Face presentation attack detection. A comprehensive evaluation of the generalisation problem

Artur Costa-Pazo1,2 | Daniel Pérez-Cabo1,2,3 | David Jiménez-Cabello4 | José Luis Alba-Castro2 | Esteban Vázquez-Fernández1

1Alice Biometrics, Vigo, Spain  
2atlanTIC research center at University of Vigo, Vigo, Spain  
3GRADIANT, Vigo, Spain  
4R&D, Nielsen IQ, Madrid, Spain  

Abstract  
Face recognition technology is now mature enough to reach commercial products, such as smart phones or tablets. However, it still needs to increase robustness against impostor attacks. In this regard, face Presentation Attack Detection (face-PAD) is a key component in providing trustable facial access to digital devices. Despite the success of several face-PAD works in publicly available datasets, most of them fail to reach the market, revealing the lack of evaluation frameworks that represent realistic settings. Here, an extensive analysis of the generalisation problem in face-PAD is provided, jointly with an evaluation strategy based on the aggregation of most publicly available datasets and a set of novel protocols to cover the most realistic settings, including a novel demographic bias analysis. Besides, a new fine-grained categorisation of presentation attacks and instruments is provided, enabling higher flexibility in assessing the generalisation of different algorithms under a common framework. As a result, GRAD-GPAD v2, a comprehensive and modular framework is presented to evaluate the performance of face-PAD approaches in realistic settings, enabling accountability and fair comparison of most face-PAD approaches in the literature.

1 | INTRODUCTION

Face recognition (FR) technology has become a key element in a wide range of applications. These systems are increasingly used in different commercial products to verify the identity of the users, such as onboarding processes, ID document verification and access controls in airports or borders. Although face recognition has reached impressive performance, even outperforming humans in some situations [1], its security is still a challenge. To enable the inclusion of this technology into trustable and scalable applications, we need to deal with several non-mutually exclusive aspects: (1) the types of presentation attacks, (2) the generalisation capabilities of the decision system, (3) the potential bias of the decisions and (4) the characteristics of the production environment. We argue that a system that aims at performing face presentation attack detection (face-PAD) needs to be evaluated considering these 4 dimensions and that the omission of any of them has a critical impact on the assessment of a face-PAD system.

Face-PAD is indeed the weakest point for commercial face recognition applications and a challenging problem for the biometrics community. Technically, a face-PAD solution must determine whether the face presented to the system is a bona fide presentation (BFP) or a presentation attack (PA). A variety of Presentation Attack Instruments (PAIs) are used to perform presentation attacks (e.g. videos, printed pictures, masks, etc), aiming to both impersonate or obfuscate a bona fide presentation. But attackers will always be looking for new ways to bypass the security of FR systems, and hence, a face-PAD model should be able to detect from simple replayed videos to dynamic 3-D face avatars or custom-made flexible silicone masks. Such a complex scenario, where the tools and the dexterity of attackers are in constant change, manifests the urgent necessity of not only designing robust and scalable algorithms but also strong evaluation frameworks that let us evaluate and anticipate all these vulnerabilities.

Current state-of-the-art approaches have achieved unprecedented results in standard benchmarks and datasets [2, 3]. The reasons are clear: (a) the available data is orders of magnitude under the amount of the model's parameters and (b) all the presentation attacks are known beforehand and are completely represented by the training data. The problem...
arises when the trained face-PAD models have to operate in a real-world scenario. To list some of them, a face-PAD solution should be robust to (i) different domains (e.g. indoor, outdoor), (ii) different PAIs (e.g. replay, print, masks) with (iii) different objectives (e.g. impersonation or obfuscation), (iv) using a wide variety of devices (e.g. from low- to high-quality capture devices), (v) fairly considering people from different ethnicities, sex, age or skin tones and (vi) mitigating the security breach of new appearing PAIs. 1 Behind this tremendous complexity, the generalisation problem arises as follows: state-of-the-art approaches exhibit a severe drop of performance due to the possible overfitting that maximises their accuracy for just the benchmarks used to train and validate them [5].

Another missing piece is demographic bias, recently revealed as a critical ethical issue in commercial face recognition systems [6] that forced even big sellers to stop providing this technology to law enforcement. Fairness is a critical aspect for any deployable solution aiming at creating models that are agnostic to different social and demographic characteristics. Recently, the authors in [7] presented for the first time an attempt to evaluate the ethnic bias in a self-designed dataset and protocol. However, there is no prior work that evaluates in detail different sources of bias in face-PAD approaches, such as the influence of different demographics groups. In this work, we incorporate for the first time protocols devoted to evaluating the influence of demographics in face-PAD approaches by casting the bias evaluation as a generalisation problem over the aforementioned characteristics.

Moving from research to production is not a trivial task. Apart from the generalisation problem, it is important to consider the penetration of the different PAIs into a deployment environment. On the one hand, when placing a fully automated (but assisted by a human) FR-based access control in an airport, it implicitly restricts the possible range of PAIs that could appear (e.g. make-up or professional silicone masks) for obvious reasons. On the other hand, FR systems that might appear in online onboarding procedures to enroll in an application (no additional human surveillance) pose a challenging scenario, where attackers could try any PAI and invest the time needed to perform high-quality presentation attacks. This prior knowledge about the deployment scenario should also be injected in face-PAD assessments, requiring a fine-grained categorisation of all possible PAIs that threaten the security of an FR system. In this work, we increase the granularity of the largest aggregated dataset extending the GRAD framework [8] to cover a wider range of PAIs for both impersonation and obfuscation attacks. We also review all publicly available datasets and categorise them accordingly with the suggested schema.

In conclusion, despite the large efforts of the community to build representative datasets, all of them appear to have strong limitations. From a generalisation perspective, these limitations come from the restricted sources of information in current isolated datasets: environment, background, illumination, acquisition devices, demographics, and specifically for the face-PAD problem, the wide range of known and unknown PAIs. We propose to analyse the status quo of the evaluations of face-PAD approaches with a proposal centred around the generalised face-PAD issue. Additionally, we analyse the bias of the face-PAD systems from a demographic perspective, analysing the performance of the algorithms depending on three main aspects: (1) sex (2) age and (3) skin tone. The main contributions of this work include the following:

- We introduce a new version of the largest aggregated dataset (GRAD-GPAD [5]) with the categorisation of three new datasets: UVAD [9], HKBU v2 [10], and SiW-M [4]. The new version of GRAD-GPAD (v2) aggregates 13 datasets, increasing the number of identities (300% more) as well as the number of samples (181% more).
- We provide a novel categorisation and labelling in GRAD-GPAD v2 for demographics groups: sex, age and skin tone. Based on that, we introduce novel demographic protocols, visualisation tools and metrics to detect and measure the existence of biases in the face-PAD performance among different demographic groups.
- We introduce two metrics to evaluate generalisation on face-PAD and a novel visual evaluation tool (the ‘face-PAD-radar’) to help with face-PAD explainability. In order to show the benefits of using these novel tools over the former ones, we evaluate and compare two state-of-the-art face-PAD methods.
- We release an open-source framework2 providing tools to reproduce the research. In addition, the labelling of all GRAD-GPAD samples is provided in JSON format.

The remainder of this work is organised as follows: Section 2 provides an overview of the challenges in the field of generalisation for face-PAD systems. The new version of GRAD-GPAD is presented in Section 3, where novel categorisations, protocols, visualisation proposals and metrics are described in detail. In Section 4, we analyse in detail the benefits of new proposals as well as the challenges that still lie ahead using as examples two state-of-the-art face-PAD systems with quite different algorithmic approaches, one based on image quality measurements and the SVM classifier and the other based on DNNs with auxiliary information. Conclusions are presented in Section 5.

2 | RELATED WORK

In [11], Freitas et al. posed the following question: ‘Can face anti-spoofing countermeasures work in a real world scenario?’ But yet there is no clear answer to it. There are several unresolved open generalisation issues and many questions that arise when the system has to be deployed in production. In this

---

1 We adopt the same intuition as in [4] where the authors differentiate between unknown and unseen PAIs. The former refers to those PAIs that we are not aware of its existence and there is no chance to anticipate them. The latter refers to those PAIs that we do not have samples to learn from but we know about their existence.

