IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 December 2025, accepted 29 December 2025, date of publication 12 January 2026, date of current version 15 January 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3651853

RESEARCH ARTICLE

FaceSpoofLDM: Language-Guided Synthesis of Face Presentation Attacks Based on Latent Diffusion

ANDRÉ DÖRSCH[1], MARCEL GRIMMER[1], LAZARO JANIER GONZALEZ-SOLER[1], ROBERTO CASULA[2], (Member, IEEE), GIAN LUCA MARCIALIS[2], (Senior Member, IEEE), CHRISTOPH BUSCH[1], (Fellow, IEEE), AND CHRISTIAN RATHGEB[1]

[1]da/sec–Biometrics and Security Research Group, Hochschule Darmstadt, 64295 Darmstadt, Germany
[2]Saifer Laboratory, Department of Electrical and Electronic Engineering, Biometric Unit, University of Cagliari, 09123 Cagliari, Italy

Corresponding author: André Dörsch (andre.doersch@h-da.de)

This work was supported in part by the European Union (EU) Interoperable Applications Suite to Enhance European Identity and Document Security and Fraud Detection (EINSTEIN) under Grant 101121280; in part by the German Federal Ministry of Education and Research; and in part by the Hessian Ministry of Higher Education, Research, Science, and the Arts within their joint support of the National Research Center for Applied Cybersecurity (ATHENE).

ABSTRACT Presentation Attacks (PAs) pose a serious threat to face recognition (FR) systems. These attacks cover a broad range of scenarios, including images replayed on various devices, printed photographs, or more sophisticated approaches such as 3D masks used to impersonate another identity. Recent advances in deep neural networks have led to an increasing number of face presentation attack detection (PAD) methods, replacing traditional approaches with great success. However, these methods are highly data-intensive and require large amounts of training data for reliable decision-making. Although several face PAD datasets have been introduced, they often come with restricted usage, limited subject and attack diversity and privacy or legal constraints. In this work, we introduce FaceSpoofLDM, a latent diffusion model (LDM) for language-guided image synthesis to generate synthetic face PAs and non-attacks across various demographic groups. Our approach reduces the need for manually crafting physical presentation attack instruments (PAI) while increasing scalability and attack diversity. Extensive experiments demonstrate the effectiveness of our model and show that incorporating synthetic PAIs, on average, enhances security against PAs.

INDEX TERMS Synthetic face presentation attacks, presentation attack detection, biometric security, image synthesis, diffusion models.

I. INTRODUCTION

Biometric technologies such as face recognition (FR) systems have become an integral part of our society due to their increasing social acceptance, wide-ranging application scenarios, and the capability to authenticate individuals in an efficient and trustworthy manner. Despite the numerous advantages of FR systems [1], [2] over traditional authentication approaches, targeted attacks on the system policy may result in unwanted decisions. One attack category that falls within the scope of FR systems are so-called face presentation attacks (PAs). According to the international standard ISO/IEC 30107-3 [3], a PA refers to the presentation of human characteristics or an artefact to a biometric capture subsystem in a manner intended to interfere with the operation of the biometric system. These attacks range from face images replayed on various hardware devices, printed photographs, make-up attacks or silicone mask attacks primarily intended to impersonate another identity [4], [5], [6]. To this end, numerous face presentation attack detection (PAD) methods have been proposed in the literature [7], [8], [9], [10], [11], [12] to mitigate the risk of PAs and preserve security. For a more comprehensive discussion of the development of PAD methods, the interested reader is referred to [13]. In general, PAD detection methods can be categorized into two groups: sensor-based methods,

The associate editor coordinating the review of this manuscript and approving it for publication was Lei Shu.

which utilize dedicated hardware such as spectral sensors to verify the authenticity of a captured biometric sample, and software-based methods, which extract relevant features directly from the captured sample [14]. A major advantage of software-based PAD methods over hardware-based ones is that they do not require additional sensors, making them device-independent and more flexible in their application. As software-based PAD methods are more widely studied in the literature, our work focuses exclusively on these methods. In recent years, significant advances in deep neural networks have led to more advanced deep learning-based PAD methods replacing traditional approaches with great success. However, these algorithms require large amounts of training data to make reliable and accurate decisions.

To address this need, several publicly accessible face PAD databases have been made available [15], [16], [17], [18]. However, most face PAD databases lack subject diversity and mainly include prevalent attack types, such as print or replay attacks, whereas sophisticated or unconventional attack types, such as wearing silicone masks [19], are generally available in a more limited variety. In addition, as many PAD databases include authentic biometric samples, their use is often constrained by data protection regulations and privacy concerns. Especially with the introduction of regulatory and legal frameworks such as the General Data Protection Regulation (GDPR) [20] or the European AI Act [21], the use of synthetic data is becoming increasingly necessary, as the use of real biometric data is accompanied by strict restrictions on collection, storage and usage. As a result, synthetic data is increasingly being incorporated in the training process of deep learning models. An example of the application of synthetic data in the biometric field is its use to improve biometric fairness [22], [23]. Additional benefits of incorporating synthetic data for FR systems are demonstrated in the FRCSyn-onGoing challenge [24], which underlines its potential to improve recognition performance, whereas training exclusively on synthetic data generally results in lower accuracy. Based on these findings, no experiments were conducted in this paper in which PAD algorithms are trained exclusively on synthetic data. In addition, we would like to emphasize that the cross-domain face pad evaluation experiments presented in Section IV-B are intended as supporting analysis to demonstrate how synthetic PAs impact PAD robustness under domain-shift. Consequently, no intra-dataset evaluations are performed. A more detailed discussion of synthetic data for biometric applications is provided in Section II-A.

Just a few years ago, the aforementioned image synthesis seemed almost infeasible. However, groundbreaking development in generative models [25], [26] have made synthetic images indistinguishable from authentic ones.

In the scope of FR systems, synthetic data refers, for example, to a synthetically generated identity whose biometric characteristics are not linked to an existing individual. This property is, among other benefits, advantageous over the usage of authentic biometric samples, as it offers an efficient and privacy-friendly alternative [28]. Therefore, the generation and utilization of synthetic face PAs is of high interest for PAD research, as it enables high data diversity and scalability. In addition to the general advantages of synthetic data over authentic ones, synthetic PAs offer significant cost and resource-saving advantages, as they avoid time-consuming and costly processes. Synthetic PAs reduce the need for crafting physical attack types that require special materials, while ensuring high diversity and scalability.

