This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Deep Anomaly Detection for Generalized Face Anti-Spoofing

Daniel Pérez-Cabo
Gradiant - UVigo, Spain
dpcabo@gradiant.org

David Jiménez-Cabello
Gradiant, Spain
djcabello@gradiant.org

Artur Costa-Pazo
Gradiant, Spain
acosta@gradiant.org

Roberto J. López-Sastre
University of Alcalá, Spain
robertoj.lopez@uah.es

Abstract

Face recognition has achieved unprecedented results, surpassing human capabilities in certain scenarios. However, these automatic solutions are not ready for production because they can be easily fooled by simple identity impersonation attacks. And although much effort has been devoted to develop face anti-spoofing models, their generalization capacity still remains a challenge in real scenarios. In this paper, we introduce a novel approach that reformulates the Generalized Presentation Attack Detection (GPAD) problem from an anomaly detection perspective. Technically, a deep metric learning model is proposed, where a triplet focal loss is used as a regularization for a novel loss coined "metric-softmax", which is in charge of guiding the learning process towards more discriminative feature representations in an embedding space. Finally, we demonstrate the benefits of our deep anomaly detection architecture, by introducing a few-shot a posteriori probability estimation that does not need any classifier to be trained on the learned features. We conduct extensive experiments using the GRAD-GPAD framework that provides the largest aggregated dataset for face GPAD. Results confirm that our approach is able to outperform all the state-of-the-art methods by a considerable margin.

1. Introduction

Whether we like it or not, we are in the era of face recognition automatic systems. These solutions are now beginning to be used intensively in: border controls, on-boarding processes, accesses to events, automatic login, or to unlock our mobile devices. As an example of this last technology, we have the *Intelligent Scan*[1] that comes with Samsung mobiles, or the *FaceID*[2] for iPhones. All these systems are highly valued by consumers because of their usability and its non-intrusive nature. However, there remains one major challenge for all of them, Presentation Attacks (PA).

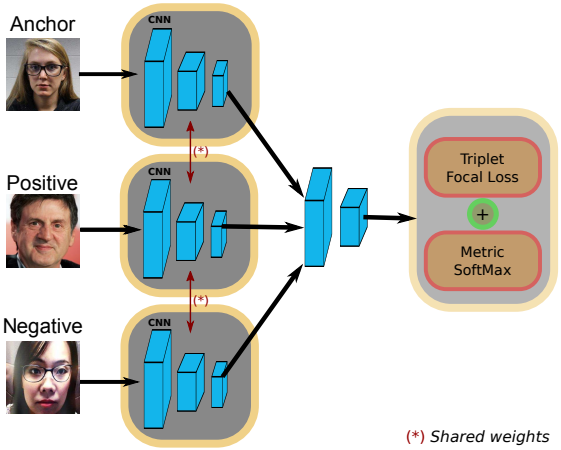

Figure 1: We propose a deep metric learning approach, using a set of Siamese CNNs, in conjunction with the combination of a triplet focal loss and a novel "metric softmax" loss. The latter accumulates the probability distribution of each pair within the triplet. Our aim is to learn a feature representation that allows us to detect impostor samples as anomalies.

These commercial systems rely on specialized hardware such as 3D/IR/thermal cameras entailing a far easier option to detect presentation attacks. Besides, this situation restricts the use case to a few specialized devices, incrementing costs dramatically. For the sake of accessibility and costs, we focus on the ubiquitous 2D-camera case, available in almost all mobile devices and easy to acquire and integrate on different checkpoints.

[1] https://www.samsung.com/my/support/mobile-devices/what-is-intelligent-scan-and-how-to-use-it/

[2] https://www.apple.com/lae/iphone-xs/face-id/

Although face recognition technologies achieve accuracy ratios above human performance in certain scenarios, consumers should be aware that they also introduce two new challenges that compromise their security: the Presentation Attack Detection (PAD) and the generalization capability of these solutions. With respect to the former, for example, a face recognition system with an outstanding 99.9% of accuracy fails simply by presenting a page with your face printed on it. These presentation attacks stand as a major threat for identity impersonation where illegitimate users attempt to gain access to a system using different strategies, *e.g*. video replay, make-up. Note that it is really easy to obtain audio-visual material from almost every potential user (*e.g*. Facebook photos, videos on YouTube, *etc*.), which allows the creation of tools to perform these PAs.

But the generalization problem is also relevant. In a nutshell, the scientific community has failed to provide an efficient method to detect identity impersonation based on face biometrics that is valid for real-world applications. Normally, the state-of-the-art models suffers a severe drop of performance in realistic scenarios, because they exhibit a sort of overfitting behaviour maximizing the results for just the dataset they have been trained on.

In this paper we explicitly address these two challenges. First, we introduce a deep metric learning based approach to deal with the PAD problem. As it is shown in Fig. 1, our solution is trained to learn a feature representation that guarantees a reasonable separability between genuine and impostor samples. Then, the generalization problem is tackled from an anomaly detection approach, where we expect to detect the attacks as if they were out-of-distributions samples that naturally exhibit a higher distance in the embedding space with respect to the real samples in the dataset.

The generalization capability of our solution and its state-of-the-art competitors is thoroughly evaluated using the recent GRAD-GPAD framework [11]. We use the aggregated dataset provided in GRAD-GPAD, which comprises more than 10 different datasets for face anti-spoofing. This aspect results fundamental, because it allows us to deploy extensive inter-dataset experiments, to address the Generalized Presentation Attack Detection problem.

