for active authentication in [26]. Perera *et al*. [31] also deploy the one class classification method for active authentication. Authors in [33] learn sophisticated deep features focused for one class classification. Open set algorithms also aim to detect anomalies or out of distribution data: Perera *et al*. [32] introduce membership loss for novelty classification. Authors in [24, 27] use data distribution shift and patch level information respectively for multiple class novelty detection. Authors in [30] learn discriminative deep features for open set recognition. Zhang *et al*. [41] induce a sparsity constraint on features for open set classification. Further, Shao *et al*. [36] defend against open set adversarial attacks.

3. Proposed Method

For one class problems, the absence of negative class data makes it difficult to train a deep network in an end-to-end fashion. Hence, most work follow a methodology where deep features are extracted using a pre-trained network and a one class classifier such as one class SVM (OC-SVM) is trained for classification. To enable end-to-end learning for deep networks in one class setting, Oza *et al*. [25] proposed a method named one class convolutional neural network (OC-CNN) which attempts to mimic the idea behind OC-SVM where a separating hyperplane is learned to separate one class features from the origin. Specifically, they use samples from a Gaussian distribution centered at the origin with a small standard deviation as the pseudo-negative class. Inspired from this work, we utilize similar framework for training the deep networks. However, the problem setting considered in [25] considers anomaly data that are visually distinct as compared to the training class data we have in fPAD. Hence, the concept of separating training data in the feature space from a zero centered Gaussian might work well for their problem setting. Whereas, in the case of fPAD the attacked samples during testing are very similar to the non-attacked samples used during training. As a result there exists only a subtle visual difference between the two classes (bonafide presentation and attacked images of the same person). Hence, im-

ages from both classes can be in a very close proximity to each other in the feature space. To deal with this issue, we propose an adaptive mean estimation strategy to generate pseudo-negative data for training. The goal of this adaptive strategy is to create the pseudo-negative Gaussian distribution such that it lies in the close proximity to the features from the bonafide presentation class data. The differences among OC-SVM, OC-CNN and the proposed method are illustrated in Figure 2. In the following sub-sections we discuss the proposed method in detail. The proposed method consists of two main parts, a feature extractor network (V) and a classifier (G).

3.1. Feature Extractor V

Let $\{x_i\}_{i=1}^{N}$ be a set of N training images consisting of bonafide presentation (i.e. non-attacked) images. These images are passed through a face recognition network V which produces a set of d-dimensional features $\{f_i\}_{i=1}^{N}$ such that $f_i = V(x_i) \in \mathbb{R}^d$. These features are then fed into the classifier, G, as described in the following Section.

3.2. Classifier G

As discussed earlier, we utilize an adaptive strategy to estimate the mean of pseudo-negative Gaussian distribution. The adaptive mean estimation makes sure that samples from pseudo-negative Gaussian distribution lies in the proximity of bonafide presentation class features. More formally, we define a pseudo-negative distribution, $\mathcal{N}(\mu^*, \sigma I)$ where $\mathcal{N}$ is a Gaussian distribution whose mean is μ^* and covariance matrix is $\sigma I (I \in \mathbb{R}^d)$. We consider displacement of sample feature mean of bonafide presentation data across two iterations when defining μ^*. More precisely, let us consider a batch (B) that contains features extracted from the non-attacked images using network V, denoted as $f^B = \{f_i\}, \forall i \in B$. Let μ_{new} be the mean of feature vectors f^B and μ_{old} be the mean of features of the previous batch. Both μ_{new} and μ_{old} are d-dimensional. For training a classifier G a batch B' of attacked features $f^{B'} = \{f_j\}, \forall j \in B'$ are sampled from the pseudo-negative class, whose center μ^* is calculated as

