F. Boutros, N. Damer, M. Fang, M. Gomez-Barrero, P. Neto,
K. Raja, C. Rathgeb, A. Sequeira, and M. Todisco (Eds.): BIOSIG 2025,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2025

The Impact of Sample Quality Stratification on Generalization in Face Presentation Attack Detection

Eva María Benito Sanz[1]

Abstract: We investigate the impact of sample quality on cross-dataset generalization in Face Presentation Attack Detection (PAD). Using quality metrics derived from MagFace embeddings, we propose a stratification protocol that partitions training data into low, medium, and high quality tiers. Through experiments on two benchmark datasets, we show that training on different quality levels yields varied generalization performance, with low-quality samples in some cases promoting better cross-domain robustness. Our framework is modular, requires no additional labels, and can be combined with causal learning in future work.

Keywords: Face Presentation Attack Detection, Sample Quality, Domain Generalization, Quality Stratification, MagFace, Causal Learning

1 Introduction

Facial recognition systems have been an indispensable security feature in a variety of diverse application settings, such as mobile devices and border security (JRN16). This wide-spread application are mainly attributed to the developments in deep representation learning and the accessibility of large-scale facial data.

But such systems are also inherently susceptible to PAs, where the adversary manipulates the biometric data so that his/her identity cannot be recognized using biometric sensors, such as presenting printed images, video playbacks, or sophisticated 3D mask imitations, to impersonate the genuine users.

Face Presentation Attack Detection (PAD) has been recognized as an important countermeasure (RB20). It is particularly challenging due to the fact that current PAD methods perform remarkably well under controlled intra-dataset evaluations, with AUC scores up to 98% (LJL20). Nevertheless, such models often significantly degrade in performance under unseen acquisition conditions and novel attack types – with mean Half Total Error Rates (HTER) over 25% in cross-dataset evaluation (Wa22a). This non-generalization is a major impediment for secure face recognition in real deployment.

A Key approach to improve the generalization performance of targetaware multitask learning is to employ domain adaptation and domain generalization approaches (li2018learning). Domain adaptation, in general, is based on the premise that training and target domains have to share any amount of labeled data that is big enough to learn effectively, whereas

[1] Denmark Technical University, Computer Science and Engineering Master, s243313@dtu.dk

domain generalization often assumes access to domain annotated data, or is achieved by dedicated multi-domain training recipe (Wa22a). Neither of them completely resolves the impact of training sample quality to inform model generalization. Another possibility is variations in image quality, such as through motion blur, compression artifacts, low-light conditions, or sensor noise, which can inadvertently create domain boundaries (Be20). Such variations can hide important texture cues for the purpose of spoofing detection, or induce meaningless correlations, which degrade the learning performance.

In this paper, a generic architectural framework that includes sample quality in PAD systems is proposed. Our main contributions are:

1. A MagFace-based quality assessment method.
2. A direct quality stratification pipeline, which identifies the higher and the lower quality datasets.
3. Multi-quality Evaluation Dataset (MED) with causal learning approach using transferred knowledge.
4. A modular design that highlights the impact of quality-aware learning on PAD system robustness.

2 Related Work and Background

2.1 Evolution of Face PAD Techniques

Face Presentation Attack Detection (PAD) techniques have been developed in parallel to the advances in computer vision and machine learning. Initial approaches were based on hand-engineered feature descriptors including Local Binary Pattern (LBP) (MHP11), Histogram of Oriented Gradients (HOG), and spectral features extractors (Pea16). Even tough these approaches work fine in restricted settings, they were fragile for different lighting conditions, devices and advanced spoof attacks.

The application of deep learning greatly changed PAD by introducing end-to-end trainable models (LJL20). In the early works like DeepPixBis (GM19), pixel-wise supervision was proposed for spoof detection. Newer models also used supplementary cues including depth maps (At18) and remote photoplethysmography (rPPG) signals (Lea21b) to improve robustness. Other methods consider generalization using meta-learning (Sh20), few-shot adaptation (Lea21a), and adversarial training (WD19). Despite these progresses, transferring PAD models between domains is still a challenging problem with apparent performance drop, which drives us to further explore domain-agnostic and generalisable approaches in this paper.

2.2 Domain Generalization in PAD

Domain generalization is the task of learning representations that generalize well across unseen domains, without access to data from target domains (Wa22a). The recent state-of-the-art methods involve MixStyle that excites feature statistic disturbance (Zh21), DDG that learns disentangled domain-invariant and domain-specific features (Sh20), and MVDG that taps into multiple domain views (Lea21b).