2 https://github.com/acostapazo/gradgpad
section, we discuss the problem of face-PAD with a focus on generalisation (face-GPAD) and all the efforts that the community is making to solve this issue.

2.1 | Generalisation in face-PAD

Generalisation in face-PAD has become more relevant in recent years. Current state-of-the-art solutions [2–4] are making great strides with significant performance improvements on face-PAD. However, these approaches continue to experience a severe drop of performance when tested in realistic scenarios [8, 12, 13]. In [5, 8], the authors show that the major challenges of face-PAD models are the following: (i) the lack of data, as current datasets do not represent realistic deployment scenarios, (ii) no clear consensus on the evaluations protocols, where almost each proposal comes with its own dataset and its own evaluation protocols, making it very difficult to perform fair comparisons between algorithms, and (iii) generalisation, as current state-of-the-art algorithms are highly penalised when they are evaluated on generalisation-specific protocols (e.g. Cross-Dataset, Cross-PAI, Cross-Device, etc), which means that they are not yet ready to be used in uncontrolled conditions.

In a clear attempt to mitigate the aforementioned issues, some approaches that [5, 8, 12, 13] face the problems of data scarcity and the lack of common evaluation protocols propose a set of countermeasures in the form of new evaluation protocols, datasets and/or baseline methods to address generalised face-PAD. Besides, in [8] the authors present a complete taxonomy of current face-PADs and propose the largest aggregated dataset with a diverse set of protocols, focused on the generalisation issue: the so-called GRAD-GPAD benchmark.

On the other hand, face-PAD has been tackled from different perspectives for a few years now. One of the first works to raise this issue was [12], where the authors reformulated the traditional binary classification problem as an anomaly detection using texture-based features. With the rise of neural networks, different approaches have been applied based on the domain adaptation technique [14], learning discriminative features in the embedding space [14–16], using generative models [15], or even rethinking the formulation of anomaly detection using deep learning as in [3]. Although some of them present more realistic evaluation protocols, none of these approaches has been able to present production-ready performances in terms of usability and security.

Face-PAD is still an open problem and even though there is a shift towards interesting and powerful research directions to face the problem of generalisation [2, 3], there is a need to rapidly adapt to new scenarios and the dexterity of impostors when designing new and sophisticated attacks to fool the system. In this work, we extend the GRAD-GPAD framework by adding new protocols, metrics and visualisation tools focussed on helping and explaining generalisation on face-PAD.

2.2 | Bias

Bias analysis has recently become a central problem in decision-making systems. It is critical to assess whether our AI models are providing fair decisions. As pointed out in [17], there is an urge to improve AI fairness by tackling the main drivers of bias: skewed datasets and the algorithms. In this work, we provide a comparative study of the bias introduced by two state-of-the-art methods in the literature, centred on a detailed analysis of the built-in bias that each dataset incorporates due to the different capture conditions and target objectives, and then the demographic bias.

2.2.1 | Demographic bias

Variation in performance on sex, age or skin tone has recently become a controversial topic [17–20]. Unequal performance across demographic groups may affect the adoption of face recognition and face-PAD technology. Further, estimating the performance of our system following a different demographic ratio from the one that appears in the target domain can lead to unexpected errors in a real scenario. To the best of our knowledge, the first work to introduce some demographic bias analysis for face-PAD is [7], where the authors introduced the Cross-Ethnicity protocol over the CASIA-3Fe dataset. Three different ethnicities (i.e., Africa, East Asia, and Central Asia) are labelled. One ethnicity is used for training and evaluation, and the rest of the two ethnicities are used for testing. Additionally, this dataset contains 1607 subjects and four attack types (i.e., print attack, replay attack, 3D print and silica gel attacks). Unfortunately, this multi-modal dataset is not compatible with the GRAD-GPAD framework as the 2-D images are pre-processed with a 3-D reconstruction and density alignment on the detected faces. If the original RGB images are released, it could be tremendously interesting to categorise and label this dataset to be available in GRAD-GPAD. Nevertheless, thanks to the new demographic labels incorporated in GRAD-GPAD, we were able to design three novel protocols to evaluate the demographic bias: Cross-Sex, Cross-Age and Cross-SkinTone.

2.2.2 | Public dataset built-in bias

If we analyse all the available public datasets, we can easily observe that each one focusses on a small portion of the general face-PAD domain. This specificity in the domain covered by most of the datasets can be observed in different scenarios: (i) some datasets focus on a single type of presentation attack (e.g. in 3DMAD, HKBU(s) and CSMAD, we can only find Mask PA), (ii) others focus on the study of different imaging devices (depth/NIR/thermal) like CSMAD, (iii) some of them try to simulate a certain scenario like unrestricted mobile accesses where the user holds the device (e.g. Replay-Mobile or OULU-NPU), or a webcam setting, where the user stands in front of a fixed camera (e.g. Replay-Attack or SiW), or even a stand-up
scenario where users are registered further away from the camera (e.g. UVAD) etc.

In [21], Torralba et al. proposed a very interesting, simple experiment to reveal how easy it is to classify images from different object recognition datasets. They propose an experiment called ‘Name The Dataset’. The goal is to guess which group of images correspond with each dataset by just looking at some examples of each one. Intrigued by Torralba et al.’s experimentation, we trained a very simple classifier to play ‘Name The face-PAD Dataset’. We randomly sampled images from 13 face-PAD datasets and trained a 13-way linear SVM classifier on top of the feature extractor proposed in [12] that concatenates the quality measures (IQM) introduced in [22, 23] to extract features using the whole image or the cropped face area.

As depicted from Table 1, we obtained fairly good accuracy considering the simplicity of the classifier, surpassing by a large margin the human accuracy obtained in our lab (Check Appendix 6.1). Despite the best efforts of the community, these results indicate that individual datasets have a strong built-in bias. The rationale behind this bias originates on the experimental setting as follows: each dataset has different objectives and evaluates the performance of the models in different scenarios (e.g. using only mobile devices, outdoor/lab environments, synthetic/natural illumination, using a simulated onboarding setting etc). Even if we try to incorporate every possible scenario, we find that the biases are still present in some form.

Figure 1 shows a clearly pronounced diagonal of the confusion matrix in two experiments: (1) using the whole image and (2) using only the face region. From the first experiment, we can analyse the classification ability from a holistic perspective, taking into account the background of the image. This lets the classifier learn some features from the capture environment which is meant to help the classifier perform really well. On the other hand, the second experiment analyses how well the dataset classifier performs when isolating a specific object of interest (in this case, the face region). Interestingly, the bias is still there. Furthermore, we can observe the relationship between the accuracy of the classifier on each dataset and the intra-dataset variability. The takeaway message of this experiment is clear: if we want to have a model that generalises in the real world, we cannot train and evaluate it in individual datasets. There is a

| TABLE 1 | Computer plays Name the face-PAD Dataset with a quality-based approach with two configurations: using whole image and using only the face region. The experiment has been repeated using all PAIs and filtering each one individually. |
|---|---|
| | Accuracy (%) | Accuracy (%) |
| | Whole Image | Face Region |
| Only real | 0.85 ± 0.04 | 0.70 ± 0.05 |
| Only print attacks | 0.95 ± 0.01 | 0.68 ± 0.07 |
| Only replay attacks | 0.94 ± 0.02 | 0.69 ± 0.07 |
| Only mask attacks | 0.93 ± 0.03 | 0.88 ± 0.09 |
| Real and all attacks | 0.84 ± 0.03 | 0.68 ± 0.06 |

Abbreviations: PAD, presentation attack detection; PAI, presentation attack instrument.

The classifier was tested with a stratified K-Fold cross-validator (with K = 5) preserving the percentage of samples for each class.

The classifier is classified with very high accuracy as it has only eight identities and one attack, while SW is classified with less accuracy due to its higher capture variability, higher number of identities and attacks.

---

---
high chance that our model overfits the dataset, not solving the real challenge of face-PAD, that is, generalisation. In this work, we extend the GRAD-GPAD framework to cover most of these limitations and provide a starting point to design trustworthy evaluation frameworks for face-PAD approaches that covers not only research improvements but also production-ready environments.