$$\mu^* = \alpha \mu_{old} + (1 - \alpha)\mu_{new}, \tag{1}$$

where α is a hyper-parameter controlling the influence of old mean in the estimation of μ^* and $|B'| = |B| = k$. During the first iteration μ_{new} is used to calculate μ^*. The data $f = \left(f^B, f^{B'}\right)$ is concatenated across the batch dimension and fed into the classifier G. The corresponding label vector, is of size $2k \times 1$, where the first k elements are all zeros followed by all ones denoting bonafide presentation and attacked class, respectively. This process is illustrated in Fig. 3. The classifier G uses these features and produces probability vector p of size $2k \times 1$. The network can be trained using the cross-entropy loss defined as follows

$$\ell_{ce} = -\sum_{i=1}^{2k} \{y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\}, \tag{2}$$

where, y_i denotes the label (0 or 1) of the i^{th} input to the classifier G and p_i denotes the probability of the i^{th} input being from the pseudo-negative class. Similarly, $1 - p_i$ denotes the probability of the i^{th} input to the classifier G being extracted from the non-attacked class data. Our classifier G consists of 3 fully-connected layers of 8192, 1000, 500 neurons, respectively. The dimension of the input layer is the same as the output layer of V. The final output has two neurons, one for the probability corresponding to the bonafide presentation class and other for the attacked class.

As in most cases the network V is pre-trained on a face recognition dataset and in this case it's VGGFace [28]. Although pre-trained weights serve as a good initialization for training the initial layers of V, the high level features (which will be used as input to G) are not suitable for fPAD. The final layer features of a pre-trained network are more suitable for face recognition. Furthermore, feature representations are usually clustered by identities. Eliminating the identity information from the features helps in improving the fPAD performance as we are only interested in spoof detection irrespective of the identity information. To this end, we utilize the following Pairwise Confusion (PC) loss [11]

$$\ell_{pc} = \sum_{i} \sum_{j \neq i} \|f_i - f_j\|_2^2, \tag{3}$$

where f_i is the feature vector corresponding to image x_i. This loss induces a Euclidean confusion in the pre-trained feature space and tries to remove the identity information from it. Disrupting the feature space in this manner results in better feature representation for fPAD. Note that, this loss is only calculated using the bonafide presentation features and not from the pesudo-negative Gaussian samples.

The overall loss function used to train the network is a combination of both loss functions and is defined as follows

$$\ell_{final} = \ell_{ce} + \lambda \ell_{pc}, \tag{4}$$

where λ is a constant. This loss function is used to train the spoof detection network in an *end-to-end* manner.

4. Datasets and Evaluation Protocols

We demonstrate the effectiveness of the proposed model for fPAD on four datasets: Replay Attack [9], Rose-youtu [17], OULU-NPU [7] and Spoof in Wild [20]. Sample images from each dataset are shown in Figure 4.