As a summary, in this paper we make the following contributions:

1. We introduce a novel anomaly detection strategy based on deep metric learning for face anti-spoofing using just still images.

2. Our model leverages the use a triplet focal loss as a regularizer of a novel "metric softmax" loss, to ensure that the learned features allow for a reasonable separability between real and attacks samples in an embedding space.

3. A thorough experimental evaluation on GRAD-GPAD shows that our anomaly detection based approach outperforms the state-of-the-art models.

4. Finally, we propose a novel few-shot a posteriori probability estimation that avoids the necessity of training any classifier for decision making.

The remainder of this paper is organized as follows. Section 2 reviews the main progress and challenges on the problem of generalization for anti-spoofing systems. We introduce our anomaly detection deep model in Section 3. Sections 4 and 5 provide the experimental evaluation and the conclusions, respectively.

2. Related Work

Face-PAD approaches can be categorized regarding the following standpoints: i) from the required user interaction as *active* [19] or *passive* [20, 33] methods; ii) from the hardware used for data acquisition as *rgb-only* [14, 25, 33], *rgb-infrared-depth* [3, 37] or *additional sensors* [30] approaches; iii) from the input data type as *video-based* [1, 28] or *single-frame* [33] approaches; iv) and, finally, depending on the feature type, from classical hand-crafted features [5, 33] to the newer ones based on automatic learned deep features [18, 20]. These deep models are precisely the responsible for a considerable increase in accuracy for face-PAD, defining the new state of the art.

However, recent studies reveal that the current approaches are not able to correctly generalize [21] using fair comparisons. Actually, the main difficulty for the inclusion of anti-spoofing systems in realistic environments is the Generalized Presentation Attack Detection (GPAD) problem. Some works [11, 12, 25] propose new evaluation protocols, datasets and methods to address the GPAD.

Overall, generalization has been addressed from different perspectives: i) applying *domain adaptation* techniques [21]; ii) learning *generalized deep features* [20, 21]; or even iii) using generative models [18]. All these methods are able to slightly mitigate the drop of performance when testing on new unseen scenarios, but they are still far from being suitable for real scenarios.

Traditional methods for face anti-spoofing use a two-class classifier to distinguish between real samples and attacks. Recently, some works suggest that formulating the problem of anti-spoofing as an anomaly detection approach could improve their generalization capacity [2, 25]. In [2], the authors assume that real-accesses share the same nature, in contrast to spoofing attempts that can be very diverse and unpredictable. They present a study to determine the influence of using only genuine data for training and compare it with traditional two-class classifiers. From the experimental results the paper concludes that: i) anomaly detection based systems are comparable to two classes based systems; and ii) neither of the two approaches perform well enough in the

evaluated datasets (CASIA-FASD [39], Replay-Attack [8] and MSU-MFSD [33]). On the other hand, the authors of [25] propose a more challenging experiment based on an aggregated dataset that comprises Replay-Attack, Replay-Mobile [10] and MSU-MFSD. They propose a GMM-based anomaly classifier which outperforms the best solutions reported in [2].

In this paper, we reformulate the anomaly detection scheme using a deep metric learning model for face-PAD that highly reduces the problem of generalization. Experiments are performed over the largest aggregated publicly available dataset, the GRAD-GPAD framework [11]. This framework allows us to reinforce the assumption that real access data shares the same nature, provided that the number of identities is large and the capture conditions and devices are diverse enough; that is, the genuine class is well represented by data. Additionally, the highly representative embeddings obtained using the proposed metric learning approach permits outperforming prior works, distinguishing genuine amongst an open-set class of attacks in the most challenging dataset so far.

3. Deep Anomaly Detection for Face GPAD

3.1. Review on Metric Learning

Many works rely on a softmax loss function to separate samples from different classes in deep learning models. However, class compactness is not explicitly considered and samples from different classes might easily overlap in the feature space. Instead, metric learning based losses are designed to address these issues, by promoting inter-class separability and reducing intra-class variance. Note that several metric learning approached have been applied to multiple tasks such as face recognition [26], object retrieval [17] or person re-identification [38], obtaining outstanding generalization performance. In this section we introduce the mathematical notation and our formulation for the problem of deep anomaly detection for face GPAD, from a metric learning perspective.

Let $f_\theta(x_i)$ be the feature vector in the embedding space of a data point $x_i \in \mathbb{R}^N$, where the mapping function $f_\theta : \mathbb{R}^N \rightarrow \mathbb{R}^D$ is a differentiable deep neural network of parameters θ, and let $D_{i,j}$ be the squared l2-norm between two feature vectors defined by $D_{i,j} = \|f_\theta(x_i) - f_\theta(x_j)\|_2^2$. Usually, $f_\theta(x_i)$ is normalized to have unit length for training stability. In a deep metric learning based approach, the objective is to learn a deep model that generates a feature representation $f_\theta(x_i)$ to guarantee that samples from the same class are closer in the embedding space, than samples from different categories. For doing so, different loss functions can be found in the literature.

For instance, the center loss proposed in [34] concentrates samples around their class centers in the embedding space (see Eq. 1). It is used in conjunction with the softmax loss to increase intra-class compactness, however the latter does not guarantee a correct inter-class separation.

$$\mathcal{L}_c(\theta) = \frac{1}{2}\sum_{i=1}^{b} \|f_\theta(x_i) - c_{y^i}\|_2^2, \tag{1}$$

where b is the number of input tuples in a batch and c_{y^i} is the class center corresponding to the ground truth label y^i of sample x_i.