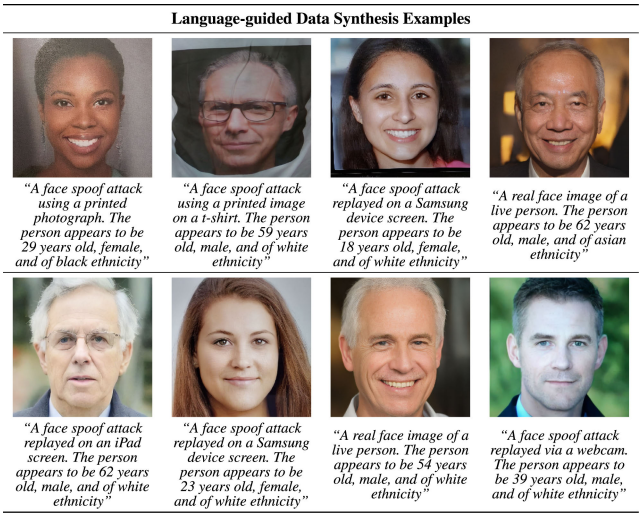

FIGURE 1. Synthetic identities (PAs and non-attacks) generated via text-guided image synthesis. Samples generated with 250 DDIM steps [27] and $\eta = 1$. FaceSpoofLDM allows the generation of specific PAI species across various soft biometric characteristics such as age, gender, and ethnicity.

In this work, we introduce FaceSpoofLDM, a novel approach based on a latent diffusion model (LDM) [29], allowing precise image synthesis of both PAs and non-attacks across diverse demographic groups to address the limitations of existing PA databases. FaceSpoofLDM allows language-guided data synthesis via precise prompts, as demonstrated in Figure 1. The novelty of the proposed approach lies in its ability to manipulate soft-biometric characteristics (age, gender, and ethnicity) in the latent space to generate diverse synthetic PAs and non-attacks. This allows the generation of specific attack scenarios across various demographic groups. In this work, the generation and utility of the synthetic identities and the model performance will be demonstrated in extensive experiments. In summary, this work makes the following contributions:

- Training a latent diffusion model for image synthesis in the Face PAD domain, incorporating a novel language-guided conditioning mechanism to enable controlled generation of synthetic PAs and non-attacks across various demographic groups.
- Conducting an in-depth evaluation based on the models quality, including whether our proposed model generates synthetic data that align with the intended soft-biometric characteristics specified in the text-prompt and potential memorization effects.

- Performing an extensive cross-domain performance evaluation in line with metrics defined in the international standard ISO/IEC 30107-3 [3] benchmarking various PAD methods trained on authentic identities enriched with synthetic samples, demonstrating the effectiveness of the proposed approach.

The remainder of this work is organized as follows: Section II briefly summarizes the utilization of synthetic data in the biometric domain, its advantages over real data, but also associated challenges, including the generation and usage in the PAD context. In Section III, we introduce FaceSpoofLDM and describe its underlying architecture and components, including the soft-biometric conditioning mechanism, in a detailed manner. Section IV summarizes the experiments and evaluations conducted to demonstrate the proposed model's capabilities, such as its application to PAD. Section V discusses the limitations of our proposed model and highlights associated challenges. Finally, Section VI summarizes and concludes on outcomes of our proposed model.

II. RELATED WORK

A. SYNTHETIC DATA FOR BIOMETRIC APPLICATIONS

Synthetic data has proven beneficial for various human-related analysis tasks, including biometric recognition [30]. Among other benefits, synthetic data is mainly used for eliminating drawbacks of real data, such as privacy concerns, annotation costs, scalability and diversity. There are numerous studies that have achieved aforementioned benefits through the usage of synthetic data in a biometric context. In [31], Wood et al. have shown that face-related tasks (e.g. face detection) can be effectively performed using exclusively synthetic data, with results comparable to real data. Additionally, the authors showed that the utilization of synthetic data opens up new possibilities for tasks where manual data annotation would be practically unfeasible [32]. Another work by Colbois et al. [33] investigated the use of synthetic data on FR systems and came to similar conclusions: Synthetic identities are a valuable replacement for authentic data as they provide similar benchmarking results. Other biometric modalities also benefit from incorporating synthetic data, such as synthetic fingerprints [34], [35], synthetic gait images [36] or synthetic iris images [37], [38].

Synthetic data can also be used to supplement the training of neural networks, which typically have limited data access to minority demographic groups such as children [39], [40] or individuals with disabilities [41]. To this end, the application of synthetic identities also has the potential to improve fairness of existing biometric systems and increase recognition performance for minority groups. However, the generation of synthetic data is also associated with several challenges, such as potential identity leakage [42], ethical and legal considerations [43] and the risk of reproducing existing data biases [44]. For a more comprehensive discussion of the role of synthetic data in human analysis including its benefits and challenges, the interested reader is referred to [30].

B. APPLICATION OF SYNTHETIC DATA IN THE PAD DOMAIN

Recent advances in deep learning-based PAD algorithms have led to more robust and secure systems against targeted attacks. Simultaneously they pose new challenges. On the one hand, deep learning PAD methods require large amounts of diverse training data to cover a broad variety of possible PAs. At the same time, authentic PAs are often only available to a limited extent and come with aforementioned drawbacks. In addition, high development costs, expert knowledge, and the effort required to manually craft PA attacks make the contribution of authentic PAs a challenging and time-consuming task. To this end, various work on synthetic data across multiple biometric modalities in the PAD domain has been contributed.

In [45], Fang et al. developed and contributed the privacy-friendly (semi) synthetic face PAD dataset SynthASpoof, containing 25,000 bona fide and 78,800 attack samples. The bona fide samples in SynthASpoof were generated using StyleGAN2-ADA [46], while the PAIs were collected manually by presenting these synthetic bona fides to capture systems in real attack scenarios. While SynthASpoof exclusively covers print and replay attacks, Ibsen et al. [47] proposed a T-shirt Face PA (TFPA) database of 1,608 T-shirt attacks using 100 unique synthetic PAIs. Although synthetic PAD datasets address existing limitations of authentic PAD datasets (e.g. privacy restrictions), they also exhibit limitations: While SynthASpoof and TFPA each focus on specific PAIs (e.g. TFPA exclusively covers T-shirt attack scenarios), they also remain constrained in scale and subject diversity. FaceSpoofLDM is designed to overcome both limitations: privacy restrictions and a broader (combined) attack coverage, by enabling demographically controlled and scalable synthetic face PAD generation across multiple PAI species. In addition to provided synthetic PAD databases, there are also approaches to directly generate pseudo-negative samples in the feature space. In Ma et al. [48] the authors analyse the feature distributions of bona fide and PA samples to synthesise new pseudo-negative features that expand the PA space to increase the model's robustness to unseen attacks. Another approach using synthetic data in the PAD domain was very recently proposed by Tapia et al. [49], presenting a novel synthetic passport dataset for identifying fake identity documents. In addition to the Face PAD systems, the generation and utilization of synthetic PAs for various other biometric modalities was investigated, such as synthetic fingerprint PAs [50], synthesized iris PAs [51], [52] or speech synthesis attacks [53].