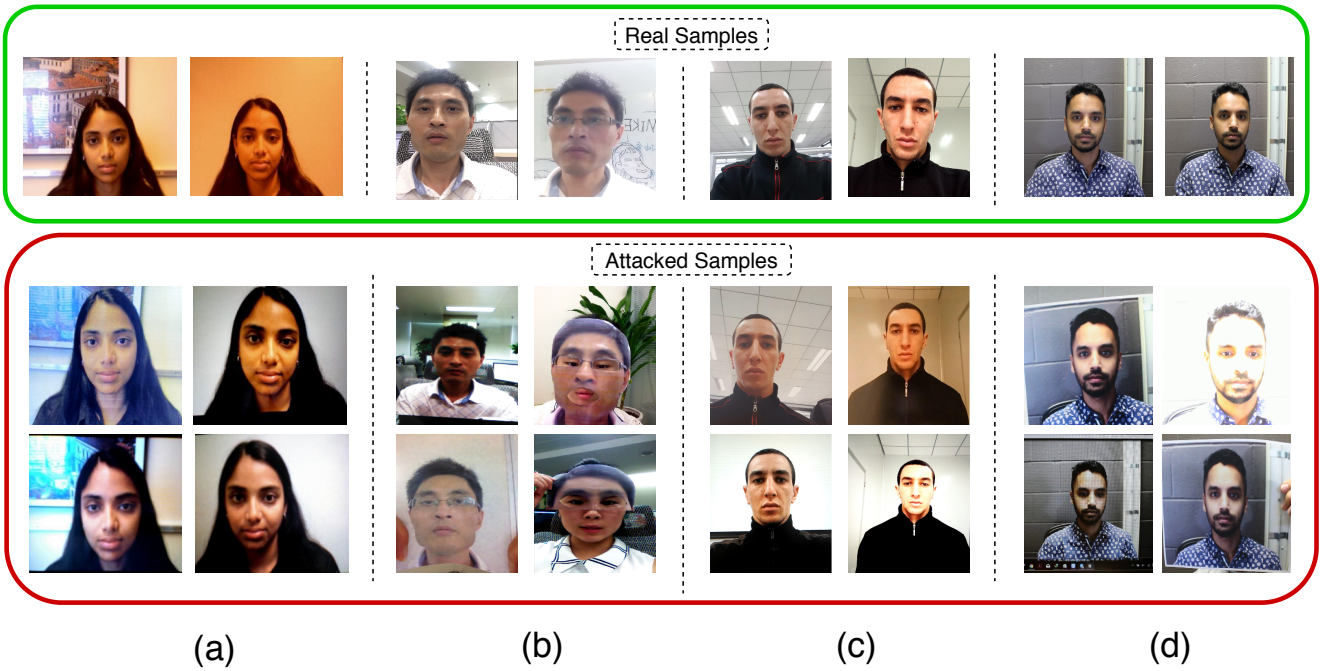


Figure 4. Sample images from each dataset. (a) Replay Attack [9], (b) Rose-Youtu [17] (c) Oulu-NPU [7] and (d)Spoof in Wild [20]. Green part represents the bonafide presentation samples and red corresponds to the attacked samples of same identity.

4.1. Replay Attack

The Replay Attack dataset [9] is an fPAD dataset of 1300 videos consisting of photo and replay attacks. It contains 50 identities and four sets namely, train, development, test and enroll. The train, development and test sets contain both bonafide and attacked presentation data of 15, 15 and 20 identities, respectively. We extract 30-40 frames per video with a gap of at least 10 frames in between to carry out our experiments. HOG-based face detection [10] algorithm is then used to find the face. We train our model and the baselines on only bonafide presentation data from the training set and test on the bonafide presentation and all attacked data from the test set.

4.2. Rose-Youtu

The Rose-youtu dataset [17] is an fPAD database which covers a large variety of illumination conditions and attack types. There are in total 3350 frontal face videos of 20 subjects captured using 5 mobile phones: Hasee smart-phone, Huawei smart-phone, iPad 4, iPhone 5s, and ZTE smartphone. On average each subject has around 25-50 bonafide presentation videos. There are three types of attacks in the dataset: print attack, replay video attack and mask attack. Every 15th frame from the video under consideration is sampled and then corresponding face is extracted using the HOG-based face detector [10]. We use 10 train identities' bonafide presentation data for training the baselines and the proposed method. Testing is done on the data from the remaining 10 identities using all attacks.

4.3. OULU-NPU

The OULU-NPU dataset [7] consists of mobile attacked videos. There are in total 4950 videos, 55 identities recorded using six mobile devices in different sessions. There are mainly two types of attacks in the dataset – print and replay attack. The print attacks are done using two types of printers and video replays are done using two display devices. We consider all attacks in the test set. For training, only bonafide presentation data corresponding to the identities in training set are used. For extracting faces from videos, we sample every 10th frame from each video and then use the face locations given in the dataset. There are 20 and 15 different identities in the training set and the test set, respectively.

4.4. Spoof in Wild

The Spoof in Wild dataset [20] is a collection of 4478 videos distributed among 165 identities. All videos are captured in minimum 1080p resolution. Each subject has 8 bonafide presentation and around 20 attacked videos. There are two types of attacks in the dataset – print attack based on glossy and matt paper, and replay attack created using 4 different devices. All attacks are considered during testing. Faces are extracted from each video by sampling at every 10th frame. We use the locations provided in the dataset to extract faces which are used to carry out experiments. We use only bonafide presentation data from the training set for

training the baselines and the proposed model. Data from 90 subjects are used during training and the remaining data from 75 different users are used during testing.