2.3 | Aggregated datasets

Although there is no common categorisation among all the datasets, it seems to be clear that the community is aware of the importance of data representativeness in the evaluation of face-PAD models, and every day we find more and more investigations that propose new datasets or different combinations of inter-dataset evaluations. Recent works [8, 12, 13] incorporate aggregated datasets to further expand the scope of their evaluations. In [12], the authors introduced an aggregated dataset composed of three publicly available datasets (Replay-Attack [24], Replay-Mobile [25] and MSU-MFSD [23]) while in [13], the authors proposed another dataset combining four different datasets (CASIA [26], Replay-Attack, MSU-MFSD and Oulu-NPU [27]). Both works tackle the problem of generalisation from an anomaly detection perspective. However, despite the fact that using larger datasets leads to better performance, it is not yet sufficient to solve the generalisation issue in a real production environment. In [8], the authors introduced the largest aggregated datasets so far, the GRAD-GPAD framework. In this evaluation framework, the authors incorporated the 10 most representative face-PAD datasets (CASIA, Replay-Attack, 3D-MAD [28], MSU-MFSD, Replay-Mobile, HKBU v1 [29], Oulu-NPU [27], Rose-YouTu [14], SiW [15], and CS-MAD [30]), presenting a set of protocols to evaluate generalisation from different angles. In this work, we extend GRAD-GPAD, increasing the data representation by adding three new datasets (UVAD [9], HKBU (v2) [10], and SiW-M [4]).

2.4 | Anticipating new approaches

The fast incorporation of face recognition systems in real products poses new challenges. In this context, attackers can easily invest time to learn new sophisticated ways to fool the systems by just acquiring one of these devices that incorporate FR technology, resulting in new unknown PAIs that appear so often. PAI variation over time is a key aspect that was recently introduced by [31]. In this work, the authors hypothesise that in a realistic scenario we would expect both unseen and unknown PAIs to appear through time and propose a continual meta-learning approach to tackle this problem. Face-PAD approaches should consider this aspect to fully understand the behaviour under such a common real scenario. The authors of [31] extended GRAD-GPAD with a continual few-shot protocol to evaluate face-PAD approaches using a meta-learning approach [32–34] and to measure the impact of catastrophic forgetting in a continual setting.

Learning through time without forgetting is a great quest of artificial general intelligence systems. Classical offline supervised deep neural networks struggle with building upon previous experience and usually rely on training from scratch or fine-tuning the model as new data is collected. The former suffers from the burden of increasing computational effort through time, whereas the latter comes under catastrophic forgetting that results in exponential loss of the retained knowledge. We refer to [35] for further details.

Meta-learning in the context of image classification tries to solve the problems of data scarcity and generalisation, trying to mimic the learning process of humans in two aspects: (1) using few samples to learn a new task and (2) rapidly adapting its knowledge to behave well under new scenarios. These two features, few-shot learning and rapid adaptation, make meta-learning to be on the spot of recent works, surpassing the previous literature by a considerable margin [32, 36, 37]. In this work, we provide protocols to take advantage of the GRAD-GPAD categorisation in new learning methodologies based on meta-learning [38] and lifelong learning [31].

3 | THE GRAD-GPAD EVALUATION FRAMEWORK

In this section, we describe the proposed new version of the GRAD-GPAD framework. For the sake of reproducibility and comparison, we have versioned the GRAD-GPAD as follows: (v0) for the initial version presented in [8], (v1) for the continual and few-shot learning extension proposed in [31] and (v2) for the present work. GRAD-GPAD v2 is aimed to help the research community address the problem of generalisation in face-PAD. This new version provides the following: (1) more tools and resources for reproducible research; (2) the categorisation of three new datasets following the existing taxonomy; (3) a novel labelling approach that enables us to design new evaluation protocols to assess the demographic bias and new learning methodologies such as lifelong learning; (4) new metrics on generalisation and demographic bias, and (5) new ways of graphically representing performance that makes it easier to compare methods at a glance. GRAD-GPAD v2 allows the evaluation of face-PAD algorithms from different angles, revealing hidden generalisation issues in traditional evaluations.

First, we present the framework in detail including the PAI categorisation and new demographic labelling. Then, new protocols, metrics and visualisation proposals are introduced.

3.1 | Reproducible research

This framework is publicly available as a Python package and it provides tools to reproduce our research and even extend some presented experiments. GRAD-GPAD v2 is extendable

https://github.com/alice-biometrics/gradgpad
and scalable by design. This will help researchers create new protocols, define categorisations over different datasets in order to make them compatible, and enable the addition of new datasets. We encourage the biometrics community to adopt GRAD-GPAD as an evaluation standard and to collaborate with the extension of its face-PAD. Face-PAD, by nature, is in constant evolution, and this framework can help us all to understand the real challenges that we are facing and anticipate the performance of our models in production-like scenarios.

3.2 The aggregated dataset

Previous versions of the GRAD-GPAD framework (v0 and v1) are composed of the 10 most representative RGB-based publicly available face-PAD datasets at the time of its publication: CASIA-FASD, Replay-Attack, 3DMAD, MSU-MFSD, Replay-Mobile, HKBU (v1), Oulu-NPU, Rose-YouTu, SiW and CS-MAD. The number of global identities is around 454 (there is no information about identities’ collision between datasets). The aggregate dataset contains 4548 real samples and 9655 attacks, with a wide range of Print, Replay and Mask P AIs. Three new datasets have been added for the extension provided with this work: UVAD, HKBU (v2), and SiW-M. With this extension, we increase the number of identities by a factor of 3 (909 new identities out of 1363 in total). Therefore, the number of samples also increases, reaching 6520 for bona fide presentations (43% more) and 27,395 for presentation attacks (183% more).

The GRAD-GPAD evaluation framework is focused on the most challenging scenario in assessing face-PAD software-based approaches, when data is captured using RGB-only sensors using passive approaches that avoid any challenge-response. Recent multi-modal datasets such as WMCA [39], CASIA-SURF [40] or CASIA-SURF CeFA [7] cannot be added due to data incompatibilities: images publicly available are a pre-processed version of the raw capture. WMCA samples are in grayscale format, and both CASIA-SURF and CASIA-SURF CeFA directly provide the segmentation of the face based on a prior depth estimation and density alignment of the detected faces. In the future, if raw data is publicly available, these three datasets would be important candidates to extend the GRAD-GPAD v2 benchmark.

3.3 Categorisation

For the sake of compatibility between datasets, in GRAD-GPAD v0 [8] the authors introduced a common taxonomy to categorise the different types of accesses represented in current datasets. In this work, the categorisations are extended, allowing, in this way, an in-depth analysis of the weaknesses and strengths of the algorithms. In addition, the three newly added datasets have been adapted to fit into this categorisation. The categorisation proposed in GRAD-GPAD v0 was based on three general categories: (1) Common Capture Device, that represents the variability of capture devices divided in three categories (Webcam, Mobile/Tablet and Digital Camera); (2) Common Lighting, based on illumination information already labelled as Controlled and Adverse conditions; and (3) Common Face Resolution, which categorises the access based on the size of the inter-eye distance of the detected faces, subdividing them into three types (Small, Medium and Large Face).

Face recognition systems can be attacked by a large or indeterminate number of potential presentation attack instruments (PAI). Each of these PAIs can be included in a PAI species categorisation that provides a semantic value and allows us to group the different types of attacks by common characteristics (e.g. High-Quality Replay PAI species groups every PAI displayed on a screen with a resolution over 1080 pixels). As pointed out in the ISO/IEC 30,107-3:2017, it is very difficult to have a comprehensive model of all possible PAIs, so it could be impossible to find a representative set of PAI species for the evaluation. GRAD-GPAD v2 introduces two new hierarchical levels of categorisations for PAI species in order to increase the representativeness of PAI species and to allow more in-depth analysis of face-PAD approaches.

- Coarse-grained PAI species, which represents the five coarse-grained PAIs available on the aggregated dataset (Print, Replay, Mask, Makeup and Partial);
- Fine-grained PAI species, which represents the fine-grained PAIs available on the aggregated dataset: (a) Print (Low, Medium and High Quality); (b) Replay (Low, Medium and High Quality); (c) Mask (Paper, Rigid and Silicon); (d) Makeup (Cosmetic, Impersonation and Obfuscation); and (e) Partial (Periocular, Upper Half, Lower Half, Paper Glasses and Funny Glasses).