The contrastive loss [9] (see Eq. 2) forces all images belonging to the same class to be close, while samples from different classes should be separated by a margin m. It uses tuples of two images as different image pairs $\{p, q\}$: i) *positive*, if both belong to the same class and ii) *negative*, otherwise. However, one needs to fix a constant margin m for the negative pairs, separating all negative examples by the same margin regardless their visual appearance:

$$\mathcal{L}_{ct}(\theta) = \sum_{i=1}^{b} y_{p_i,q_i} D_{p_i,q_i} + (1 - y_{p_i,q_i}) \max(0, m - D_{p_i,q_i})^2, \tag{2}$$

where $y_{p_i,q_i} = 1$ for the positive pair and $y_{p_i,q_i} = 0$ for the negative.

Following the same idea, the authors of the triplet loss [32] extend the contrastive loss to consider positive and negative pairs simultaneously by using a tuple of three images: i) anchor, ii) positive and iii) negative. The goal of the triplet loss in Eq. 3 is to reduce the intra-class variance defined by the anchor-positive pair, while simultaneously increase the inter-class separation by maximizing the euclidean distance between the anchor-negative pair. Despite avoiding a constant margin for the negative pair and obtaining highly discriminative features, it suffers from the complexity of the triplet selection procedure. Nevertheless, it has been successfully addressed in many recent approaches [15, 31, 35].

$$\mathcal{L}_t(\theta) = \sum_{i=1}^{b} \max(0, D_{a_i,p_i} - D_{a_i,n_i}) + m), \tag{3}$$

where $\{a_i, p_i, n_i\}$ sub-indexes are the anchor, the positive and the negative samples for each triplet within the batch, respectively.

Prior works successfully applied the triplet loss (or any of its variants) using a large number of classes, *e.g.* face recognition models use thousands of identities, for instance in VGG2 Face data set [7] there are more than 9000 different identities. Such a diversity of classes encourages embeddings to generalize when the number of samples is large enough. In this paper, we show that a triplet loss based model, following an anomaly detection perspective, can actually outperform existing methods for face GPAD.

3.2. Triplet Focal Loss for Anomaly Detection for Face GPAD

We address the face GPAD problem from a metric learning approach with a Triplet focal loss. Technically, we propose to use a modified version of the triplet loss described in [29] that incorporates focal attention, see Eq. 4. The triplet focal loss automatically up-weights hard examples by mapping the euclidean distance to an exponential kernel, penalizing them much more than the easy ones.

$$\mathcal{L}_{\text{tf}}(\theta) = \sum_{i=1}^{b} \max\left(0, e^{\left(\frac{D_{a_i,p_i}}{\sigma}\right)} - e^{\left(\frac{D_{a_i,n_i}}{\sigma}\right)} + m\right), \quad (4)$$

where σ is the hyper-parameter that controls the strength of the exponential kernel.

The triplets generation scheme is a critical step that highly impacts the final performance. Traditional methods run their sample strategy over the training set in an off-line fashion, and they do not adapt once the learning process starts. Alternatively, we use an approach for triplets selection based on a semi-hard batch negative mining process, where triplets examples are updated during the training process in each mini-batch, avoiding models to collapse.

The goal of the implemented semi-hard batch negative mining (based on [26]) is to choose a negative sample that is fairly hard within a batch but not necessarily the hardest nor the easiest one. For each training step, we select a large set of samples of each class using the current weights of the network. Next, we compute the distances between all positive pairs within this population, *i.e.* $D_{a,p}$, and, for each positive pair, we compute the distance between the corresponding anchor $f(x_a)$ and all possible negative samples $f(x_n)$. Finally, we randomly pick a negative sample that satisfies the following margin criteria, $D_{a,p} - D_{a,n} < m$, to build the final tuples that are used for training at each step, in the so called mini-batch. This mining strategy has two important benefits: 1) we ensure that all the samples included in a training step are relevant for the learning process; and 2) we improve training convergence thanks to the random selection over the negative samples.

In real face anti-spoofing, attackers are constantly engineering new ways to cheat PAD systems with new attacks, materials, devices, *etc*. Thus, a classification-like approach is prone to over-fitting to the seen classes and will not generalize well. On the contrary, we follow an anomaly detection based strategy. First, we do not consider the identity of the users as different classes. We define two categories in an anomaly detection setting: 1) the *closed-set*, referring to the classes that can be correctly modeled during training; and 2) the *open-set*, referring to all the classes that cannot be fully modeled by the training set. In face GPAD, genuine samples belong to the closed-set category, while impostors belong to the open-set class, motivated by the scarcity or even the lack of training samples to model certain types of attacks. To achieve this, we fix during training the anchor-positive pair to always belong to the genuine class (*i.e.* the closed-set category) while selecting negative samples from any type of attack (*i.e.* the open-set category) regardless their identity.