C. DIFFUSION MODELS IN THE FACE PAD DOMAIN

Early work by Ngyuen et al. [54] explored a CycleGAN-based approach for the synthetic generation of PAs. Their model learns to reproduce the characteristics of captured bona fides and PAs and the mapping between these two domains, but does not allow for explicit control over PAI species, identity factors or soft-biometric characteristics. While such

GAN-based approaches introduced initial directions for PA synthesis, they have also motivated the development of more modern diffusion-based methods for PA synthesis. However, the application of diffusion models in the face PAD domain remains widely understudied and related literature is limited.

Among these, Ge et al. [55] recently introduced the diffusion-based DiffFAS framework, which synthesizes face PAs from authentic ones. The DiffFAS framework introduces a ''Bona fide to PA'' transformation to incorporate PAI properties into authentic face images while preserving their identity. While the goal of DiffFAS is primarily to address domain shift and robustness to cross-domain scenarios, our work takes a more flexible and granular approach by allowing the targeted generation and control of soft-biometric attributes in the latent space. This results in the generation of synthetic data points and biometric samples that are not necessarily available in the training dataset. In [56], Zhang et al. employ a diffusion model to denoise PA images to reconstruct bona fide images from the extracted PA noise patterns. Consequently, the authors utilize the extracted PA noise as a discriminating feature for PAD. While this work focuses on improving PAD classification performance through noise modelling, it does not address the generation of diverse synthetic training data. Very recently, Ko et al. [57] proposed SpoofFusion a text-to-image diffusion framework to address domain robustness with synthetic spoofs. Their approach fine-tunes Stable Diffusion to capture bona fide face features and employs Low-Rank Adaptation (LoRA) [58] for spoof detection. Another recent work by Zhang et al. [59] utilized a diffusion model for synthetic spoof generation additionally demonstrated the potential of image synthesis in the face PAD domain. However, unlike our proposed FaceSpoofLDM, their methods do not explicitly offer control over soft-biometric characteristics in their spoof generation.

To the best of our knowledge, FaceSpoofLDM is the first diffusion-based PAD framework that (1) enables explicit and fine-grained control over both PAI species or non-attack and the soft-biometric characteristics of the generated face image, (2) is trained exclusively on synthetic identities, and (3) supports the generation of multiple PAI species including print, replay and t-shirt attacks, as well as synthetic non-attack face images across various demographic groups.

III. FaceSpoofLDM

As summarized in Section II, several existing PAD databases have limitations due to the use of authentic biometric samples. To address this, our objective is to enable a scalable and diverse generation process of different PAI species across various demographic groups. However, many existing PAD databases lack these properties, potentially leading to biased decisions [60]. To ensure diverse image synthesis to generate synthetic face PAs and non-attacks across various demographic groups, we employ our soft-biometric conditioning mechanism, introduced in Section III-B. This allows the generation of specific attack scenarios by manipulating soft-biometric characteristics in the latent space, as illustrated in Figure 2.

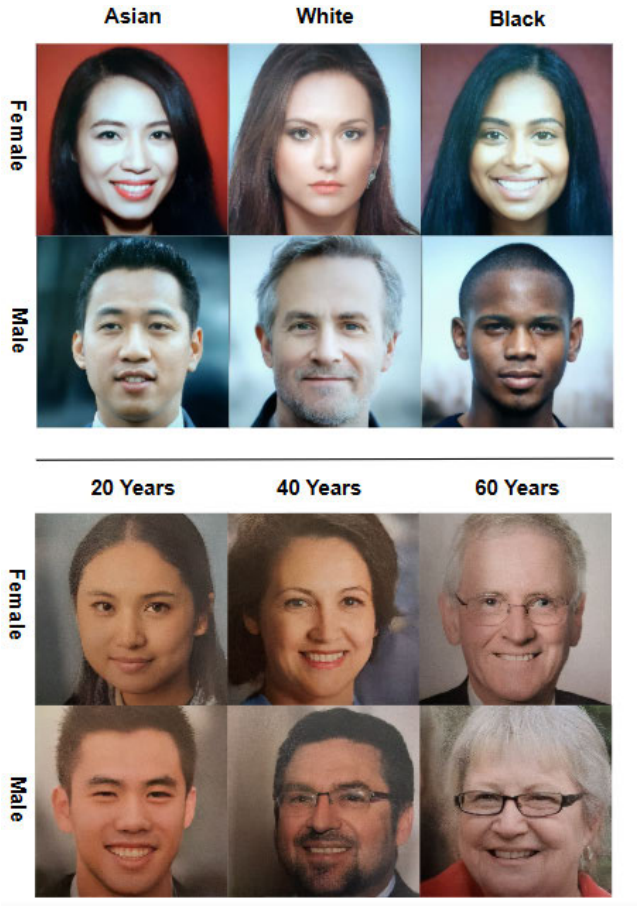

FIGURE 2. Synthetic PAs generated using FaceSpoofLDM. Our model allows image synthesis across different demographic groups, including variations in age, gender, and ethnicity. The upper section of the figure showcases synthetic webcam attacks categorized by ethnicity and gender, while the lower section organizes synthetic print attacks by age and gender.

A. LATENT DIFFUSION MODEL

FaceSpoofLDM is based on the LDM architecture introduced by [29], which reduces computational complexity and memory allocation by conducting the diffusion and denoising process in the latent space, instead of pixel space. Several LDM-based architectures have recently been explored in the biometric domain related to certain facial manipulation tasks, such as face age editing [61] or being used for privacy-preserving adversarial generation [62]. A visual overview of the underlying architecture is illustrated in Figure 3.