A fundamentally notable progress is CF-PAD (Wa22b), that incorporates causal reasoning with counterfactual feature learn- ing. Without relying on explicit domain labels or extra model parameters, CF-PAD achieves a cross-dataset HTER of 17.3%, rendering it particularly attractive for deployment on resource-limited devices.

2.3 Face Quality Assessment Models

Most of the existing IQMs including NIQE and BRISQUE (MMB12) are not applicable for the problem of face biometrics because they are developed for universal IQA and they are not sensitive to the biometric-specific variances.

To fill this blank, the Face Quality Assessment (FQA) models have been proposed. such as SerFiQ (He21) that predicts the quality from the classification uncertainty; FaceQnet (He21) from human-annotated quality annotations; and SDD-FQA (Tea20) adding semantic concepts of the accessory facial descriptors. MagFace (Me21) is a holistic model, where the quality of facial quality is associated with the magnitude of the feature embedding, and it is capable of learning the two tasks concurrently.

3 Methodology

3.1 Quality Assessment Framework

We use MagFace as our quality assessment framework base and do not modify the original MagFace setting. For a face image I, we extract its MagFace embedding $\mathbf{f}(I) \in \mathbb{R}^d$ and the magnitude $|\mathbf{f}(I)|$ is used as our basic quality metric for the registered image.

This scale magnitude quantifies the quality of facial images in recognition property and is consistent with that in MagFace. We directly employ this criterion to evaluate the photographic quality of facial images:

$$Q(I) = |\mathbf{f}(I)| \tag{1}$$

We avoid incorporating any task-specific adaptation (e.g., for spoof detection, PAD, etc.), as the quality measure provided by MagFace directly provides a task-invariant yet meaningful signal correlated with the recognition performance. This enables us to maintain the

simplicity, effectiveness, and interpretability of the original method and to exploit its core concept of image quality.

3.2 Quality Stratification Protocol

We introduce a quality-based stratification scheme that labels face images into one of the three quality levels, namely, low, medium, and high based on the quality metric, Q, computed on MagFace embeddings. It is an indication of the embedding strength, and strong correlation with face image quality and biometric performance can be established (Me21).

For all the datasets, $Q(I)$ is first computed for all the images at which the 33rd and 66th percentiles of the distribution are obtained. These percentiles, labelled Q_{33} and Q_{66}, are subsequently used to split the data into 3 stratifications:

$$\textbf{Low quality (LQ)}: Q(I) < Q_{33}$$
$$\textbf{Medium quality (MQ)}: Q_{33} \leq Q(I) < Q_{66}$$
$$\textbf{High quality (HQ)}: Q(I) \geq Q_{66}$$

To avoid this issue and to make our experiments reproducible, we set these thresholds per dataset according to the initial distribution of quality. The final magnitude bins implemented for stratification are listed in Table 1.

Tab. 1: Quality thresholds used to stratify the datasets based on the magnitude of MagFace.

Dataset	**LQ**	**MQ**	**HQ**
CASIA-MFSD	< 27.5052	[27.5052, 29.2154)	≥ 29.2154
LCC-FASD	< 23.786	[23.786, 25.750)	≥ 25.750

This per-dataset stratification is aimed at handling changes in acquisition settings, and guarantees that quality groups correspond to relative differences within each dataset, making possible to carry out meaningful intra-dataset and cross-quality analysis.

3.3 Model Architecture and Training

Our training pipeline employs a ResNet-18 architecture initialized with pretrained ImageNet weights (He16). The final fully connected layer is replaced with a linear classifier that outputs two logits for binary classification (bona fide vs. attack). No pixel-wise supervision or auxiliary branches are used. The model is trained using the standard cross-entropy loss:

$$\mathscr{L} = \mathscr{L}_{\mathrm{CE}}(y, \hat{y}) \tag{2}$$

To improve generalization, data augmentation is applied during training using random horizontal flips, rotation, color jitter, and random resized crops. All images are normalized

using standard ImageNet statistics. Validation uses center crops and no random augmentations.

All models are trained for 10 epochs using the AdamW optimizer (LH19), with a learning rate of $\eta = 10^{-5}$, batch size $B = 32$, and weight decay of 10^{-4}. The learning rate follows a cosine annealing schedule via the OneCycleLR policy. Class imbalance is mitigated through dynamic weighting based on label frequency, and training is accelerated using automatic mixed precision.