3.3. Triplet Loss Regularization for a Metric-Softmax

Recent work [17] demonstrates that the triplet loss, acting as a regularizer of the softmax function, achieves more discriminative and robust embeddings. In our deep anomaly detection based model, we do not focus on the classification task, but instead we aim at obtaining highly representative embeddings to distinguish genuine samples amongst an open-set class of attacks. We thus propose to add the triplet focal loss as a regularizer of a novel softmax function adapted to metric learning, see Eq. 5. The proposed softmax formulation, coined as *metric-softmax* ($\mathcal{L}_{metric\_soft}$ in Eq. 6), accumulates the probability distribution of each pair within a triplet to be highly separated in an euclidean space. We thus prevent from guiding the learning process towards a binary classification and thus avoiding the well known generalization issues.

$$\mathcal{L}_{\text{anomaly}} = \mathcal{L}_{metric\_soft} + \lambda\, \mathcal{L}_{tf}, \quad (5)$$

$$\mathcal{L}_{\text{metric\_soft}} = -\sum_{i=1}^{b} \log \frac{e^{D_{a_i,p_i}}}{e^{D_{a_i,p_i}} + e^{D_{a_i,n_i}}}, \quad (6)$$

where λ is the hyper-parameter to control the trade-off between the triplet focal loss and the softmax loss.

The metric learning model proposed obtains a discriminative embedding for every input image. However, we need to provide a posterior probability of whether the image belongs to a genuine sample or to an impersonation attempt. In the experiments, we simply propose to train an SVM classifier with a Radial Basis Function to learn the boundaries between both classes in the feature space.

3.4. Few-shot a Posteriori Probability Estimation

Often, the inherent dynamic nature of spoofing attacks and the difficulty to access data requires to adapt rapidly to new environments where few samples are available. To deal with this problem, we propose a *few-shot a posteriori estimation* procedure, that does not need any classifier to train on the learned features for decision making in metric learning.

Technically, we proceed to compute the probability of being genuine (see Eq. 7) as the accumulated posterior probability of the input sample (x_t) given two reference sets in the target domain, corresponding to the genuine class $\mathcal{G}$ and the attacks $\mathcal{H}$, respectively.

$$P(x_t \mid \{\mathcal{G}, \mathcal{H}\}) = \sum_{i=1}^{M} \frac{e^{D_{t,g_i}}}{e^{D_{t,g_i}} + e^{D_{t,h_i}}}, \quad (7)$$

where M is the total number of pairs in both reference sets for every attack and for each dataset involved, t sub-index refers to the test image and $\{g_i, h_i\}$ sub-indexes refer to each of the reference samples in the genuine and attack sets, respectively. In order to satisfy the few-shot constraints we choose M to be small in our experiments.

4. Experimental Results

In this section we present the experiments where our novel approach is compared against three state-of-the-art methods from the literature. The approach in [25] computes hand-crafted features based on *quality* evidences. They obtain a 139-length feature vector from the concatenation of the quality measurements proposed in [14] and [33]. For the second method, we choose [4], which consists in computing a color-based feature vector of high dimensionality (19998-length) by concatenating texture features based on Local Binary Patterns (LBPs) in two different color spaces (*i.e*. YCbCr and HSV). Finally, the third method is the one proposed in [23], which introduces a two-branch deep neural network that incorporates pixel-wise auxiliary supervision constrained by the depth reconstruction for all genuine samples (attacks are forced to belong to a plane) and the estimation of a remote PhotoPlethysmoGraphy (rPPG) signal to add temporal information. Despite being the state of the art for face anti-spoofing, this model requires to pre-process genuine samples in order to compute the depth estimation and the corresponding rPPG signal, that impacts in the usability and bounds the performance to the methods for depth reconstruction and rPPG estimation. The code for the first two algorithms is based on the reproducible material provided by the authors[3] [4]. Results for [23] are obtained using our own re-implementation of their approach.

4.1. GRAD-GPAD Framework

Regardless almost every paper comes with its own reduced dataset [21, 23, 24, 37], there is *no agreed upon a PAD benchmark*, and as a consequence, the generalization properties of the models are not properly evaluated. During a brief inspection of the capture settings of available face PAD datasets, one can easily observe that there is no unified criteria in the goals of each of them, leading to a manifest built-in bias. This specificity in the domain covered by most of the datasets can be observed in different scenarios: i) some of them focus on a single type of attacks (*e.g*. masks - 3DMAD [13], HKBU [22], CSMAD [3]); ii) others focus on the study of different image sources (depth/NIR/thermal) such as CASIA-SURF [37] or CSMAD; iii) others attempt to simulate a certain scenario like a mobile device setting, where the user hold the device (*e.g*. Replay-Mobile [10], OULU-NPU [6]), or a webcam setting, where the user is placed in front of a fixed camera (*e.g*. Replay-Attack [8], SiW [23]), or even a stand-up scenario where users are recorded further from the camera (*e.g*. UVAD [27]).

For our experiments, we propose to use the recently published GRAD-GPAD framework [11] that mitigates the aforementioned limitations. GRAD-GPAD is the largest aggregated dataset that unifies more than 10 datasets with a common categorization in two levels, to represent four key aspects in anti-spoofing: attacks, lightning, capture devices and resolution. It allows not only a fair evaluation of the generalization properties, but also a better representativity of the face-PAD problem thanks to the increased volume of data. For the sake of the extension of the paper we focus on the evaluation based on the instruments used to perform attacks (*i.e*. PAI - Presentation Attack Instruments) using the categorization in Table 1 (*i.e*. the *Grandtest* protocol).

Category	Types	Sub-type	Criteria
Presentation Attack Instrument	print	low	dpi $\leq$ 600pix
		medium	600 $<$ dpi $\leq$ 1000pix
		high	dpi $>$ 1000pix
	replay	low	res $\leq$ 480pix
		medium	480 $<$ res $<$ 1080pix
		high	res $\geq$ 1080pix
	mask	paper	paper masks
		rigid	non-flexible, plaster
		silicone	silicone masks

Table 1: Two-tier common PAI categorization in GRAD-GPAD.