We evaluate the performance of anomaly detection-based fPAD methods on two protocols. The details of these protocols are described as follows:

• **Protocol 1:** In this protocol models are trained using the identities in the training set with only bonafide presentation image data and evaluated on the identities of the test set having both bonafide presentation and attacked images. Both sets contain non-overlapping identities.

• **Protocol 2:** For this protocol, enrollment split is used for model training and evaluation [12]. fPAD models are trained using a training set and evaluated on the test set. Both training and test set are derived from the enrollment split of the datasets and both have bonafide presentation and attacked images of same identities. Unlike [13], we do not utilize any identity labels and hence for fair comparison consider only class-independent models of [13].

5. Implementation Details

In all datasets we only use the bonafide presentation data from the training set without using any identity label information for training the model. The test data consists of both bonafide and attack presentation data. For datasets without face locations, we use HOG-based face detection mechanism [10]. The value for λ is set to 3. The value of α is set equal to 0.8. The standard deviation σ of the pseudo-negative Gaussian is set as 1. VGGFace [28] is used as the base feature extraction module. The layer just after the convolution layers, that is 'fc6'(4096 dimension) of VGGFace is used to extract features. Thus output of the 'fc6' layer is used as the input to the classifier G. The last two convolution layers, 'fc6' of V along with all the parameters of G are trained using the loss equation (4). The model is trained for 100 epochs with 1e-4 learning rate, batch size of 80 and results are reported in Section 6.

6. Results and Analysis

In this section, we discuss results on the aforementioned datasets and present an in depth analysis of the proposed method. We use Average Classification Error Rate (ACER) as the metric to compare the performance of different methods. ACER is the average of the Attack Presentation Classification Error Rate (APCER) and Bonafide Presentation Classification Error Rate (BPCER) at a particular threshold [34]. The ACER is defined as follows

$$ACER = \frac{APCER + BPCER}{2}\%.$$

In our comparisons we aim to find the threshold for which the average of APCER and BPCER is least. ACER is calculated for each identity in the test set and the average of ACER is reported in Table 1.

6.1. Comparison with Baseline One Class classifiers

For our baseline comparisons we consider the following five methods:

1. **OC-SVM** - One class SVM (OC-SVM) [35] maps the one class data into a feature space and tries to learn a hyperplane between the training data and origin. Implementation from sklearn library [29] is used with default parameters and ν=0.1. Signed distance from the hyperplane is used as an anomaly score.

2. **SVDD** - Support Vector Data Description (SVDD) [37] is an extension to the one class SVM where it tries to learn a hypershpere around the training data as opposed to a hyperplane separating the origin. Optimization is done to have minimum spherical radius that can contain all data. Default implementation of LibSVM library [8] is used with parameter ν=0.1. Distance from the center is used as the anomaly score.

3. **MD** - Mahalanobis Distance (MD) assumes data is coming from a Gaussian distribution with single mode and computes the distance based on the parameters of the Gaussian model. The bonafide presentation data is used to estimate the parameters of Gaussian model. The distance calculated based on the estimated mean and variance of Gaussian model is used as the anomaly score. Higher distance increases the likelihood of a test image being an attacked image. MD is implemented using the Sklearn library.

4. **OC-GMM** - A Gaussian Mixture Model (GMM) is learned using bonafide presentation class data. Parameters of the GMM are estimated using the Sklearn library. For consistency across all datasets, the number of components are set equal to 50. Log likelihood of a test sample belonging to GMM is used as the anomaly score for calculating ACER.

5. **OC-CNN** - One Class CNN [25] trains a classifier using one class data and pseudo-negative Gaussian distribution centered at the origin. The probability that a test sample is a bonafide presentation is used as the anomaly score.

For the proposed algorithm, we use the probability for class 0, *i.e.*, being a bonafide presentation as the anomaly score to calculate ACER.