Additionally, we introduce a holistic categorisation to define the Presentation Attack Scenario (PAS). We define three PAS for GRAD-GPAD v2: PAS I represents a scenario where spoofers have the freedom to try to impersonate an identity completely (as with a stolen cell phone or in an isolated access environment); PAS II represents partial identity impersonations, where attackers only use a part of someone else’s identity; and finally PAS III, where users try to hide their identity without impersonating anyone in particular. In Figure 2, we present a detailed visual representation of this specific categorisation.

This comprehensive categorisation by PAS allows us to consider what scenarios we should take into account both when training and evaluating the model. This decision will be determined by the algorithm parametrisation and the target environment where we want to evaluate our face-PAD system. For example, if you are evaluating a system that will be used in border controls with identity alerts, it may be interesting to study the behaviour of identity obfuscation (PAS III) as certain people may deliberately not want to reveal their identity. However, it makes no sense to evaluate this type of attack in a mobile scenario, where the face-PAD system works in conjunction with a facial verification system. On the other hand, this is an interesting attack type to test in a controlled environment.
hand, PAS II contains presentation attacks and impersonation features at the same time, so including these types of attacks for training should be decided as a per case basis, depending on the type of face-PAD algorithm and deployment scenario.

3.4 | Demographic distribution

Unequal performance across demographic groups can potentially undermine public acceptance of face-PAD methods. Also, estimating the performance based on images with a different demographic distribution than the target users of the technology can lead to unexpected results in the operational scenario. Therefore, it is important to know whether or not demographic groups are well represented and understand if accuracy differences actually exist as well as the reasons for them. For that purpose, we have added an additional demographic categorisation to GRAD-GPAD v2 by introducing additional labelling to all the identities: Sex, Age and Skin Tone.

Figures 3–5 show the distribution of demographic attributes over all the 13 datasets of GRAD-GPAD v2. The use of public datasets, normally captured by research centres or universities, leads to obtaining demographically skewed data. The aggregation of the 13 datasets in GRAD-GPAD helps to alleviate the scarcity of some demographic groups, but the skewness of the aggregation might not be solved. Let us see it in more detail.

3.4.1 | Sex

Amongst the 1124 labelled identities, 61% are male and 39% are female. The aggregated distribution reflects a higher number of male identities in all datasets except SiW-M (see Figure 3). In fact, without adding SiW-M (presented in 2019), the skewness would be much higher, since the number of SiW-M identities is the highest of all GRAD-GPAD datasets and this is the only dataset with more female than male samples. Undoubtedly, Figure 3 shows the need to capture more balanced datasets. Furthermore, aggregated data as in the GRAD-GPAD framework is crucial if we really want to palliate the lack of representative data in public available datasets, but regarding specific demographic groups, additional resampling might be needed.

3.4.2 | Skin tone

One of the most controversial sources of bias originates from the skin tone diversity around the globe. It is not surprising that collected datasets do not equally represent the different skin tones. In Figure 4, we show how unbalanced the public datasets are with regard to the skin tone, by grouping them into 6 categories8: (1) Light Yellow, (2) Medium Yellow

---

8We have agreed to use the same Skin Tone Enumeration as the IJCB: https://www.nist.gov/system/files/documents/2017/12/26/readme.pdf
Brown, (3) Light Pink, (4) Medium Pink Brown, (5) Medium Dark Brown and (6) Dark Brown. As expected, many of the datasets that are normally used as benchmarks for comparing algorithms either have large data imbalance or they have not represented the different ethnicities and skin tones. For example, the Oulu-NPU has become one of the most widely used datasets, thanks to its challenging generalisation protocols. However, despite including 55 identities, 52 correspond to Light Yellow and Medium Yellow Brown, and the remaining three to Medium Pink Brown. Therefore, when comparing algorithms with the Oulu-NPU benchmark, we have to consider this demographic skewness and we need to be aware that the comparison is not fair and demographically reliable if the target users in the production scenario are not mainly Asian. The aggregation in GRAD-GPAD solves the scarcity problem in some skin-tone groups (i.e. ‘TOTAL’ in Figure 4) but still presents a distribution less homogeneous than, for example, SiW or MSU-MFSD. So, if skin tone is a critical issue in deployment, an appropriate resampling might be needed as well.

3.4.3 | Age

GRAD-GPAD v2 incorporates a coarse-grained age categorisation of the different samples into three ranges: (1) young, identities between 18 and 25 years (there are no images for kids or babies), (2) adult, people between 25 years and 65, and (3) senior, people older than 65 years. Similar to skin tone, age is unevenly distributed, with more adults than young people and very residual, or even zero, seniors for all the 13 datasets. Figure 5 shows the shortcomings of the datasets when it comes to representing the different ages of potential users. Despite the fact that the SiW dataset is one of the best balanced datasets in terms of skin tone distribution, it is not the same situation when it has to do with age distribution: SiW does not have any senior user.

As discussed in the previous sections, each dataset represents a small portion of the actual PAD scenario. Aggregating the datasets as it is done in GRAD-GPAD v2 not only helps to mitigate the bias coming from the data but also helps us understand the bias distribution within our training dataset, allowing us to incorporate compensations in the learning process and to improve future dataset captures. In Section 4, we use two state-of-the-art face-PAD approaches to analyse how demographic bias is affecting usability.

3.5 | Protocols

The unified categorisation added in GRAD-GPAD v2 brings the opportunity both to create novel protocols and to visualise the results from different perspectives. Also, the extended
GRAD-GP AD v2 dataset allows a better statistical significance of the results of previous protocols, leveraging their added-value for assessing face-PAD generalisation on current and future algorithms. In addition to former protocols presented in [8, 31] (Grandtest, Cross-Dataset, One-PAI, Cross-PAI or Unseen-Attack, Cross-Device, Cross-FaceResolution, Cross-Condition and Lifelong-Learning), two novel groups of protocols are introduced in this version: the Leave-Other-Dataset-Out (LODO) and the Demographics protocols.

3.5.1 | Leave other dataset out protocol

The LODO protocol is a variant of the Cross-Dataset protocol. We are used to comparing algorithms with the leave-one-out protocol, where the named dataset is used for testing the generalisation of the models and the remaining datasets are used for training and tuning the algorithms. We also propose the opposite, the LODO protocol that uses only one dataset for training and tuning and then tests it using the rest of the datasets. What we want to show with this protocol is the gap that training in an isolated dataset and then deploying it in the real world involves. This protocol shows us the performance of our systems in a more global scenario, but perhaps the most interesting thing about this protocol is to show the representativeness of the dataset evaluated. We show that a common scenario in traditional evaluations is that we present impressive results when we evaluate our approach using our dataset, but we fail to move these models into different domains.

3.5.2 | Demographic protocols

Using the proposed demographic labelling allows us to propose new protocols to evaluate the demographic bias in face-PAD: Cross-Sex, Cross-Age and Cross-Skin-Tone. These protocols evaluate the performance variation of each demographic group. We train two state-of-the-art models following the Grandtest protocol (we use the Train split for training and the Devel split for hyper-parameter search); however, we evaluate the performance of the models using a Test split filtered by Demographic label, that is, samples are taken from each class in the same proportion (same number of males and females). In this way, we can observe the impact of face-PAD approaches on a real scenario when training has been carried out with typical skewed datasets and optimisation techniques are democratic with the samples. The aim of these protocols consists of testing whether a particular face-PAD approach is agnostic to the different demographic groups or not.
3.6 | Meta-learning and lifelong learning

GRAD-GPAD v2 enables the design of new protocols adapted to new approaches such as meta-learning and/or lifelong learning allowing the incorporation of the latest advances of the AI community. In this work, we propose a baseline protocol to assess the performance of the model in a continual meta-learning setting, where samples for training are fed to the model sequentially following a non-independent and identically distributed distribution. In this work, we greatly expand GRAD-GPAD v1 to cover a wider scope of face-PAD evaluations: generalisation, bias, finer PAI species categorisation and production settings, facilitating its use for testing these novel approaches. An example of how to take advantage of these protocols to train and evaluate face-PAD approaches from sequential tasks is shown in Appendix 6.4.