We conduct all the experiments using the GRAD-GPAD framework, where we add the UVAD dataset [27] to further increase the total number of samples in more than 10k images. In Fig. 2 we show the population statistics of the whole GRAD-GPAD dataset (left figure) and the training split of the *Grandtest* protocol (right figure).

4.2. Experimental Setup

Network Architecture We use as our backbone architecture a modified version of the ResNet-50 [16]. We stack both RGB and HSV color spaces in the input volume, and feature dimension is fixed to 512. We use *Stochastic Gradient Descent* with *Momentum* optimizer. We start training with a learning rate of 0.01 using a maximum of 100 epochs. Batch size is fixed to be 12 *triplets*, *i.e*. 36 images per batch. As suggested in the original works, σ and m values in Eq. 4 are set to 0.3 and 0.2, respectively.

Pre-processing Since our approach follows a frame-based procedure, instead of using the full videos we only pick the central frame of each video. We use as inputs of the network the cropped faces detected using the method proposed in [36].

[3] https://github.com/zboulkenafet/Face-anti-spoofing-based-on-color-texture-analysis

[4] https://gitlab.idiap.ch/bob/bob.pad.face/

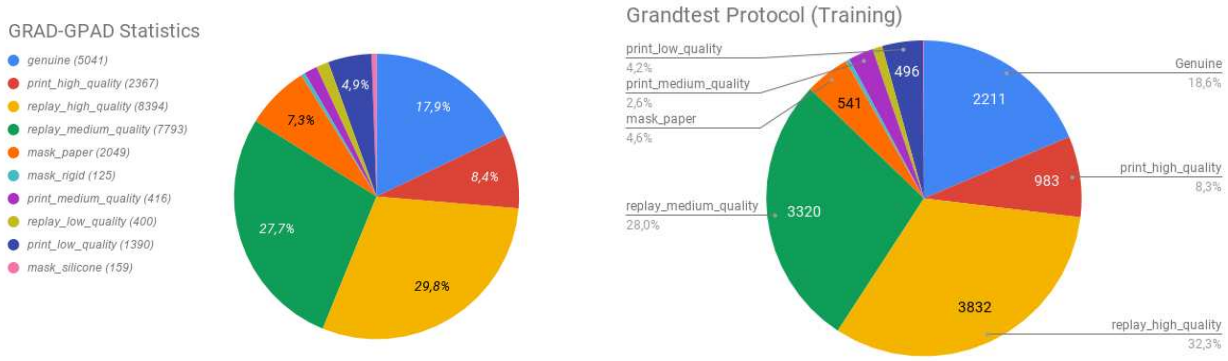

Figure 2: Population statistics for the whole dataset provided in GRAD-GPAD (left) and the training samples statistics for the *Grandtest* protocol (right).

Metrics To compare our method with prior works we use the metrics that have been recently standardized in the *ISO/IEC 30107-3*[5]: *i.e*. False Acceptance Rate (FAR), False Rejection Rate (FRR), Half Total Error Rate (HTER $= \frac{\text{FAR}+\text{FRR}}{2}$), Attack Presentation Classification Error Rate (APCER), Bonafide Presentation Classification Error Rate (BPCER) and Average Classification Error Rate (ACER). We would like to highlight the importance of the ACER metric because it entails the most challenging scenario, where performance is computed for every attack independently, but it only considers the results for the worst scenario. Thus it penalizes approaches performing well on certain types of attacks. HTER reflects the overall performance of the algorithm in a balanced setting where FAR is equal to FRR, *i.e*. for Equal Error Rate (EER).

Protocols We evaluate our method on two settings within the GRAD-GPAD framework: 1) intra-dataset; and 2) inter-dataset. For the intra-dataset setting we use the *Grandtest* protocol and for the inter-dataset evaluation we use the leave-one-dataset-out protocols, provided by the framework: the *Cross-Dataset-Test-On-CASIA-FASD* and the *Cross-Dataset-Test-On-ReplayAttack*. In these protocols, one of the datasets (CASIA-FASD and Replay-Attack, respectively) is excluded during training. Results are provided by evaluating the models in the excluded dataset (Test split).

4.3. Ablation study

The scientific contribution of our work is twofold. First, we introduce a reformulation of the face PAD problem from a deep anomaly detection perspective using metric learning. Second, we propose to use a triplet focal loss as a regularization for a novel softmax loss function adapted to metric learning, coined as "metric-softmax". To show the influence of each of these contributions, we conduct the following experiments. We start from a classification-like triplet loss based model, *i.e*. without the anomaly detection approach. This first approach is named as *Baseline* in Table 2, where tuples for the triplets are selected randomly from the set of classes (genuine + 9 different attacks in GRAD-GPAD). We then incrementally incorporate our contributions. *Model 1* includes the anomaly approach using the triplet loss. In *Model 2* we included the focal attention into the triplet loss formulation. And finally, *Ours* represents the whole pipeline of our system, where the proposed metric-softmax term is added. The results reported in Table 2 show the influence of each contribution in the final performance.

Note that for this ablation study, we use the development split of the *Grandtest* protocol of GRAD-GPAD, and the performance is shown in terms: FAR, FRR and Average Error Rate (AER). Besides, performance is computed using the accumulated metric-softmax distribution described in Eq. 7 with $M = 3$ and by randomly choosing samples from the training set.