6.2. Experiments using Protocol 1

As can be seen from Table 1 the proposed algorithm (last column) outperforms all the other methods. Among

Dataset	OC-SVM ACER (APCER, BPCER)	SVDD ACER (APCER, BPCER)	MD ACER (APCER, BPCER)	OC-GMM ACER (APCER, BPCER)	OC-CNN ACER (APCER, BPCER)	Proposed ACER (APCER, BPCER)
Replay Attack [9]	31.142 (43.441, 18.843)	32.961 (46.337, 19.586)	31.747 (32.392, 31.102)	30.096 (29.922, 30.117)	35.985 (32.508, 39.462)	**20.739** (25.047, 16.539)
Rose Youtu [17]	47.165 (80.877, 13.453)	46.434 (80.599, 12.465)	38.993 (50.083, 27.903)	46.771 (69.216, 24.325)	35.808 (43.556, 28.059)	**31.623** (35.573, 27.673)
Oulu NPU [7]	47.561 (81.815,13.307)	47.517 (65.739, 29.296)	45.185 (72.821, 17.549)	46.957 (75.353, 18.562)	45.799 (54.488, 37.110)	**30.242** (38.632, 21.852)
Spoof In Wild [20]	47.165 (88.524, 5.806)	47.051 (90.277, 3.826)	42.970 (69.997, 15.943)	44.250 (76.755, 11.744)	36.994 (37.454,36.534)	**23.335** (23.482, 23.187)

Table 1. Comparison of the proposed algorithm with baseline one class classifiers for Protocol 1. The values in bold represent the lowest ACER(%) with APCER(%) and BPCER(%) in brackets.

the baseline methods, one class CNN performs the best as it helps capturing the multi-modal nature of the distribution of bonafide presentation features via deep representation learning process. However, the proposed approach outperforms one class CNN significantly, providing overall improvement of $\sim$4% across all datasets. This performance improvement can be attributed to the novel training strategy of separating the bonafide presentation features from Gaussian sampled features, which are in close proximity and helps in learning a good decision boundary. Furthermore, training a deep neural network using the given data for spoof detection helps in learning more rich features rather than using hand-crafted or pre-trained features.

6.3. Experiments using Protocol 2

The results of experiments on protocol 2 are summarized in Fig. 5. Specifically, for the Replay Attack dataset [9] the best identity independent one class classifier is one class GMM, which achieves ACER of 9.96% [13]. Whereas the proposed method achieves ACER of 7.334 % providing $\sim$2% improvement. For the Rose-Youtu dataset, one class GMM achieves ACER of 17.54% [13], whereas the proposed algorithm achieves ACER of 10.354%, resulting in an improvement of $\sim$7%. Additionally, we performed protocol 2 experiments on OULU-NPU and Spoof in Wild and the best performing classifier OC-GMM achieved ACER of 13.927% and 23.383%, respectively. In comparison, the proposed approach achieved ACER of 12.473% and 21.557% for OULU-NPU and Spoof in Wild, providing $\sim$1% and $\sim$2% improvements respectively.

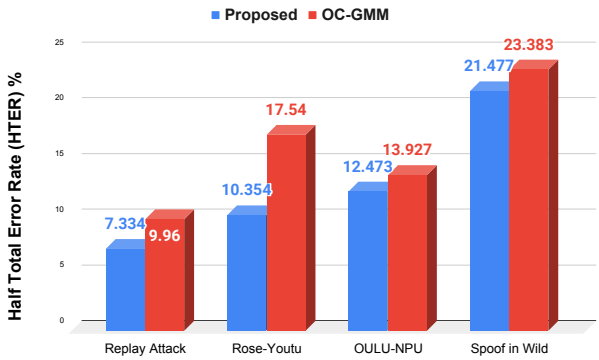


Figure 5. Comparison of the best baseline (OC-GMM) and the proposed method for Protocol 2 in terms of ACER.

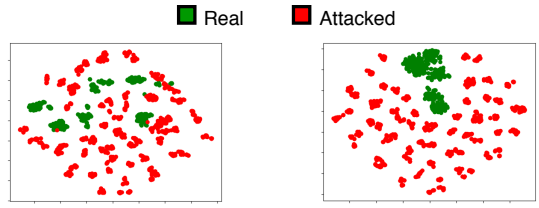
The results in Table 1 and Figure 5 show that the proposed approach outperforms other one class classifiers in both Protocol 1 and Protocol 2.