3.7 | Metrics

In this work, we propose novel metrics to complement the existing standard metrics (see Appendix 6.3). First, we review the importance of considering how we should evaluate actual applications. Second, we present generalisation and demographic metrics. Finally, we propose two novel visualisation tools.

3.7.1 | The importance of analysing at the attack instrument level

When we want to compare the performance of face-PAD methods, especially in a research stage, we are interested in finding out not only which one is better overall but also what the performance is under different attacks. Two methods with identical $APCER$ and $BPCER$ metrics can have very different behaviour with different attack instruments. For instance, method A may be very good for $Print$ PAIs and not so good for $Mask$ PAIs and method B may be very good for $Replay$ PAIs but very bad for $Print$ attacks. It is crucial to understand these differences when deploying a face-PAD solution as it will allow us to choose or combine the appropriate methods (e.g. in a final ensemble-like system) to maximise the performance in our setting. In Section 4, we centre our analysis at the PAI species level (and even subPAIs) using the proposed fine-grained sub-categorisation in Figure 2.

3.7.2 | Generalisation metrics

As noted in Sections 1 and 2, generalisation is the main challenge in the field of face-PAD today. In this work, we propose two novel generalisation metrics. These metrics may be used in
the specific protocols intended to evaluate the generalisation of the algorithms with several variants (e.g. Cross-Dataset, LODO, Cross-Device, Cross-PAI). The proposed generalisation metrics are given below:

- **HTGER**: Half Total Generalisation Error Rate. Used in generalisation protocols and defined as the average of all ACER computed for all the variants of the protocol (Equation 1). For instance, in a Cross-Dataset protocol the HTGER would be the average of the ACER computed for each dataset.

\[
HTGER = \left( \frac{1}{N_{GP}} \right)^{\frac{1}{2}} \sum_{i=1}^{N_{GP}} ACER(GP_i) \tag{1}
\]

where \( GP \) represents any generalisation protocol and \( N_{GP} \) the total number of generalisation protocols to be evaluated.

- **WCGER**: Worst Case Generalisation Error Rate. It is defined as the maximum value of all possible ACER computed for all the variants of the protocol (Equation 2). Following with the same example given for the HTGER, the WCGER for a Cross-Dataset protocol will be equal to the worst ACER of all datasets. This is the most challenging metric by far and allows us to detect important vulnerabilities over generalisation protocols selected to evaluate a particular use case.

\[
WCGER = \max_{GP} (APCER_{GP}) \tag{2}
\]

Table 3 in Section 4 is an example of how these metrics can be used to compare two face-PAD approaches in terms of generalisation.

### 3.7.3 A metric for demographic bias

In the era of trustworthy AI, a face-PAD model can be considered fair if the errors are distributed similarly across different demographic groups. There are many ways to define fairness in automated decisions [41] but in this work we just focus on the well-known definition of Demographic Parity or Statistical Parity: a predictor is unbiased if the prediction is independent of the protected attribute (sex, race, religion etc.). The lack of Demographic Parity can be measured by the Statistical Parity Difference:

\[
SPD = Pr(Y = 1 | A = 1) - Pr(Y = 1 | A = 0) \tag{3}
\]

where \( Y \) represents a binary predictor and \( A \) represents the protected attributes of individuals (e.g. \( A = 1 \) for Female and \( A = 0 \) for Male).

In this work, we have adapted the \( SPD \) definition by measuring the statistical difference of \( BPCER \) between demographic groups. Moreover, given that the model can be set up for different working points depending on the deployment scenario, we extend the definition to account for a practical range of working points (or predictors). Consequently, the difference should be always positive to account for a possible change of polarity or biased decision depending on the working point. The Demographic Bias Metric (DBM) for a specific Demographic Group (DG) is defined as follows:

\[
DBM = \frac{1}{|DG_{pairs}| |w_{points}|} \sum_{j,k \in [DG_{pairs}]} \sum_{w \in w_{points}} |BPCER(i, j) - BPCER(i, k)| \tag{4}
\]

where \( DG_{pairs} \) is the number of pairs we want the model to be accounted in. For instance, in the DG ‘sex’, only one pair can be measured: male-female, but in DG ‘age’, we might want to measure bias among young versus adult, young versus senior and adult versus senior. The DBM will yield an average of the pair-wise bias. \( w_{points} \) is the set of working points where we want to measure bias and \( BPCER(i, j) \) is the BPCER evaluated in the demographic group \( j \) with working point \( w_i \).

### 3.8 Visualisation proposal

The metrics provide us with a lot of information, but sometimes with such a large volume of experiments, the analysis becomes a very complicated task. As discussed above, it is helpful to compare systems at a finer level of detail to complement standard metrics analysis. For these reasons, we propose two new visualisation tools to compare the algorithms: (1) the ‘PAD-radar’ shows us information related to the model’s generalisation and PAI behaviour and (2) the ‘bias-percentile’ to better understand the demographic bias of the systems that we evaluate. The ‘PAD-radar’ is based on a radar chart representation that presents the metrics for the different subcategories evaluated. For example, in the case in which you want to visualise the error of a system for each of the available PAIs, we propose to present the errors—\( APCER \) at a fixed working point (e.g. \( BPCER = 10\% \))—at each vertex corresponding to an evaluated PAI (see an example in Figure 6). Besides, these plots are very useful for visualising generalisation protocols where we obtain several results. For instance, when using a Cross-Dataset protocol, all datasets appear in vertices to evaluate the influence of each sub-protocol on the algorithm (see Figure 7).

Additionally, we propose to visualise the metric defined in Section 3.7.3 using a ‘bias-percentile’ representation derived from the statistical distribution of genuine scores and attacks scores but segmented by demographic groups. With this visualisation, we can easily study a face-PAD \( BPCER \) for all demographic groups and range of working points, revealing also any polarity change in the demographic bias across them (see Figure 9).
4 | EXPERIMENTS AND GENERALISATION

In this section, we present the results of several protocols and scenarios described in Section 3. The proposed metrics and visualisation illustrates the benefits of GRAD-GPAD v2, in particular, the fine-grained categorisation and the new protocols. First, we introduce two baselines selected for face-PAD generalisation tests. Then, we show their performance on former and novel protocols, including a final demographic bias study.

Two popular approaches have been selected as the baseline. The so-called Quality approach [12] computes hand-crafted features based on quality evidences. Authors proposed to obtain a 139-length feature vector from the concatenation of the quality measurements proposed in [22, 23]. The second method is based on the Auxiliary deep-learning approach proposed in [2]. We use an adaptation of this approach presented in [31]. We evaluate the baselines from the perspective of a real system in production. Taking this into account, we selected a use case where the main attack is impersonation. Therefore, we only use two Presentation Attack Scenarios: PAS I and II (see Figure 2 in Section 3). All genuine and PAS I samples will be used for training and testing. Since PAS II samples show a large portion of the real face, they will be included just for testing. Otherwise, actual face pixels would be used both as genuine and impostors, misleading the optimisation.

4.1 | Grandtest protocol and the importance of fine-grain analysis

The Grandtest protocol summarises the model's performance from all available data. This protocol is very useful in the early stages of fitting parameters and it also provides a very useful overview for ruling out certain approaches. However, many details are lost in the metrics we use to compare the approaches. Unfair comparisons can lead to erroneous conclusions that hinder our research, so it is more than advisable to analyse the results at a fine-grained level, as it provides a deeper understanding of the model's behaviour and makes it easier to take decisions to improve our models.

For some time now, HTER, ACER and APCER metrics (fixing a specific BPCER working point) have been used to compare face-PAD algorithms. Table 2 shows the general performance of the Quality and Auxiliary approaches. Results are presented using the coarse-grained PAIs (Print, Replay, Mask, Partial, Makeup) and the fine-grained PAIs categorisation (see PAS I and II in Figure 2).
We show in Table 2 the influence of the PAI granularity (see how it affects the metrics in Appendix 6.3). Although metrics such as ACER and APCER bring us very valuable information about performance, sometimes they can mask information about generalisation of different specific attacks. The results show that both Quality and Auxiliary methods suffer a dramatic drop in performance with certain specific attacks. But, what are the PAIs that degrade the face-PAD algorithm and on which should we focus our research?