The main components of the LDM architecture can be broken down and described as follows:

- **Perceptual Encoder & Decoder**: Instead of applying the diffusion process directly to the pixel-space representation of the input image x, the perceptual encoder $\mathcal{E}$ and decoder $\mathcal{D}$ manage the mapping to and from the latent space representation z. This transition

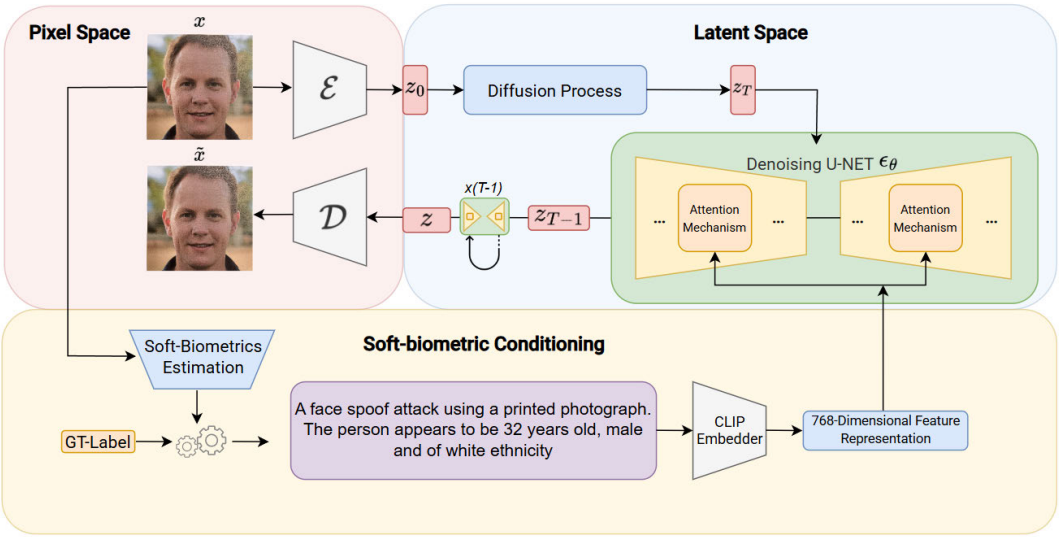

FIGURE 3. High-level model-architecture overview: A high level overview of the proposed models' workflow. A detailed component description and the models workflow can be found in Section III.

reduces computational costs while maintaining essential semantic information in a compressed representation. This component is not trained alongside the diffusion process but uses pretrained autoencoders. For more in-depth material about this component, the interested reader is referred to [63].

- **Diffusion Process & Denoising U-Net**: During the forward diffusion process, Gaussian noise is incrementally added to the latent space representation z over T steps, resulting in a sequence z_t, where z_T follows a standard normal distribution. Subsequently, a denoising U-Net [64] is then trained to estimate the noise ϵ in z_t at each step t, so that z can be recovered through recursive denoising. Our customized conditioning is incorporated into the denoising U-Net via a Cross-Attention mechanism [65], further described in Section III-B. This conditioning guides the denoising process by injecting desired features at multiple attention layers.

The training and inference workflow architecture of FaceSpoofLDM is summarized in Algorithm 1. FaceSpoofLDM was trained using the AdamW optimizer with the default weight decay of 1e-2 on a single NVIDIA Tesla A100 for 220 Epochs. Furthermore our proposed model is trained utilizing the standard LDM loss that minimizes the squared L2 norm between the Gaussian noise ϵ and the estimated noise from the denoising U-Net $\epsilon_\theta(z_t, t)$:

$$L_{LDM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t)||_2^2\right] \quad (1)$$

where t timesteps are sampled uniformly during the training process. An overview of FaceSpoofLDM hyperparameters is provided in Table 1. Learning Rate was manually adjusted during the training process ranging from 3e-5 to 8.5e-5. For a more comprehensive discussion of latent diffusion and its computational aspects, the interested reader is referred to [29].

TABLE 1. Overview of the FaceSpoofLDM hyperparameters.

Hyperparameter	Value(s)
f	4
z-shape	$64 \times 64 \times 3$
$\|Z\|$	8192
Diffusion steps	1000
Noise Schedule	linear
Channels	224
Num. Res. Blocks	2
Channel Multiplier	1,2,3,4
Spatial Transformer	True
Attention resolutions	8,4,2
Head Channels	32
Batch Size	64
Optimizer	AdamW
Learning Rate	3e-5 to 8.5e-5
Conditioning	Cross-Attention
Transformer-Depth	1
Embedding Dimension	768

B. CONDITIONING

To enable controllable text-to-image generation, FaceSpoofLDM utilises a conditioning mechanism based on cross-attention. Therefore, the model is guided by both the subject's PAI species and soft-biometric characteristics (age, gender, ethnicity), allowing the generation of synthetic samples reflecting these attributes. During training, the soft-biometric characteristics (age, gender, and ethnicity) of an input image x were estimated using a Commercial Off-The-Shelf (COTS) system (Cognitec's FaceVACS technology (Version 9.8.0)). We would like to point out that, as with any other automated demographic estimation method, the extraction of soft biometric attributes can lead to prediction errors or biases. These estimated demographic variables complement the ground truth PAI label (e.g. Print-Attack) and

Algorithm 1 Training and Inference Workflow of FaceSpoofLDM

Training Phase

1: Given input image x and ground-truth PAI species.
2: Extract soft-biometric characteristics from x (age, gender, ethnicity) using COTS system (Cognitec).
3: Construct a fixed text-prompt by inserting (PAI species + soft-biometric characteristics) into a predefined template (see Figure 4).
4: Encode the text-prompt into embedding vector representation c via CLIP text encoder.
5: Encode x into latent representation $z_0 = E(x)$.
6: Apply forward diffusion at timestep t to obtain z_t
7: Condition denoising U-Net on c via cross-attention.
8: Predict noise $\hat{\epsilon} = \epsilon_\theta(z_t, t)$.
9: Minimizes the squared L2 norm between the Gaussian noise ϵ and the estimated noise from the denoising U-Net $\epsilon_\theta(z_t, t)$ (see Equation 1).

Inference Phase

1: Input: Text-prompt (e.g. "A face spoof attack using a printed image on a T-shirt. The person appears to be 43 years old, female and of asian ethnicity").
2: Encode text-prompt into embedding vector representation via CLIP text encoder.
3: Sample initial noise $z_T \sim \mathcal{N}(0, I)$.
4: **for** $t = T, T-1, \ldots, 1$ **do**
5: Iteratively denoise z_t using conditioned U-Net.
6: **end for**
7: Decode via D to obtain synthetic PA or non-attack image.