Datasets are loaded using custom PyTorch `Dataset` classes that support three formats: folder-based (e.g., CASIA), hierarchical directory structures, and text-based image path lists (e.g., LCC). Each model is trained and validated on the same dataset split.

Cross-domain evaluation is performed by testing models trained on one dataset against another, using preprocessed quality-stratified subsets. All experiments are orchestrated using a custom automation script that handles training and evaluation across dataset combinations and quality levels (see the figure 1).

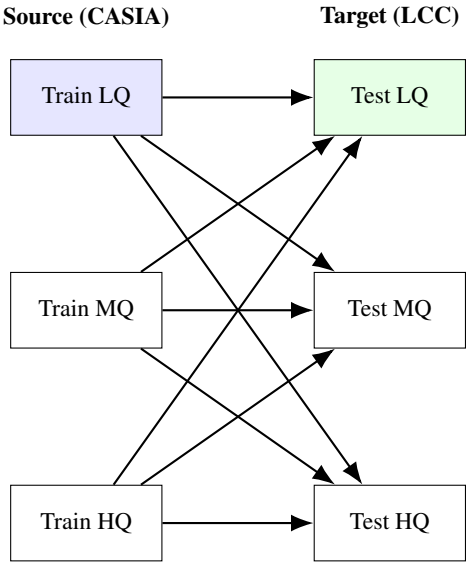

Fig. 1: Cross-domain evaluation protocol: models trained per quality level on CASIA are evaluated across all quality levels in LCC. This setup is repeated in reverse for LCC→CASIA.

3.4 Evaluation Framework

The evaluation framework follows a structured cross-domain and cross-quality protocol. For each trained model, performance is assessed on various quality strata and across different datasets. The evaluation strategy includes:

1. **Intra-quality**: Training and testing are conducted within the same quality tier.

2. **Cross-quality**: The model is trained on one quality tier and evaluated on a different tier, either within the same dataset or across datasets.
3. **Cross-dataset**: The model is trained on one dataset (e.g., CASIA) and evaluated on another (e.g., LCC).

All evaluations are performed on held-out test sets using center-crop normalization. Models are loaded in inference mode without any fine-tuning. Predictions are compared against ground-truth labels. Misclassified samples are optionally saved for visual inspection, and confusion matrices are automatically generated and saved.

Metrics. We report the following binary classification metrics derived from the confusion matrix:

- **FAR** (False Acceptance Rate): Proportion of attack samples incorrectly classified as bona fide.
- **FRR** (False Rejection Rate): Proportion of bona fide samples incorrectly classified as attacks.
- **HTER** (Half Total Error Rate): Mean of FAR and FRR.

$$\text{FAR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{3}$$

$$\text{FRR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{4}$$

$$\text{HTER} = \frac{\text{FAR} + \text{FRR}}{2} \tag{5}$$

All predictions and computed metrics are exported as CSV files for further quantitative analysis.

4 Experimental Evaluation

4.1 Quality Stratification Analysis

This section analyzes the impact of training on different quality strata, Low Quality (LQ), Medium Quality (MQ), and High Quality (HQ), on cross-dataset generalization. We conduct experiments where models are trained on a single quality tier from one dataset (either CASIA or LCC) and evaluated on the full-quality test set of the other dataset. This protocol allows us to assess how training quality influences the ability to generalize across domains.

4.1.1 Numerical Results

To isolate the effect of the training domain, we present results in two parts.

Training on CASIA, Testing on LCC. Table 2 summarizes the performance of models trained on CASIA. While the medium-quality model achieves the highest validation accuracy (97.08%), it does not yield the best generalization. The high-quality model achieves the lowest HTER (41.23%), suggesting that training on clearer samples may help in handling challenging cross-domain attacks. However, all models exhibit HTERs above 40%, highlighting the difficulty of generalizing to LCC.

Tab. 2: Models trained on CASIA and evaluated on LCC

Model	**Val Acc (%)**	**Train Quality**	**HTER (%)**
text_low_quality	96.81	Low	42.42
text_medium_quality	97.08	Medium	43.13
text_high_quality	91.89	High	**41.23**

Training on LCC, Testing on CASIA. Table 3 presents results for models trained on LCC. The low-quality model performs best both in terms of validation accuracy (99.29%) and cross-dataset generalization (HTER 39.99%). In contrast, the medium-quality model underperforms considerably (HTER 46.76%), suggesting that training on medium-quality data may not offer enough robustness to cope with domain shifts. High-quality training offers a modest compromise.