In this work, we propose a detailed performance visualisation for each of the evaluated PAIs using the proposed ‘PAD-radar’ plot such as the one presented in Figure 6. With this granularity and our novel representation, we can visualise the data separately and see what the generalisation weaknesses of the evaluated models are. This valuable information will lead to a better understanding of the models, which is essential to tackling the generalisation problem. Additionally, we publish a video to show the evolution of this ‘PAD-radar’ along the working points on the reproducible research resources.1

1https://github.com/acostapazo/gnudgpad/blob/master/samples/pad_radar.gif

![Figure 7](https://github.com/acostapazo/gnudgpad/blob/master/samples/pad_radar.gif)

**Figure 7** Fine-grained results for Cross-Dataset protocol

| PAI Granularity         | Approach | ACER (%) | APCER@BPCER = 10% (%) |
|-------------------------|----------|----------|-----------------------|
| Coarse-grain (5 PAI     | Quality  | 36.26    | 73.17                 |
| species)                | Auxiliary| 28.43    | 68.30                 |
| Fine-grain (15 PAI      | Quality  | 59.60    | 100.0                 |
| species)                | Auxiliary| 55.35    | 90.00                 |

Table 2 Results for Grandtest with different PAI granularity. HTER values are 17.75% for quality and 8.68% for auxiliary.

Some interesting conclusions can be drawn from Figure 6. We can observe the influence of testing unseen PAIs (PAS II). Partial Periocular and Lower-Half PAIs obtain very bad results in both Quality and Auxiliary. This is to be expected since most of the region corresponds to a genuine user, and therefore, both holistic algorithms classify it as such. On the other hand, we see that the Auxiliary seems to be able to better generalise for Partial PAIs using the information learnt from Print-based attacks. Partial Upper-Half PAIs (only present in Rose-
4.2 | Generalisation protocols

In this section, ‘PAD-radar’ visualisation, generalisation metrics and protocols help us evaluate some specific scenarios that are particularly relevant for the analysed case.

4.2.1 | Generalisation metrics

Table 3 summarises generalisation metrics for four generalisation protocols and shows as well a final result for the two face-PAD baselines. HTGER results, which represent an average metric for all generalisation protocol variants, show that the deep-learning-based algorithm (Auxiliary) generalises much better than the handcrafted-method (Quality), but these tough generalisation protocols show clear weaknesses also for the DNN-based baseline.

In addition, if we consider WCGER results, which represent the worst error rate over all generalisation protocol variants, our two baselines collapse with 100% error. This metric brings to light the critical vulnerabilities of a system whose objective is to operate in uncontrolled and generalist environments. In any case, knowledge of the constraints of the deployment scenario would allow us to pay more attention to one or other protocol and even the plausible PAI species in them. In the following subsections, we present an in-depth evaluation of two generalisation protocols available in GRAD-GPAD v2: Cross-Dataset and LODO.

4.2.2 | Cross-dataset protocol

As discussed in Section 3, the Cross-Dataset protocol has traditionally been used to analyse generalisation. However, doing this analysis with only two, three or four datasets limits the conclusions we can draw. In addition, there is no standardisation on which datasets we are to analyse, so in many cases, it is difficult to have a fair comparison between different approaches.

Figure 7 shows the result for Cross-Dataset protocols. High APCERs for Quality in every protocol is very representative of the fact that the algorithm does not generalise well. This result is consistent with the 93.24% HTGER shown in Table 3. Likewise, the Auxiliary method obtaining a HTGER of 55.81% presents better results in certain datasets but performs poorly in other ones, which can be critical for deployment environments.

Moreover, Figure 7 shows which datasets are consistently more challenging for both the evaluated face-PAD approaches. For the Quality approach, almost every Cross-Dataset protocol reaches 100% showing critical model overfitting. Although the Auxiliary approach also presents very high overfitting, this is not as evident as that of Quality. Nevertheless, there is a method where the trained model obtains very good results, the Cross-Dataset for HKBU v1. It seems that mask samples learnt from mask dataset samples (3DMAD, CSMAAP, SIW-M and especially HKBU v2) and enabled the system to generalise in this concrete sub-protocol. Additionally, we have to consider that HKBU v1 has very few samples and there is no variability. So, it does not seem prudent to determine whether or not an application can work on real scenarios with only a few Cross-Dataset protocols. These results evidence the necessity of evaluation approaches from different points of view before taking production-oriented research decisions.

4.2.3 | LODO protocol

The LODO protocol allows us to analyse how the algorithms behave when trained with limited data. The LODO ‘PAD-radar’ shows in each vertex the generalisation performance when the algorithm is trained with only a specific dataset and tested with all the others. So, each vertex plots a slightly different potential deployment scenario.

| Approach | Generalisation Protocol | HTGER (%) | WCGER (%) |
|----------|-------------------------|-----------|-----------|
| Quality  | Cross-dataset           | 93.24     | 100.00    |
|          | LODO                    | 98.83     | 100.00    |
|          | Cross-device            | 99.88     | 100.00    |
|          | Cross-PAI               | 87.89     | 100.00    |
|          | TOTAL                    | 94.96 ± 5.53 | 100 |
| Auxiliary | Cross-dataset           | 55.81     | 100.00    |
|          | LODO                    | 98.88     | 100.00    |
|          | Cross-device            | 81.41     | 91.67     |
|          | Cross-PAI               | 73.16     | 100.00    |
|          | TOTAL                    | 77.32 ± 17.89 | 100 |

Abbreviations: PAI, presentation attack instrument; HTGER, half total error rate; LODO, leave-other-dataset-out.
The results of the LODO protocol (see Figure 8) show that, given the built-in bias of each dataset (discussed earlier in Section 2), it is not possible to generalise by training with one isolated public dataset since each one of them only represents a small portion of the real world problem. We encourage face-PAD researchers to use these generalisation protocols or even extend them to obtain more information both to improve the generalisation results and to design new datasets to alleviate the lack of representativeness in the studied domain. As mentioned in Section 3, this evaluation framework is available as an open-source framework and any contributions and updates with new datasets and protocols are welcome.

4.3 | Demographic protocols

In these protocols, we try to detect if there is any demographic bias in the baseline models or not, as introduced in Section 3.7.3 and 3.8. So, we plot the BPCER curves for different demographic groups and the common APCER curve. For a specific working point, a BPCER difference among the represented demographic curves indicates a demographic bias. We focus the analysis on three demographic groups (Sex, Age and Skin Tone) in terms of usability, evaluating BPCER fixing a working point range from APCER 5% to APCER 15% (see grey area in Figure 9).

4.3.1 | Cross-sex protocol

Figure 9 shows that the Quality method has a worse overall performance than that of the Auxiliary method. However, the bias present in the Quality method is less evident than in the Auxiliary method. In the former, the male and female curves overlap in a large part of the working range, separating only in the points of very little usability (high rejection of genuine). In the Auxiliary method, on the contrary, male and female curves are separated in the whole range of possible working points. The performance in the female group is always worse. There is, therefore, a very significant bias. The higher bias of the Auxiliary method can be

---

**FIGURE 8** Fine-grained results for LODO protocol. LODO, leave-other-dataset-out
explained by the nature of the face-PAD algorithms analysed. The Quality method is based on designed image quality characteristics (not automatically learnt), and therefore, they do not seem to be much penalised by an unbalanced training set of males and females. However, the Auxiliary method is based on deep learning and thus, very dependent on the skewness of the demographic distribution present in the training sets, which, as we saw in Figure 2, is unbalanced with a predominance of male samples.

Our visualisation proposal allows us to observe the bias over all working points. In addition, this visualisation can be very useful to check if by balancing the training data, it is possible to mitigate the bias in testing. Figure 10 shows the result of training the Auxiliary method on sex-balanced data. We can see how by balancing the training data, the curves are much closer in both sets. This seems to confirm the hypothesis that an unbalanced distribution in training negatively affects the bias of the Auxiliary method.

Table 4 shows the Demographic Bias Metric (DBM) for the sex attribute in the range of working points that yield APCER from 5% to 15% (includes the EER). The DBM for Quality and Auxiliary summarises the observed behaviour in Figures 9 and 10.