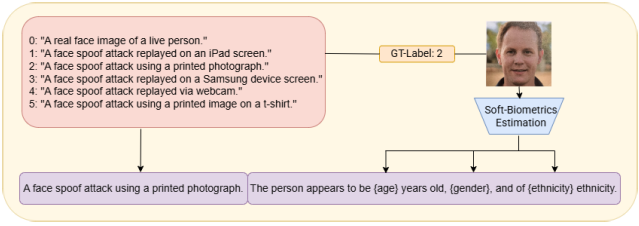

FIGURE 4. High-level overview of the creation of a fixed text prompt: The soft-biometric characteristics (age, gender, and ethnicity) are estimated from the input image using a COTS system. Simultaneously, the ground truth label is mapped to a fixed prefix. Both parts are combined into a text-prompt and passed to the CLIP text embedder.

are combined into a fixed text-prompt by inserting both the PAI species and estimated soft-biometric characteristics into a predefined prompt template (see Figure 4).

This prompt is then encoded using a CLIP text embedder [66], resulting in a 768-dimensional feature vector. The feature vector is incorporated into the denoising U-Net via a cross-attention mechanism and subsequently guides the denoising process. Figure 5 illustrates different synthetic PAs of examples for the selected PAI species. The benefit of using this proposed soft-biometric text-guided conditioning approach over class-based conditioning mechanisms is that it uses prior semantic knowledge by incorporating the contextual depth of the CLIP embedder, rather than relying on discrete class labels with limited semantic information. To this end, encoding the text-prompt *"A face spoof attack using a printed photograph. The person appears to be 29 years old, female and of black ethnicity."* using CLIP embedder leads to a context-aware embedding enriched with prior knowledge of facial semantics typical to the selected demographic attributes. During training, these feature embeddings incorporate inductive biases to facilitate generalizability of the denoising U-Net, thus, leading to better controllability of demographic factors during inference.

TABLE 2. Overview of the databases used for training FaceSpoofLDM.

Dataset	Year	Images	PAI species
SynthASpoof [45]	2023	103,800	Print & Replay
TFPA [47]	2023	1,608	T-shirt

C. CHOICE OF TRAINING DATABASES

Our model was trained exclusively on a combined training dataset including the SynthASpoof [45] database and the TFPA database [47] which were further described in Section II-B. Although these databases exclusively contain synthetic identities, these were embedded on real PA artefacts. An overview with statistics on the databases used for training FaceSpoofLDM is shown in Table 2. The combined training dataset consists of 93,430 synthetic face images from SynthASpoof and 1,141 images from TFPA (total 94,571 training images). Each training epoch used all 94,571 synthetic training images. The remaining images were held-out to validate training stability. No real identities were used at any point during the training of FaceSpoofLDM. We opted for these two databases, as both contain synthetic data, offering the previously mentioned advantages over authentic identities. In addition, the combined training dataset enables our model to be trained on a wide range of PAI species, including common PAIs such as print or replay attacks, as well as unconventional PAIs such as T-shirt attacks.

However, the limited number of images in the TFPA database could lead to weaker generalization compared to other PAI species. This potential limitation will be further investigated in Section IV-A. During training, we utilized a Weighted Random Sampler [67], to assign each training sample a weight proportional to the inverse frequency of its PAI species.

D. MODEL BOUNDARY CONDITIONS

FaceSpoofLDM was developed and evaluated under a set of boundary conditions that define the model's operational scope, which are summarized as follows:

- Synthetic training domain: FaceSpoofLDM was trained exclusively on synthetic identities from SynthASpoof [45] and TFPA [47]. Consequently, the

model's boundaries for image generation are limited to (1) the PAI species present in these datasets and (2) the learned semantic feature knowledge of the synthetic identities.

- Soft-biometric Conditioning: During model training, our proposed conditioning mechanism relies on soft-biometric characteristics (age, gender, and ethnicity) estimated using a COTS system. As a result, the model's generation knowledge is limited to demographic combinations represented in the synthetic training domain.
- Architectural boundaries: As our model is based on latent-diffusion [29], architectural components, such as the use of a pretrained autoencoder [63] for latent-space compression and the CLIP text embedder [66] for prompt encoding further define the representational and semantic capabilities for image generation.

IV. EXPERIMENTS

A. MODEL QUALITY

To evaluate whether our proposed model generates synthetic face PAs that align with the intended soft-biometric characteristics specified in the text-prompt, we generated 5,000 prompts across various PAI species. Specifically, the soft-biometric characteristics were randomly sampled from predefined ranges: age (5-60 years), gender (male, female) and ethnicity (White, Black, Asian) across the PAI species. These sampled attribute combinations were then embedded into a fixed text-prompt template and provided as input to our proposed FaceSpoofLDM, as illustrated in Figure 4. The soft-biometric characteristics (age, gender, ethnicity) of the generated face PAs were subsequently estimated using a Commercial Off-The-Shelf (COTS) system[1] and compared against the sampled demographic characteristics from the 5,000 prompts. We then analysed the accuracy of ethnicity and gender predictions by comparing the characteristics from the text-prompts with the estimated characteristics of the COTS system. For age, we evaluated the Mean Absolute Error (MAE) between the age from the text-prompt and the estimated age of the synthetic PA sample by the COTS system: A low MAE indicates that the COTS system accurately predicts the age of the synthetic individuals with minimal average error, while a high MAE suggests that the system often estimates an age that is very different and consequently does not align with the text-prompted age.

As shown in Table 3, it is noticeable that the percentage of estimated soft-biometric characteristics that align with the characteristics specified in the text-prompt provide reliable results across print and replay attacks, with minimal deviations. Across print and replay attacks, the gender and ethnicity specified specified in the text-prompt and those estimated by the COTS system remain consistently high, ranging from 96.1% to 97.43% for gender and 93.94% to 96.81% for ethnicity, respectively. The mean absolute error (MAE) for age estimation remains low, ranging from 2.48 years to 3.13 years. A greater deviation between the characteristics specified in the text-prompt and the characteristics estimated by the COTS system can be observed for T-shirt attacks. The MAE for age increases to 9.69 years, while the ethnicity estimation drops to 83.13%. This behaviour is to be expected, as the TFPA database contains a limited number of unique subjects. Moreover, when subject diversity is strongly limited, generative models tend to replicate existing training data rather than generating novel patterns [68].