Algorithm	AER	FAR	FRR	ΔAER
Baseline	17.02 %	10.72 %	23.33 %	-
Model 1	12.46 %	24.43 %	**0.49 %**	**26.8 %**
Model 2	9.80 %	14.87 %	4.74 %	**42.42 %**
Ours	**5.07 %**	**6.38 %**	3.77 %	**70.21 %**

Table 2: Performance evaluation in the development set of GRAD-GPAD for the different models involved in the ablation study. We also show the relative improvement ΔAER with respect to the baseline.

We show in Table 2 that, when we incorporate the focal attention into the triplet, *i.e*. *Model 2*, we achieve a relative improvement of 42.42% in terms of AER. This aspect

[5]https://www.iso.org/standard/67381.html

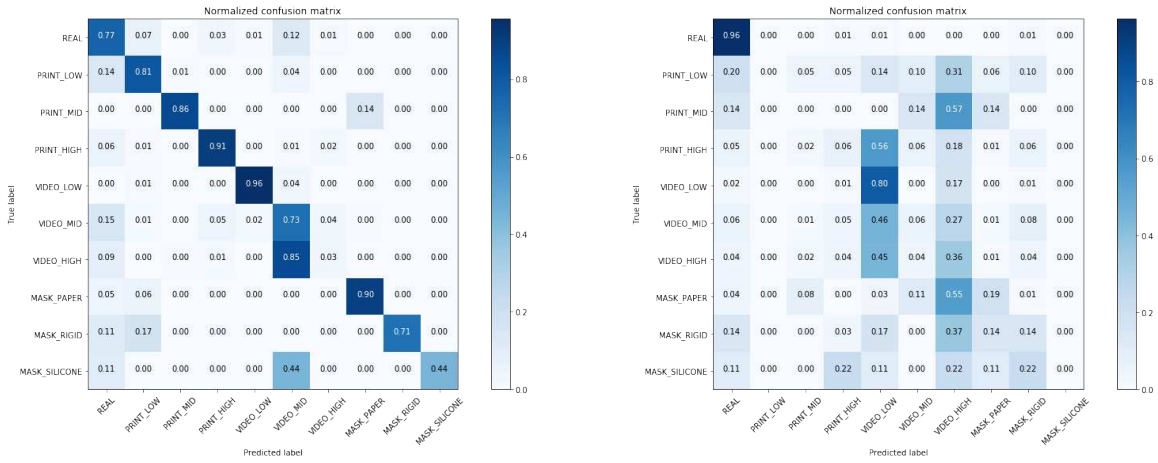

Figure 3: True Positive Rate confusion matrices for the baseline (left) and our approach (right).

reveals the importance of a mining strategy in the learning process. Finally, the introduction of the proposed metric-softmax term, achieves a remarkable relative improvement of 70.21% of AER.

Furthermore, we show in Fig. 3 the True Positive Rate (TPR) confusion matrices for the *Baseline* (left) and our approach (right). We assess that, with the anomaly detection approach, we are able to highly differentiate genuine from impostor samples, regardless the classification of the attack instrument. Note that the *baseline* obtains poor performance for genuine samples classification, despite classifying correctly the different attacks, which highly penalizes its global performance.

4.4. Intra-Dataset Evaluation

In order to fairly compare our approach with the state-of-the-art methods, we train an SVM-RBF classifier for each of them using their corresponding features. Additionally, for the *Auxiliary* model [23], we report the results just using the L2-Norm from the depth map (*Auxiliary**), as it is proposed by the authors in their original work. For all the experiments, we use $M = 3$ in Eq. 7 for the few-shot a posteriori probability estimation (*Ours*†) experiment. Note that, both the original method proposed in [23] (*Auxiliary**) and our approach with a posteriori estimation, do not need to use any classifier with the learned features.

Results in Table 3 demonstrate that both our novel approaches outperform the state-of-art methods, even using the most challenging metric (ACER). These results highlight that the learned feature space has a high discrimination capability and that our model performs the best.

4.5. Inter-Dataset Evaluation

In order to assess the generalization capabilities, we perform two cross-dataset evaluations where a whole dataset is excluded from the training step. In the first experiment, we leave out CASIA-FASD [39] for the test set. In the second one, ReplayAttack [8] is excluded during learning. In both experiments, none of the samples from the test dataset are used neither in the training set nor in the development set.

Algorithm	HTER	ACER	APCER	BPCER
Quality [25]	23.21 %	36.96 %	50.51 %	23.42 %
Color [4]	7.87 %	19.21 %	28.57 %	9.84 %
Auxiliary [23]	5.92 %	37.89 %	66.67 %	8.55 %
Auxiliary*	6.52 %	31.81 %	53.33 %	10.44 %
Ours	**5.41 %**	**10.14 %**	14.29 %	**5.99 %**
Ours†	5.45 %	10.42 %	**14.28** %	6.55%

Table 3: Intra-dataset results on the *Grandtest* protocol.

4.5.1 Test on CASIA-FASD

As it is shown in Fig. 4, the training set for this experiment includes all types of attacks, however the domain is different (*i.e.* different environments, lighting conditions, capture devices, *etc.*). CASIA-FASD is one of the smallest datasets for face anti-spoofing samples. Therefore, considering only its test set for the evaluation would highly penalize the performance in case of miss-classification. This fact is reflected in Table 4, where performance significantly drops in all methods, except for our approach, where we are able to keep a reasonable good performance: from an ACER of 10.14% (see Table 3) to 16.8%.