4.3.2 | Cross-skin-tone protocol

In this case, both methods present bias (see Table 5). However, the behaviour of the Quality method is clearly different from that of the Auxiliary method (see Figure 11). In the Quality approach, the three curves, brown, pink and yellow, are slightly separated in a large part of the range. The worst performance is found in the Yellow skin tones, while the best performance is found in the brown tones. It is not clear why the designed image quality characteristics favour the generalisation of some skin colour over others, but this visualisation and metric clearly disclose the problem. On the other hand, in the Auxiliary method, the performance for Yellow and Pink skin tones is quite similar, while for Brown tones it is clearly worse across the range. Again, as in the previous evaluation, we can appreciate a greater bias due to training in the case of Auxiliary than in the case of Quality. Even when brown and yellow skin samples have roughly the same total number of training samples (Figure 4), Brown skin appears in less scenarios (8 vs. 12) and Yellow skin appears more demographically mixed with Pink skin, so it is quite likely that the DNN favours more characteristics to tell apart these two demographic groups over the Brown-skin one.

4.3.3 | Cross-age protocol

Here again, we see a different behaviour for Quality and Auxiliary methods (see Table 5 and Figure 12). In the case of Quality, the performance of the young age group is worse than that of the adult and senior groups in the whole range of working points. In the case of the Auxiliary method, it is the adult group that performs the best throughout the range, while the senior and young groups behave the worst for thresholds between 0.1 and 0.6.
TABLE 4 Sex demographic bias metric results for selected working point range APCER 5% – APCER 15%

| Approach         | DBM(Sex) |
|------------------|----------|
| Quality          | 0.77     |
| Auxiliary        | 8.37     |
| Balanced auxiliary | 0.68   |

Abbreviations: APCER, attack presentation classification error rate; DBM, demographic bias metric.

TABLE 5 Age and skin tone demographic bias metric results for selected working point range APCER 5% – APCER 15%

| Approach         | DBM(Age) | DBM(SkinTone) |
|------------------|----------|---------------|
| Quality          | 14.24    | 8.38          |
| Auxiliary        | 2.46     | 11.13         |

Abbreviations: APCER, attack presentation classification error rate; DBM, demographic bias metric.

Better overall performance of the adult group in the Auxiliary method is aligned with the dependence on the number of training samples observed in the previous experiments. The adult group has the highest number of samples in training. As we saw in the previous section, the Quality method is based on the analysis of designed (not learnt) quality characteristics. In this case, the worst performance for the young group seems to come from the selection of characteristics themselves. It is quite likely that wrinkles and other face marks that are more common in adults and seniors than in young people were better represented by the quality metrics, so SVMs place a more discriminative boundary between genuine and attackers in those regions of the feature space.

With this detailed visualisation of the demographic protocols, we can verify that, in general, data-hungry approaches with millions of free parameters will find more ad hoc features in demographic groups that are better represented in the training sets, so they are prone to suffer due to demographic bias and perform better on those groups. However, for approaches with programmed features and/or much lower number of free parameters for feature extraction and classification, the demographic bias has less dependence on unbalanced distributions (as long as the samples are enough to avoid overfitting) and can be traced back to the design of the features.

5 | CONCLUSION

In this work, we present a new version of GRAD-GPAD, a holistic evaluation framework for face-PAD. The benefits of GRAD-GPAD v2 include the following: (1) the largest aggregated benchmark with most publicly available datasets following an unified taxonomy, (2) new protocols covering realistic settings, including demographic labelling to evaluate different sources of bias, (3) novel metrics to understand the
FIGURE 11  Skin tone percentile-bias visualisation for both baselines. APCER, attack presentation classification error rate; BPCER, bona fide presentation classification error rate.

FIGURE 12  Age Percentile-bias visualisation for both baselines. APCER, attack presentation classification error rate; BPCER, bona fide presentation classification error rate.
generalisation of face-PAD models, (4) a detailed representation of the results to visualise the expected behaviour of the models and (5) reproducible material to encourage the adoption of GRAD-GPAD in the community to fairly evaluate the different approaches. The information provided by GRAD-GPAD v2 enables the analysis of explainability of face-PAD models, which is crucial when deploying these types of models.

For the first time, we analyse the demographic bias of face-PAD models for Sex, Age and Skin Tone. In this work, we conducted these experiments using an aggregation of 13 public datasets and found that they are prone to generating biases in both evaluation and training, so they need to be improved and expanded. GRAD-GPAD v2 bridges the gap of the lack of representativeness of state-of-the-art works and moves a step towards fair evaluations between methods not only from the perspectives of the different instruments used to perform the attacks but also considering realistic settings in production.

ACKNOWLEDGEMENTS
This research was funded by Doctorados Industriales (Agencia Estatal de Investigación/AEI/10.13039/S01100011033) and also by the Xunta de Galicia (Centro de investigación de Galicia accreditation 2019-2022) and the European Union (European Regional Development Fund - ERDF).

ORCID
Artur Costa-Pazo https://orcid.org/0000-0001-5791-3482

REFERENCES
1. Lu, C., Tang, X.: Surpassing human-level face verification performance onlfw with Gaussian face. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI-15), pp. 3811–3819. AAAI Press, Austin (2014)
2. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: binary or auxiliary supervision. In: Proceeding of IEEE Computer Vision and Pattern Recognition, Salt Lake City (2018)
3. Pérez-Cabò, D., et al.: Deep anomaly detection for generalized face anti-spoofing. CoRR (2019). abs/1904.08241
4. Liu, Y., et al.: Deep tree learning for zero-shot face anti-spoofing. In: CVPR, Long Beach (2019)
5. Costa-Pazo, A., et al.: Challenges of face presentation attack detection in real scenarios. In: Handbook of Biometric Anti-Spoofing, pp. 247–266. Springer International Publishing, Cham (2019)
6. Buolamwini, J., Gebru, T.: Gender shades: intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA. vol. 81 of Proceedings of Machine Learning Research (PMLR), pp. 77–91, New York (2018)
7. Li, A., et al.: Castia-surf cefa: a benchmark for multi-modal cross-ethnicity face anti-spoofing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Aspen (2020)
8. Costa-Pazo, A., et al.: Generalized presentation attack detection: a face anti-spoofing evaluation proposal. In: 2019 International Conference on Biometrics, ICB 2019, Crete (2019)
9. Pinto, A. et al.: Using visual rhythms for detecting video-based facial spoof attacks. IEEE TIFS. 10(5), 1025-1038 (2015)
10. Liu, S., et al.: A 3d mask face anti-spoofing database with real world variations. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1551–1557, Las Vegas (2016)
11. deFreitas Pereira, T., et al.: Can face anti-spoofing countermeasures work in a real world scenario? In: ICB 2013, pp. 1–8, Madrid (2013)
12. Nikins, O., et al.: On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. In: ICB 2018, pp. 75–81, Gold Coast (2018)
13. Xiong, F., Abdalmageed, W.: Unknown presentation attack detection with face rgb images. In: BTAS, Redondo Beach (2018)
14. Li, H., et al.: Unsupervised domain adaptation for face anti-spoofing. IEEE TIFS. 13(7), 1794–1809 (2018)
15. Jourabloo, A., Liu, Y., Liu, X.: Face de-spoofing: Anti-Spoofing Via Noise Modelling. arXiv preprint arXiv:180709968. 1(2), 3 (2018)
16. Li, H., et al.: Learning generalized deep feature representation for face anti-spoofing. IEEE TIFS. 13(10), 2639–2652 (2018)
17. Zhou, J., Schiebinger, L.: Ai can be sexist and racist — it’s time to make it fair. Nature. 559, 324–326 (2018)
18. Albiero, V., et al.: Analysis of gender inequality in face recognition accuracy. In: The 2nd Workshop on Demographic Variation in the Performance of Biometric Systems at WACV 2020, Aspen (2020)
19. Vicente Garcia, R., et al.: The harms of demographic bias in deep face recognition research. In: International Conference on Biometrics (ICB), Crete (2019)
20. Drozdowski, P., et al.: Demographic bias in biometrics: a survey on an emerging challenge. IEEE Trans. Techn. Soc. 1(2), 89–103 (2020)
21. Torralba, A., Efros, A.: Unbiased look at dataset bias. In: CVPR 2011, pp. 1521–1528, Colorado Springs (2011)
22. Galbally, J., Marcel, S., Fierrez, J.: Image quality assessment for fake biometric detection: application to iris, fingerprint, and face recognition. IEEE Trans. Image Process. 23(2), 710–724 (2014)
23. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE TIFS. 10(4), 746–761 (2015)
24. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG), Darmstadt (2012)
25. Costa-Pazo, A., et al.: The replay-mobile face presentation-attack database. In: BioSIG, Darmstadt (2016)
26. Zhang, Z., et al.: A face anti-spoofing database with diverse attacks. In: ICB 2012, pp. 26–31, New Delhi (2012)
27. Boulenafet, Z., et al.: OULU-NPU: A mobile face presentation attack database with real-world variations. In: 12th IEEE International Conference on Automatic Face & Gesture Recognition, Washington (2017)
28. Erdogmus, N., Marcel, S.: Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In: IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), Washington (2013)
29. Liu, S., et al.: 3d mask face anti-spoofing with remote photoplethysmography. In: Leibe, B., et al. (eds.) Computer Vision – ECCV 2016, pp. 85–100 Cham: Springer International Publishing, Amsterdam (2016)
30. Boulenafet, Z., Marjere, S., Molinari, A., Marcel, S.: Spoofing deep face recognition with custom silicone masks. In: BTAS, Redondo Beach (2018)
31. Pérez-Cabò, D., et al.: Learning to learn face-pads: a lifelong learning approach. In: 2020 IEEE International Joint Conference on Biometrics (IJCB), IEEE, Houston (2020)
32. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning, vol. 3, pp. 1126–1135, JMLR. org., Sydney (2017)
33. Nichol, A., Achiam, J., Schulman, J.: On First-Order Meta-Learning Algorithms. arXiv preprint arXiv:1803.02099. (2018)
34. Rajeswaran, A., et al.: Meta-learning with implicit gradients. In: Advances in Neural Information Processing Systems, pp. 113–124, Vancouver (2019)
35. De Lange, M., et al.: Continual learning: a comparative study on how to defy forgetting in classification tasks. arXiv preprint arXiv:1909.08383 (2019)
36. Bertaina, L., et al.: Meta-learning with differentiable closed-form solvers. In: International Conference on Learning Representations, New Orleans (2019)
37. Rusu, A.A., et al.: Meta-learning with latent embedding optimization. In: International Conference on Learning Representations, New Orleans (2019)
38. Eshtiatir, A.E., et al.: A meta-learning approach for custom model training. CoRR abs/1809.08346 (2018)
In Section 2, we study the face-PAD datasets’ built-in bias with a toy experiment based on a very interesting approach [21]. In addition, we performed a simple experiment with 8 lab researchers who are familiar with face-PAD. The experiment consisted of matching representative images from each of the datasets with their names. The accuracy for our team was over 65% of hit rate. Check out your classification skills with Figure A1 If you are familiar with face-PAD datasets you may have a good result. Check the solution here.10