To further investigate whether FaceSpoofLDM is prone to a memorization effect and thus replicates existing training data, we conducted additional experiments. To this end, we randomly sampled 500 images from our proposed model's training database. We incorporated the images ground-truth label (PAI species or bona fide) and the estimated soft-biometric characteristics into 500 text-prompt. These prompts were then passed into FaceSpoofLDM to mimic the aforementioned data distribution with synthetic data. Subsequently, we utilized MagFace [69] to extract facial feature embeddings from both the 500 sampled images of the training distribution and the synthetic ones.

[1]Demographic characteristics (Age, Gender, and Ethnicity) have been estimated with Cognitec's FaceVACS technology (Version 9.8.0)

TABLE 3. Comparison of text-prompted vs. estimated soft-biometric characteristics across PAI species.

Attack Type	Age Prediction	Gender Prediction	Ethnicity Prediction
	MAE (Years)	Accuracy (%)	Accuracy (%)
Print (Photograph)	3.13	96.10 ± 0.19	93.94 ± 0.23
Replay (Samsung)	2.70	97.47 ± 0.15	95.15 ± 0.21
Replay (Webcam)	2.53	97.24 ± 0.16	96.57 ± 0.18
Replay (iPad)	2.48	97.43 ± 0.16	96.81 ± 0.18
T-Shirt	9.69	96.02 ± 0.20	83.13 ± 0.37

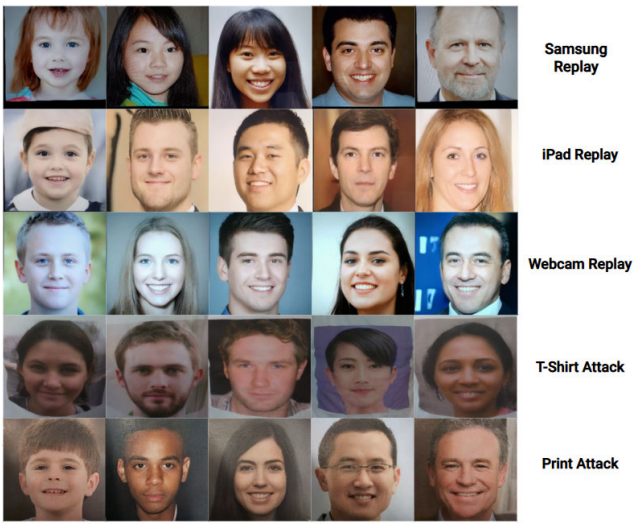

FIGURE 5. Synthetic PA samples generated via text-guided image synthesis: the generated identities can be precisely controlled via text-prompts allowing a scalable and diverse sampling approach. Samples generated with 250 DDIM steps and $\eta = 1$.

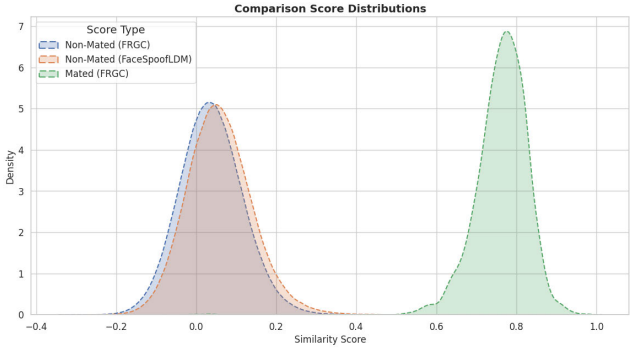

FIGURE 6. Comparison score distributions for non-mated and mated comparison trials. The FaceSpoofLDM score distribution is closely aligned with the non-mated score distribution from FRGCv2. We observe only minor differences between the non-mated comparison score distributions, indicating no significant identity leakage on our validation set.

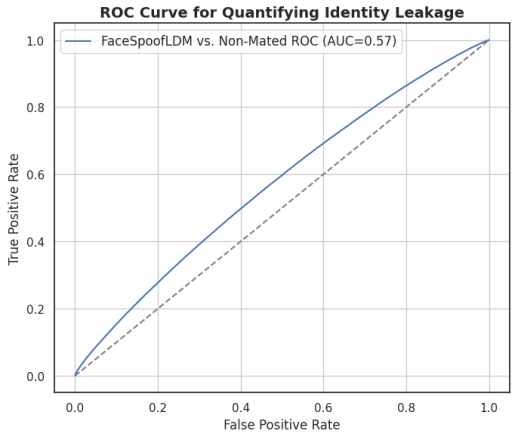

FIGURE 7. ROC curve comparing FaceSpoofLDM similarities against FRGC non-mated similarities.

TABLE 4. Comparison score statistics for non-mated and mated data.

Database	Comparison Type	Image Type	Mean	Std
FRGC	Mated	Bona fide	0.76	0.06
	Non-Mated	Bona fide	0.04	0.08
FaceSpoofLDM	Non-Mated	Non-Attacks + PAs	0.06	0.08
		Non-Attacks	0.06	0.09
		Samsung Replay	0.06	0.09
		iPad Replay	0.08	0.09
		Webcam Replay	0.07	0.08
		Print Attack	0.06	0.08
		T-shirt Attack	0.10	0.10

Following, we performed an N:N-comparison of the feature embeddings from the 500 sampled training images and the synthetic images to measure cosine similarity for obtaining a comparison score distribution. The expected distribution of the obtained comparison scores is expected to be close to a non-mated comparison score distribution. This is because the synthetic identities share the same soft-biometric characteristics and labels as the authentic images, but should not, in general, share the same identity. As a reference point for the obtained score distribution, we computed the mated and non-mated score distributions from the FRGCv2 [70] database.

Looking at the comparison score distributions from Figure 6, it is noticeable that the FaceSpoofLDM comparison score distribution is closely aligned with the non-mated score distribution from FRGCv2. However, in contrast to the non-mated score distribution, a slight shift can be observed for the FaceSpoofLDM distribution. This finding may indicate a minor memorization effect, as the non-mated distributions do not fully align. This could be caused, for example, by limited subject diversity when training FaceSpoofLDM for specific PAI species, as generative models tend to reproduce existing training data when subject diversity is limited [68].

In addition to the comparison score analysis above, we further quantified potential identity leakage by computing a ROC curve, as illustrated in Figure 7. Specifically, we evaluated how the FaceSpoofLDM score distribution can be distinguished from the non-mated score distribution of FRGCv2 (see Figure 6).