6.4. Comparison with OC-CNN

The method proposed in [25], introduced a pseudo-negative class to train the one class classifier. The method works well for their problem setup of anomaly detection, where training and test data have categorical and visual differences. But considering the fPAD-based anomaly detecting problem formulation, the face images at training and test time have very subtle visual differences. In some cases, these images may belong to the same identity and therefore it will not be optimal to separate bonafide presentation features from $\mathcal{N}(0,1)$ noise. From the results in Table 1, when the model is trained with a Gaussian centered at 0, ACER value is 35.985% for the Replay Attack Dataset. Instead when trained with a Gaussian whose center is chosen by the running mean, i.e., the proposed method, we see a significant improvement performance, achieving ACER of 20.739 %. Therefore, modeling the attacked class in the proximity to the bonafide presentation features helps in building a more sophisticated one class classifier for fPAD.

6.5. Ablation study

Effect of Pairwise Confusion loss: In this experiment we study the effect of pair-wise confusion loss [11] on the performance of the proposed model. We show that when the model is trained without the pair-wise loss, the performance is lower as compared to when PC loss is used. Here, we consider Replay-Attack dataset for the experiments. When the proposed model is trained with the PC loss, the performance is 20.739% ACER. The t-SNE feature visualization for this experiment is also provided in Fig. 6(b). On the other hand, when we train the proposed method without the pair-wise confusion loss, the ACER increases to 32.860%.



(a) After Training- without PC loss

(b) After Training - with PC loss

Figure 6. t-SNE plots for data corresponding to identities 1, 2, and 4 of the Replay-Attack dataset combined. The points in green represent the bonafide presentation features, and their attacked versions are marked in red. Figure (a) shows the t-SNE plot of the feature space when extracted using a trained network without the PC Loss. It can be seen that all the bonafide presentation features are cluttered in the feature space and are closer to their attacked counterparts. Figure (b) on the other hand shows the same feature space, but now the bonafide presentation features are all packed into a single cluster which are more suited for one class seting.

The corresponding t-SNE feature visualization for this experiment is given in Fig. 6(a). When we compare the t-SNE visualization of the features provided in Fig. 6, we can see that the best the model trained with pair-wise confusion loss achieves the feature embedding most suited for the anomaly detection-based fPAD model and thereby resulting in the improved performance among all other baselines.

7. Conclusion

In this paper we tackled the problem of face presentation attack detection based on anomaly detection. We point out the shortcomings of the existing work that address this issue and propose a novel method to alleviate these shortcomings. More precisely, we proposed a new end-to-end trainable, face presentation attack detection model based on deep convolutional neural network that can be trained with data from only one class. The proposed method utilizes a novel training procedure for one class neural network with the help of pseudo-negative sampling in the feature space. To the best of our knowledge, this is the first approach to include deep networks in the learning process and leverage their power of representation learning in an end-to-end manner for anomaly-based face presentation attack detection task. Extensive experiments on four publicly available datasets show that the proposed approach outperforms all existing approaches for the task. Furthermore, an ablation study was provided to show the significance of individual components of the proposed method.

Acknowledgment

This work was supported by the NSF grant 1801435.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006. 2

[2] S. R. Arashloo, J. Kittler, and W. Christmas. An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. *IEEE Access*, 5:13868–13882, 2017. 1, 2

[3] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu. Face anti-spoofing using patch and depth-based cnns. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 319–328. IEEE, 2017. 1, 2

[4] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 2

[5] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face anti-spoofing based on color texture analysis. In *2015 IEEE international conference on image processing (ICIP)*, pages 2636–2640. IEEE, 2015. 2

[6] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face anti-spoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters*, 24(2):141–145, 2016. 2

[7] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. May 2017. 4, 5, 7

[8] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 6

[9] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. 2012. 4, 5, 7

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005. 5, 6

[11] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik. Pairwise confusion for fine-grained visual classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–86, 2018. 2, 4, 7