In Table 4 we show that HTER and ACER values for our approach are almost the same. We argue that, despite the domain shift introduced by this protocol, the learned embeddings during training are robust enough to generalize in this setting. Instead, the other methods in the experiment are highly penalized, showing that they tend to overfit on

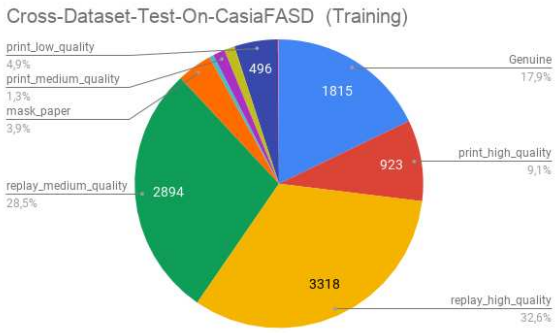

Figure 4: Training samples statistics for the *Cross-Dataset-Test-On-CASIA-FASD* protocol.

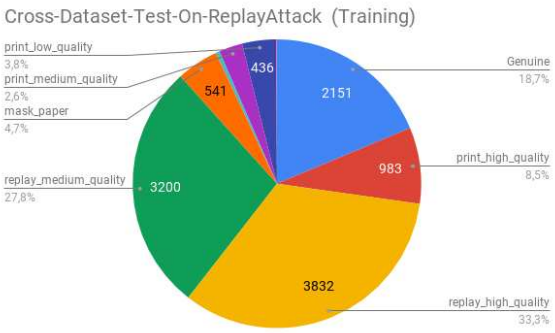

Figure 5: Training samples statistics for the *Cross-Dataset-Test-On-ReplayAttack* protocol.

Algorithm	HTER	ACER	APCER	BPCER
Quality [25]	40.90 %	47.38 %	65.56 %	29.21 %
Color [4]	22.17 %	25.69 %	26.67 %	24.72 %
Auxiliary [23]	28.60 %	29.71 %	12.22 %	47.19 %
Auxiliary*	25.42 %	26.90 %	12.22 %	41.57 %
Ours	**16.74 %**	**16.80 %**	**10.00 %**	**23.60 %**
Ours†	17.56 %	18.48 %	**10.00** %	26.97 %

Table 4: Inter-dataset results evaluated on CASIA-FASD.

Algorithm	HTER	ACER	APCER	BPCER
Quality [25]	37.35 %	47.02 %	**42.14 %**	51.90 %
Color [4]	34.51 %	**43.35 %**	51.25 %	35.44 %
Auxiliary [23]	35.62 %	45.62 %	68.75 %	22.50 %
Auxiliary*	37.87%	47.50 %	72.50 %	22.50 %
Ours	**25.00** %	45.62 %	71.25 %	**20.00** %
Ours†	25.25 %	45.62 %	71.25 %	**20.00 %**

Table 5: Inter-dataset results evaluated on Replay Attack.

the training set to a greater extent.

Besides, we show that our few-shot a posteriori estimation pipeline (Ours†) achieves similar performance compared to the SVM version in this Test on CASIA-FASD setup. Thus, we assess that the learnt embedding space generalizes enough so that we can avoid using a classifier with the feature vectors and estimate the a posteriori probability by simply using $M = 3$. This classiffier-free model is also able to outperform all state-of-the-art methods, including *Auxiliary** that neither requires a classifier.

4.5.2 Test on Replay-Attack

The motivation behind selecting to leave out the Replay-Attack dataset is to show the impact in the performance of face-PAD algorithms of unseen attacks belonging to a new domain: this dataset contains all the samples for *replay-low-quality* attacks (see Fig. 5). This entails a far more challenging scenario.

The results reported in Table 5 show a severe drop of performance for all the methods, specially for ACER, where all the approaches are highly penalized by the unseen attack and achieves performance close to random choice. This fact is due to the addition of a new attack that has never seen before in combination with a strong domain change, highly impacting on APCER (*i.e.* the attack classification). Interestingly, our proposal based on few-shot a posteriori estimation keeps exactly the same performance compared with our method with an SVM, again assessing that we can replace the classifier using a few samples. Besides, we obtain the best overall performance HTER and the best BPCER (ACER is close to random choice for all the methods).

5. Conclusions

In this work we introduce a novel approach that addresses the problem of generalization in face-PAD, following an anomaly detection pipeline. We leverage deep metric learning to propose a new "metric-softmax" loss that applied in conjunction with the triplet focal loss drives to more robust and generalized features representations to distinguish between original and attack samples. We also propose a new a posteriori probability estimation that prevents us from the need of training any classifier for decision making. With a thorough experimental evaluation in the challenging GRAD-GPAD framework we show that the proposed solution outperforms prior works by a considerable margin.

Acknowledgements We thank our colleagues of the Biometrics Team at Gradiant for their valuable contributions. Special mention to Esteban Vazquez-Fernandez, Julián Lamoso-Núñez and Miguel Lorenzo-Montoto.