The obtained AUC of 0.57 is close to random guess (0.5), indicating only minimal identity leakage. Both, the comparison-score analysis and the ROC-based identity leakage quantification demonstrate, that FaceSpoofLDM hardly reconstructs training data, with potential few exceptions.

To identify which PAI species have the greatest influence on the distribution shift (identity leakage), Table 4 demonstrates, among others, descriptive statistics of the comparison score distributions associated with Figure 6. In addition to the overall FaceSpoofLDM comparison score distribution, only individual PAI species are considered. On closer inspection, it can be observed that the T-shirt attacks from FaceSpoofLDM in particular have the greatest influence on the distribution shift. We note that although the synthetic score distribution is far from the mated comparison scores, potential failure cases may still occur, such as synthetic samples resembling existing training images. Specifically, if certain combinations of soft biometric characteristics and PAI species are barely represented in the training dataset, the probability of unintentional replication of the training data increases. This highlights a residual privacy concern, particularly for cases with limited training data (e.g. T-shirt attacks). However, in our case, this risk has only limited impact, since the training databases used for FaceSpoofLDM (SynthASpoof and TFPA) consist exclusively of synthetic identities, so that no real identities can be disclosed.

B. CROSS-DOMAIN FACE PAD EVALUATION

To evaluate the impact of the synthetic identities (PAs and non attacks) generated by FaceSpoofLDM on PAD performance, we conducted extensive cross-domain detection performance evaluations using four well-known face PAD databases. As our PAD evaluation is image-based, we sampled frames from the following video datasets. Additionally, we generated

a synthetic face PAD training database using our proposed model. The databases used are as follows:

- **OULU-NPU (O)** [71]: Contains 990 bona fide videos and 3,960 PA videos of 55 subjects, recorded with six different mobile devices across three sessions.
- **CASIA-FASD (C)** [72]: Contains 150 bona fide videos and 450 PA videos of 50 subjects, recorded in low, normal, and high quality.
- **Idiap Replay-Attack (I)** [73]: Contains 200 bona fide videos and 1,000 PA videos of 50 subjects, captured under different illumination conditions.
- **MSU-MFSD (M)** [74]: Consists of 70 bona fide videos and 210 PA videos of 35 subjects, recorded with two different mobile devices.
- **Diffused Spoof (S)**: Synthetic identities (PA samples and non-attack samples) generated by FaceSpoofLDM. The database is equally balanced based on gender and ethnicity and contains 900 synthetic non-attack images and 3,240 synthetic PA images.

Since the above-listed databases (O, C, I, and M), are commonly used to evaluate cross-dataset detection performance face PAD algorithms, exclusively focusing on variations of print and replay attacks, we aligned our synthetic database in terms of the bona fide attack ratio and PAI species. Consequently, for this experimental setup, our synthetic database S includes synthetic non-attacks as well as synthetic print and replay attacks (Samsung-Replay and Webcam-Replay).

The conducted experiments are based on a cross-domain database evaluation. To this end, we followed widely used evaluation protocols for cross-domain PAD [7], [8], [18], [75]. This approach follows a leave-one-out principle, where we train on $N-1$ databases and evaluate on the remaining one. For example, we train a PAD algorithm on the combined training data constituted by the databases O, C and I and test the algorithm performance on the remaining database M (denoted as OCI→M). Since we want to evaluate the effectiveness of our synthetic identities on the PAD performance, we conducted the experimental setup both with and without the synthetic database S for comparison (e.g. OCI→M versus OCIS→M). Since the aforementioned databases O, C, M and I contain video files, we sampled frames from each video across the video duration to collect images for algorithm training and evaluation. The number of frames sampled per video depends on the database, as database sizes vary significantly. To prevent overfitting during training to any particular database, we ensured that no database contributed disproportionately more frames than others. We sampled 25 frames for M, 12 frames for C, 2 frames for O, and 6 frames for I, maintaining a balanced training dataset across all databases used. Finally, we used the MTCNN face detector [76] to detect and crop faces for each sampled frame, as well as to crop face images from the synthetic database S. The final number of sampled images after frame exctraction per class and per database, is reported in Table 5.

TABLE 5. Final number of sampled images per PAD database after frame extraction.

Dataset	Bona Fide Images	Attack Images
OULU-NPU (O) [71]	1,849	6,752
CASIA-FASD (C) [72]	1,594	4,406
Idiap Replay-Attack (I) [73]	1,046	5,267
MSU-MFSD (M) [74]	1,708	5,227

The cross-domain performance evaluation was conducted using 3 different PAD algorithms as well as 3 general-purpose networks, including two foundation models across 8 different database setups. The general-purpose approaches were selected because of their performance at different PAD benchmarks [77]. The PAD cross-evaluation performance results (without synthetic data) differ slightly from those in the original papers. This variation is due to our decision to sample a different number of frames per database to create a more balanced training dataset, whereas the original papers typically use the same number of frames per database. Moreover, since frames are often randomly sampled from the original PAD video files, precisely replicating the training database linked to the original papers is not possible. Additionally, we did not prioritize hyperparameter optimization, as our focus was on evaluating the contribution of our synthetic data from FaceSpoofLDM.

The algorithms are described as follows:

PAD Networks

- **DeePixBiS (ICB 2019)** [9]: George et al. proposed a CNN-based framework for face PAD utilizing deep pixel-wise supervision. This approach utilises pixel-wise binary labels (bona fide or PA) depending on the input image.
- **LMFD-PAD (WACV 2022)** [7]: Fang et al. proposed a dual-stream CNN framework that combines frequency and RGB domain features for improved robustness. One data stream adapts frequency filters to learn sensor and illumination invariant features, while the other stream uses the RGB images to augment the frequency domain features.
- **CF-PAD (WACV 2024)** [8]: In this work Fang et al. proposed an efficient domain generalization (DG) approach for face PAD by modelling it as a compound DG task from a causal perspective. CF-PAD leverages counterfactual intervention to identify causal factors in high-level representations.

General Purpose Networks

- **Zero-shot-CLIP (DINOv2) (Face & Gesture 2025)** [77]: Gonzalez-Soler et al. demonstrated the effectiveness of a foundation model-based framework on an unrelated top-down task, adapting only a minimum number of parameters related to the classification header in the training phase. For this purpose, the pre-trained CLIP (Zero-shot-CLIP in this work) and DINOv2

FIGURE 8. Input image (top) and it's associated reconstruction (bottom), using a pretrained autoencoder VQ-4 [63]. Compression artifacts are most noticeable when zoomed in.