Tab. 3: Models trained on LCC and evaluated on CASIA

Model	**Val Acc (%)**	**Train Quality**	**HTER (%)**
folder_low_quality	99.29	Low	**39.99**
folder_medium_quality	84.62	Medium	46.76
folder_high_quality	80.95	High	43.35

4.1.2 Visual Analysis

To gain a deeper understanding of model behavior, Figures 2–7 present detailed performance plots for each training scenario. These include loss evolution, validation accuracy trends, and learning rate schedules, offering complementary insight beyond scalar HTER values. Below we highlight key patterns observed across different quality levels and training datasets:

- **Low-quality training (Figures 2 and 5):** These models exhibit fast and consistent loss convergence and maintain low validation error, despite being trained on degraded inputs. Notably, the model trained on LCC low-quality samples achieves the

best generalization (HTER = 39.99%), reinforcing the hypothesis that exposure to noisy data fosters more invariant feature representations.

- **Medium-quality training (Figures 3 and 6):** Although the CASIA-based medium-quality model attains the highest validation accuracy (97.08%), its validation loss plateaus early (after epoch 6), and the resulting HTER remains high (43.13%). This suggests overfitting to clean in-domain patterns, with reduced generalization capacity. The LCC-based medium model performs even worse (HTER = 46.76%) and shows signs of optimization instability.
- **High-quality training (Figures 4 and 7):** These models show more balanced loss curves and slightly lower final HTER compared to MQ, but still suffer from a noticeable drop in performance when transitioning to the target domain. Their relatively high FRR hints at a tendency to reject bona fide samples under mismatched conditions.

These visual diagnostics confirm that strong in-domain validation accuracy is not necessarily indicative of cross-domain robustness. In particular, the medium-quality assumption, often used in practice as a proxy for ”good data”, does not consistently yield generalizable models. Training with low-quality data, while counterintuitive, appears to regularize learning and reduce overfitting to spurious artifacts found in more pristine datasets.

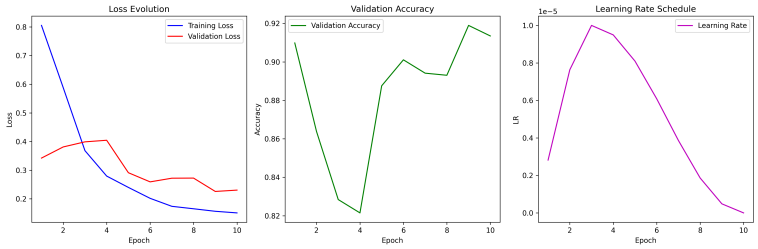

Fig. 2: CASIA → LCC: model trained on low-quality CASIA samples.

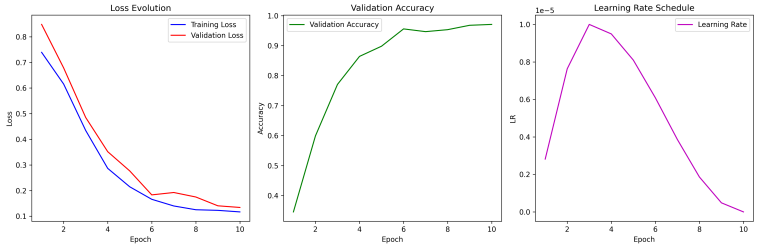

Fig. 3: CASIA → LCC: model trained on medium-quality CASIA samples.

4.1.3 Interpretation and Implications

The results from our stratified training experiments reveal several nuanced takeaways regarding the role of sample quality in cross-dataset generalization:

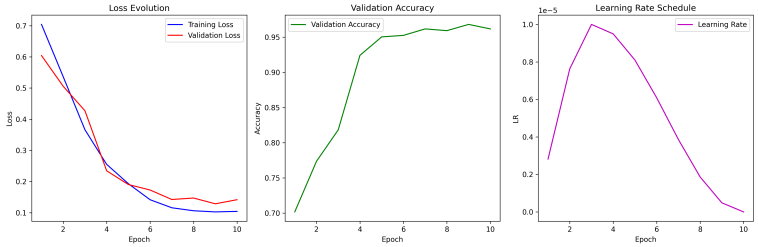

Fig. 4: CASIA → LCC: model trained on high-quality CASIA samples.