[12] S. Fatemifar, S. R. Arashloo, M. Awais, and J. Kittler. Spoofing attack detection by anomaly detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8464–8468. IEEE, 2019. 2, 6

[13] S. Fatemifar, M. Awais, S. R. Arashloo, and J. Kittler. Combining multiple one-class classifiers for anomaly based face spoofing attack detection. In *International Conference on Biometrics (ICB)*, 2019. 2, 3, 6, 7

[14] J. Galbally and S. Marcel. Face anti-spoofing based on general image quality assessment. In *2014 22nd International Conference on Pattern Recognition*, pages 1173–1178. IEEE, 2014. 2

[15] J. Galbally, S. Marcel, and J. Fierrez. Image quality assessment for fake biometric detection: Application to iris, fin-

gerprint, and face recognition. *IEEE transactions on image processing*, 23(2):710–724, 2013. 2

[16] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 15:42–55, 2019. 1, 2

[17] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7):1794–1809, 2018. 4, 5, 7

[18] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In *European Conference on Computer Vision*, pages 85–100. Springer, 2016. 2

[19] S.-Q. Liu, X. Lan, and P. C. Yuen. Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 558–573, 2018. 2

[20] Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 389–398, 2018. 1, 2, 4, 5, 7

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[22] S. Marcel, M. S. Nixon, and S. Z. Li. *Handbook of biometric anti-spoofing*, volume 1. Springer, 2014. 1

[23] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel. On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. In *2018 International Conference on Biometrics (ICB)*, pages 75–81. IEEE, 2018. 2

[24] P. Oza, H. Nguyen, and V. M. Patel. Multiple class novelty detection under data distribution shift. In *Proceedings of the European Conference of Computer Vision (ECCV), Glasgow, Scotland*, 2020. 3

[25] P. Oza and V. M. Patel. One-class convolutional neural network. *IEEE Signal Processing Letters*, 26(2):277–281, 2018. 2, 3, 6, 7

[26] P. Oza and V. M. Patel. Active authentication using an autoencoder regularized cnn-based one-class classifier. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019. 3

[27] P. Oza and V. M. Patel. Utilizing patch-level activity patterns for multiple class novelty detection. In *Proceedings of the European Conference of Computer Vision (ECCV), Glasgow, Scotland*, 2020. 3

[28] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. 2015. 4, 6

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 6

[30] P. Perera, V. I. Morariu, R. Jain, V. Manjunatha, C. Wigington, V. Ordonez, and V. M. Patel. Generative-discriminative feature representations for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11814–11823, 2020. 3

[31] P. Perera and V. M. Patel. Dual-minimax probability machines for one-class mobile active authentication. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2018. 3

[32] P. Perera and V. M. Patel. Deep transfer learning for multiple class novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11544–11552, 2019. 3

[33] P. Perera and V. M. Patel. Learning deep features for one-class classification. *IEEE Transactions on Image Processing*, 28(11):5450–5463, 2019. 3

[34] R. Ramachandra and C. Busch. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 50(1):1–37, 2017. 6

[35] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001. 2, 6

[36] R. Shao, P. Perera, P. Yuen, and V. M. Patel. Open-set adversarial defense. In *Proceedings of the European Conference of Computer Vision (ECCV), Glasgow, Scotland*, 2020. 3

[37] D. M. Tax and R. P. Duin. Support vector data description. *Machine learning*, 54(1):45–66, 2004. 2, 6

[38] wired.com. Hackers say broke face id security. *www.wired.com/story/hackers-say-broke-face-id-security/*. 1

[39] wired.com. Hackers trick facial recognition software. *www.wired.com/2016/08/hackers-trick-facial-recognition-logins-photos-facebook-thanks-zuck/*. 1

[40] F. Xiong and W. AbdAlmageed. Unknown presentation attack detection with face rgb images. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9. IEEE, 2018. 1

[41] H. Zhang and V. M. Patel. Sparse representation-based open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1690–1696, 2016. 3