References

[1] A. Anjos and S. Marcel. Counter-measures to photo attacks in face recognition: A public database and a baseline. In *International Joint Conference on Biometrics (IJCB)*, 2011. 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*, 2017. 2, 3

[3] Sushil Bhattacharjee, Amir Mohammadi, and Sébastien Marcel. Spoofing Deep Face Recognition With Custom Silicone Masks. In *BTAS*, 2018. 2, 5

[4] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security (TIFS)*, 11(8):1818–1830, Aug 2016. 5, 7, 8

[5] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. On the generalization of color texture-based face anti-spoofing. *Image and Vision Computing*, 77:1–9, 2018. 2

[6] Z. Boulkenafet, J. Komulainen, Lei. Li, X. Feng, and A. Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. In *IEEE International Conference on Automatic Face and Gesture Recognition*, May 2017. 5

[7] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018. 3

[8] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, September 2012. 3, 5, 7

[9] Sumit Chopra, Raia Hadsell, Yann LeCun, et al. Learning a similarity metric discriminatively, with application to face verification. In *CVPR (1)*, pages 539–546, 2005. 3

[10] Artur Costa-Pazo, Sushil Bhattacharjee, Esteban Vazquez-Fernandez, and Sébastien Marcel. The replay-mobile face presentation-attack database. In *BioSIG*, Sept. 2016. 3, 5

[11] Artur Costa-Pazo, David Jiménez-Cabello, Esteban Vázquez-Fernández, José Luis Alba-Castro, and Roberto J. López-Sastre. Generalized presentation attack detection: a face anti-spoofing evaluation proposal. In *International Conference on Biometrics (ICB)*, 2019. 2, 3, 5

[12] Artur Costa-Pazo, Esteban Vazquez-Fernandez, José Luis Alba-Castro, and Daniel González-Jiménez. Challenges of face presentation attack detection in real scenarios. In *Handbook of Biometric Anti-Spoofing, Springer*, 2019. 2

[13] Nesli Erdogmus and Sebastien Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *Biometrics: Theory, Applications and Systems*, September 2013. 5

[14] J. Galbally, S. Marcel, and J. Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE Transactions on Image Processing*, 23(2):710–724, Feb 2014. 2, 5

[15] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–285, 2018. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[17] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3d object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1945–1954, 2018. 3, 4

[18] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 290–306, 2018. 2

[19] Klaus Kollreider, Hartwig Fronthaler, Maycel Isaac Faraj, and Josef Bigun. Real-time face detection and motion analysis with application in liveness assessment. *IEEE Transactions on Information Forensics and Security (TIFS)*, 2(3):548–558, 2007. 2

[20] Haoliang Li, Peisong He, Shiqi Wang, Anderson Rocha, Xinghao Jiang, and Alex C Kot. Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security (TIFS)*, 13(10):2639–2652, 2018. 2

[21] 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 (TIFS)*, 13(7):1794–1809, July 2018. 2, 5

[22] Siqi Liu, Pong C. Yuen, Shengping Zhang, and Guoying Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In *European Conference on Computer Vision (ECCV)*, pages 85–100. Springer International Publishing, 2016. 5

[23] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 389–398, 2018. 5, 7, 8

[24] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar. Detecting silicone mask-based presentation attack via deep dictionary learning. *IEEE Transactions on Information Forensics and Security (TIFS)*, 12(7):1713–1723, July 2017. 5

[25] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel. On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. In *International Conference on Biometrics (ICB)*, pages 75–81, Feb 2018. 2, 3, 5, 7, 8

[26] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *British Machine Vision Conference (BMVC)*, volume 1, page 6, 2015. 3, 4

[27] A. Pinto, W. Robson Schwartz, H. Pedrini, and A. de Rezende Rocha. Using visual rhythms for detecting video-based facial spoof attacks. *IEEE Transactions on Information Forensics and Security (TIFS)*, 10(5):1025–1038, May 2015. 5

[28] A. Pinto, W. R. Schwartz, H. Pedrini, and A. d. R. Rocha. Using visual rhythms for detecting video-based facial spoof attacks. *IEEE TIFS*, 10(5):1025–1038, May 2015. 2

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 4

[30] A. Sepas-Moghaddam, F. Pereira, and P. L. Correia. Light field-based face presentation attack detection: Reviewing, benchmarking and one step further. *IEEE Transactions on Information Forensics and Security (TIFS)*, 13(7):1696–1709, July 2018. 2

[31] Evgeny Smirnov, Aleksandr Melnikov, Andrei Oleinik, Elizaveta Ivanova, Ilya Kalinovskiy, and Eugene Luckyanets. Hard example mining with auxiliary embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 37–46, 2018. 3

[32] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, June 2009. 3

[33] Di Wen, Hu Han, and A.K. Jain. Face Spoof Detection with Image Distortion Analysis. *IEEE TIFS*, 10(4):746–761, April 2015. 2, 3, 5

[34] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 3

[35] Baosheng Yu, Tongliang Liu, Mingming Gong, Changxing Ding, and Dacheng Tao. Correcting the triplet selection bias for triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–87, 2018. 3

[36] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, abs/1604.02878, 2016. 5

[37] Shifeng Zhang, Xiaobo Wang, Chenxu Zhao, Jun Wan, Sergio Escalera, Hailin Shi, Zezheng Wang, and Stan Z.Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 5

[38] S. Zhang, Q. Zhang, X. Wei, Y. Zhang, and Y. Xia. Person re-identification with triplet focal loss. *IEEE Access*, 6:78092–78099, 2018. 3

[39] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. A face antispoofing database with diverse attacks. In *International Conference on Biometrics (ICB)*, pages 26–31, March 2012. 3, 7