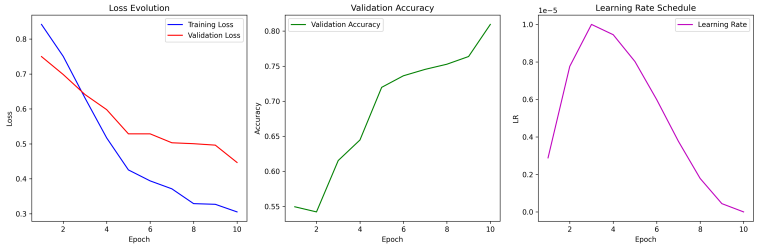

Fig. 5: LCC → CASIA: model trained on low-quality LCC samples.

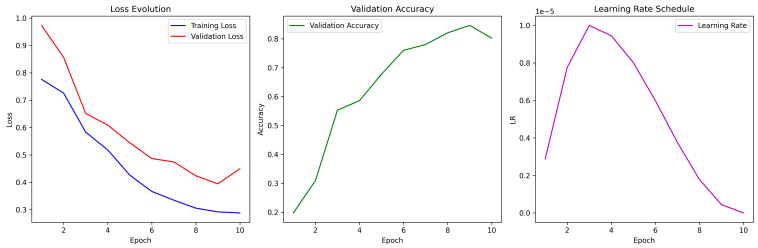

Fig. 6: LCC → CASIA: model trained on medium-quality LCC samples.

- **Low-quality training data** consistently leads to better generalization, particularly in the LCC→CASIA direction. Exposure to noise, compression artifacts, and blur may encourage the model to learn more robust, invariant features that are less sensitive to superficial variations.
- **Medium-quality models** achieve high validation accuracy in their source domain but perform poorly under distribution shift. This suggests that models trained on seemingly "optimal" samples may actually overfit to subtle spurious patterns or dataset-specific artifacts.
- **High-quality models** offer a compromise: while their validation accuracy is lower than MQ, they sometimes produce the lowest HTER (e.g., CASIA→LCC). However, their behavior is inconsistent and more prone to false rejections, indicating brittleness in unseen or noisy environments.

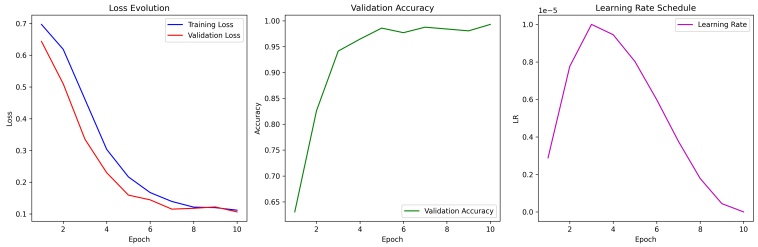

Fig. 7: LCC → CASIA: model trained on high-quality LCC samples.

Importantly, strict quality stratification reduces the number of training samples per split—particularly in the LQ and HQ categories. This introduces training instability, as seen in the slower convergence and less smooth loss curves of some models. The observed fluctuations may stem from insufficient diversity within each stratum.

These findings suggest that while quality-aware training is beneficial, future work should explore hybrid approaches: for example, adaptive quality-weighted sampling, progressive curriculum learning, or dynamic augmentation schedules that preserve both diversity and robustness without sacrificing sample count.

5 Conclusion

This work presents a comprehensive investigation into the role of sample quality stratification in face presentation attack detection (PAD), with a particular focus on cross-dataset generalization. Using two representative datasets (CASIA and LCC), we evaluated models trained on stratified quality tiers , low (LQ), medium (MQ), and high (HQ) , and observed clear differences in generalization behavior depending on the source domain and training quality.

Our results demonstrate that models trained on medium-quality samples from CASIA achieved the highest validation accuracy (97.08%) but not the best generalization, while the best overall cross-dataset HTER (39.99%) was achieved by training on low-quality LCC samples. These findings suggest that lower-quality training data may encourage more invariant feature learning, whereas higher-quality data may lead to overfitting to idealized conditions. However, quality stratification alone is not sufficient to ensure generalization: performance varied considerably depending on the direction of domain shift.

Key insights from our study include:

- Quality-aware training reveals that validation accuracy is not always predictive of robustness.
- Low-quality samples can promote better generalization, possibly due to exposure to more degraded and varied visual conditions.

- Stratifying datasets by quality reduces available training data per model, which may hinder learning, especially at quality extremes.