APPENDICES

Name the face-PAD Dataset

In Section 2, we study the face-PAD datasets’ built-in bias with a toy experiment based on a very interesting approach [21]. In addition, we performed a simple experiment with 8 lab researchers who are familiar with face-PAD. The experiment consisted of matching representative images from each of the datasets with their names. The accuracy for our team was over 65% of hit rate. Check out your classification skills with Figure A1 If you are familiar with face-PAD datasets you may have a good result. Check the solution here.10

GRAD-GPAD previous protocols

In GRAD-GPAD v0 [8], seven protocols were presented, some based on classical assessments and others with new approaches. Then, GRAD-GPAD v1 [31] added a lifelong-learning protocol over the 10 available datasets (domains). Thus, the former versions of GRAD-GPAD contain a total of eight protocols: (1) Grandest: it evaluates face-PAD algorithms without any filter where all sets include all the previous categories for Train, Dev and Test subsets. (2) Cross-Dataset: it is based on training on one or several datasets and testing in others. (3) One-PAI: a protocol that evaluates face-PAD algorithms filtering only by PAI. (4) Cross-PAI (Unseen Attacks): a protocol that evaluates the performance under unseen PAIs. (5) Cross-Device (Unseen Capture Devices): protocols that evaluate how a face-PAD algorithm works under capture devices that were excluded on training and dev stages. (6) Cross-FaceResolution: a protocol to evaluate the role of face resolution in the task of detecting fake attempts. (7) Cross-Conditions: the underlying idea is to assess the performance of face-PAD algorithms under adversarial conditions. Two variants are proposed: Cross-Conditions-Test-Adverse and Cross-Conditions-Test-Optimal. (8) Lifelong meta-learning: create a large list of diverse tasks, where each task is generated using the common categorisation (dataset, device, PAI, illumination, etc). Datasets with greater variability generate a greater number of specific categories and therefore generate a greater number of disjoint tasks.

[ACP:TODO CABO, please review lifelong definition.]

Face-PAD Standard Metrics (ISO/IEC 30107-3)

In Section 4, we use the following standard metrics:

- **HTER**: Half Total Error Rate, defined as the point along the ROC or DET curve where the False Acceptance Rate equals the False Rejection Rate.

\[
HTER = \frac{FAR(\theta) + FRR(\theta)}{2}
\]

where \(\theta\) is the threshold for a working point (typically the EER).

- **APCER**: Attack Presentation Classification Error Rate. Proportion of attack presentations using the same PAI species incorrectly classified as bona fide presentations in a specific scenario. APCER is calculated for each evaluated PAI species using the Equation (6) where \(N_{PAI}\) is the number of attack presentations for the given PAI species, and \(Res\) takes value one if the presentation is classified as an attack presentation, and value 0 if classified as a bona fide presentation given a threshold \(\theta\).

\[
APCER_{PAI} = 1 - \left( \frac{1}{N_{PAI}} \right) \sum_{i=1}^{N_{PAI}} Res_i
\]

Then, to calculate the worst scenario, we use the Equation (7). We can observe that the number of PAIS \(P\), where \(P\) is the set of selected PAIS, is determinant to the APCER result. If we use a fine-grained PAI species (up to 15 in GRAD-GPAD v2) instead of classical coarse-grained PAI (5 PAIs) there is a greater probability that a PAI will penalise the worst case scenario.

\[
APCER = \max_{PAIS \in P} (APCER_{PAI})
\]

- **BPCR**: Bona fide Presentation Classification Error Rate. Proportion of bona fide presentations incorrectly classified as attack presentations in a specific scenario.

\[
BPCR = \frac{1}{N_{BF}} \sum_{i=1}^{N_{BF}} Res_i
\]

where \(N_{BF}\) is the number of bona fide presentations, and \(Res\) takes value one if the presentation is classified as an attack.
Name the face-PAD Dataset: Given three images from each of the 13 public face-PAD datasets. Can you match the images to the dataset? Datasets: CASIA-FASD, Replay-Attack, 3DMAD, MSU-MFSD, Replay-Mobile, HKBU, HKBU-MARS, OULU-NPU, Rose-YouTu, SiW, CS-MAD, SiW-M
presentation, and value 0 if classified as a bona fide presentation given a threshold ($\theta$).

- **ACER**: Average Classification Error Rate. This metric is defined as the average of the $\text{APCER}_{\text{max}(\text{PAI})}$ and the $\text{BPCER}$ for a pre-defined decision threshold.

**Lifelong learning Protocols**

In [31], the authors based on GRAD-GPAD v1 are to build a non-independent and identically distributed task setting drawing a continuous scenario using datasets as domains. In this work, with novel GRAD-GPAD v2 PAI categorisations, we propose a lifelong learning protocol for the evolving PAIs. This protocol simulates a real world scenario where the presentation attacks evolve along the time. We have designed a protocol by which the model is trained with specific PAI species in each iteration. First, low quality PAI species are trained, then medium quality and high quality. Finally, more recently introduced PAI species (Mask, Partial and Makeup) were added sequentially for training, and improving the performance of the model to these new PAIs. (Figure A2)

![Confusion matrix of the Auxiliary face-PAD approach trained in a continual meta-learning fashion for every PAS I and II. PAD, presentation attack detection; PAS, presentation attack scenario](image)