Future work should explore dynamic quality-aware training schedules, curriculum learning strategies, and the integration of stratified training with causal disentanglement or domain-invariant representation learning. Our proposed framework requires no additional annotations, making it a practical tool to enhance the robustness of PAD systems in real-world, unconstrained scenarios.

References

[At18] Atoum, Yaman; Liu, Yaojie; Jourabloo, Amin; Liu, Xiaoming: Face anti-spoofing using patch and depth-based CNNs. In: IEEE International Joint Conference on Biometrics (IJCB). S. 319–328, 2018.

[Be20] Becker, Manuel; Damer, Naser; Boutros, Fadi; Kuijper, Arjan: On the Effectiveness of Feature-Level Domain Adaptation for Robust Spoofing Detection. In: IEEE International Joint Conference on Biometrics (IJCB). S. 1–10, 2020.

[GM19] George, Anjith; Marcel, Sébastien: DeepPixBis: Presentation attack detection via deep pixel-wise binary supervision. In: IEEE International Joint Conference on Biometrics (IJCB). S. 1–8, 2019.

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). S. 770–778, 2016.

[He21] Hernández-Ortega, Javier; Galbally, Javier; Fiebrink, Elizabeth; Fierrez, Julian: FaceQnet v2: Quality assessment for face recognition systems. IEEE Access, 9:139811–139825, 2021.

[JRN16] Jain, Anil K; Ross, Arun; Nandakumar, Karthik: Biometrics: A Tool for Information Security. IEEE Transactions on Information Forensics and Security, 1(2):125–143, 2016.

[Lea21a] Liu, Yaojie; et al.: Dual-agent GANs for photorealistic and identity preserving profile face synthesis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). S. 7553–7563, 2021.

[Lea21b] Liu, Yaojie; et al.: Multi-modal face anti-spoofing with optical flow guided feature fusion. IEEE Transactions on Information Forensics and Security, 16:3355–3367, 2021.

[LH19] Loshchilov, Ilya; Hutter, Frank: Decoupled Weight Decay Regularization. International Conference on Learning Representations (ICLR), 2019. arXiv:1711.05101.

[LJL20] Liu, Yaojie; Jourabloo, Amin; Liu, Xiaoming: Deep learning for face anti-spoofing: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3):825–845, 2020.

[Me21] Meng, Qiang; Zhao, Shichang; Huang, Zhida; Gong, Ming; Zhou, Xingyu; Wang, Lei; Liu, Xudong; Wang, Xiaogang: MagFace: A universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). S. 14225–14234, 2021.

[MHP11] Määttä, Jukka; Hadid, Abdenour; Pietikäinen, Matti: Face spoofing detection from single images using micro-texture analysis. In: International Joint Conference on Biometrics (IJCB). S. 1–7, 2011.

[MMB12] Mittal, Anish; Moorthy, Anush Krishna; Bovik, Alan C: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing, 21(12):4695–4708, 2012.

[Pea16] Patel, Vishal M; et al.: Secure face unlock using spectral features and liveness detection. In: IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS). S. 1–8, 2016.

[RB20] Ramachandra, Raghavendra; Busch, Christoph: Face recognition with presentation attack detection: Recent advances and future directions. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2(3):376–393, 2020.

[Sh20] Shao, Zhenhua; Li, Haoliang; You, Shan; Wang, Zitong; Kot, Alex C: Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: European Conference on Computer Vision (ECCV). S. 574–591, 2020.

[Tea20] Terhörst, Philipp; et al.: SDD-FIQA: Face image quality assessment based on semantic and deep descriptors distribution distance. In: International Joint Conference on Biometrics (IJCB). S. 1–10, 2020.

[Wa22a] Wang, Jiaying; Chen, Yiqing; Yu, Han; Zhang, Chunyan Miao: Domain generalization: A survey. IEEE Transactions on Knowledge and Data Engineering, 2022.

[Wa22b] Wang, Zitong; Shao, Zhenhua; Li, Haoliang; Kot, Alex C: CF-PAD: Causal Feature Learning for Face Presentation Attack Detection. In: European Conference on Computer Vision (ECCV). S. 533–550, 2022.

[WD19] Wang, Mei; Deng, Weihong: Learning a discriminative feature network for deep face recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). S. 4285–4294, 2019.

[Zh21] Zhou, Kaidi; Yang, Yongxin; Qiao, Yu; Loy, Chen Change: Domain generalization with mixstyle. In: International Conference on Learning Representations (ICLR). 2021.