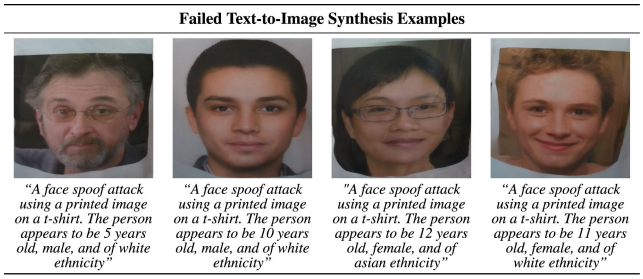

FIGURE 9. Examples of synthetic T-shirt attacks where the text prompt does not align with the generated image. The soft-biometric characteristics specified in the text-prompt are either not or barely present in the training dataset and therefore cannot be mapped to the image synthesis process. As a result, the generated image does not match the intended soft-biometric characteristics. Samples generated with 250 DDIM steps and $\eta = 1$.

(Zero-shot-DINOv2 in this work) foundation models were selected and their classification header modified to a single neuron, only optimised to produce a zero-shot PAD classification. Foundation models are large models pre-trained on large amounts of data, designed to be generalisable and easily adaptable to specific tasks.

- **EfficientNetV2 (ICML 2021)** [78]: Tan et al. introduced a novel CNN family that have faster training speed and better parameter efficiency than previous models. The authors use a combination of training-aware neural architecture search and scaling to jointly optimize training speed and parameter efficiency. It was combined with the framework proposed in [77] for zero-shot PAD. Therefore, the network weights are frozen, similar to the CLIP and DINO approaches in [77].

Following the cross-domain face PAD literature, we evaluated the results based on the Half Total Error Rate (HTER), Area Under the Receiver Operating Characteristic Curve (AUC) and the Bona fide Presentation Classification Error Rate (BPCER) [3] at a fixed Attack Presentation Classification Error Rate (APCER) [3] of 1%. All algorithms were trained for 20 epochs.

The individual cross-domain database evaluation results are demonstrated in Table 6, as the averaged results across the cross-domain setups are demonstrated in Table 7. There is a clear trend across the cross-domain setups: On average, adding synthetic data to the training set has a positive impact on the evaluation metrics. On average, the HTER slightly decreases due to the addition of synthetic data, while the AUC slightly increases across the setups. A particularly strong positive effect can be seen with the BPCER@APCER=1% (this allows us to vary the system's rigidity, i.e. how often the system falsely rejects authentic individuals when the security level is set to block only 1% of attacks), which is even more prominent for PAD algorithms (i.e. DeePixBiS, LMFD-PAD, CF-PAD) that were trained from scratch than for those that were developed for zero-shot PAD (i.e., only some network weights are updated). However, there are also a few setups where adding synthetic data does not improve all three evaluation metrics simultaneously (e.g. with DeePixBIS on ICM→O versus ICMS→O).

In [60] Fang et al. demonstrated that existing PAD databases (including the well-known databases O, C, I and M) exhibit imbalanced gender distributions, leading to biased PAD performance variations, particularly for female groups. Since subject diversity with regards to demographic variables in a PAD database are often associated to the region of the recording location, certain semantic facial patterns and demographic attributes may be overrepresented in existing PAD databases. It can therefore be assumed that in the cross-domain setups where a greater detection performance improvement is achieved by incorporating synthetic data (S), this is due to a better augmentation. However, there are no further ground truth demographic labels for the aforementioned PAD databases, which precludes a more precise analysis and therefore is beyond the scope of this paper. Another interesting finding is that in some setups the general-purpose algorithms (especially the foundation models) outperform the PAD algorithms (see. e.g. OMI→C and OMIS→C), demonstrating the generalisability of foundation models for unrelated tasks.

V. MODEL LIMITATIONS

As stated in [29], LDMs may encounter difficulties when high precision and fine-grained image synthesis is required. This is due to the autoencoders used for image compression, which introduce some degree of image quality loss. Since LDMs operate in latent space, an input image x is transformed from its pixel-space representation into its compressed latent representation z, leading to some loss of fine-grained details. Consequently, reconstructing the compressed representation back into pixel-space introduces deviations, as illustrated in Figure 8. However, it is arguable to what extent this loss of image quality affects PAD performance, as PA artefacts and not semantic facial features are usually considered to make a classification prediction. In general, when using LDM based image synthesis methods to augment high-risk scenarios where fine-grained details play a crucial role, such as medical image diagnostics, one should proceed with great caution, considering the aforementioned limitations.

Beyond architectural challenges, it is equally important to ensure sufficient data diversity to prevent the unintended

TABLE 6. Cross-domain performance evaluation. Metrics are reported in percent (%).

Method	OCI → M			OMI → C			OCM → I			ICM → O		
	HTER ↓	AUC ↑	BPCER@ APCER=1% ↓	HTER ↓	AUC ↑	BPCER@ APCER=1% ↓	HTER ↓	AUC ↑	BPCER@ APCER=1% ↓	HTER ↓	AUC ↑	BPCER@ APCER=1% ↓
DeePixBiS [9]	12.57	94.02	87.60	29.33	78.18	90.90	17.50	83.76	43.90	15.67	92.21	72.50
LMFD-PAD [7]	11.67	95.07	77.10	16.15	90.13	86.70	20.61	89.40	69.10	14.12	93.70	69.70
CF-PAD [8]	7.86	95.53	67.10	14.54	92.79	68.50	19.10	82.73	60.00	15.76	92.26	100.00
Zero-shot-CLIP (ViT-B-16) [77]	16.12	90.78	63.88	13.56	93.52	51.00	31.93	74.15	98.18	25.59	82.33	86.26
Zero-shot-DINOv2 (ViT-B-14) [77]	16.12	91.15	56.38	13.80	93.88	53.95	17.84	90.61	81.26	18.05	89.69	77.55
EfficientNetV2 (L) [78]	19.61	88.22	79.39	24.93	83.15	72.52	25.99	78.15	71.31	23.16	83.93	90.31
	OCIS → M			OMIS → C			OCMS → I			ICMS → O		
	HTER ↓	AUC ↑	BPCER@ APCER=1% ↓	HTER ↓	AUC ↑	BPCER@ APCER=1% ↓	HTER ↓	AUC ↑	BPCER@ APCER=1% ↓	HTER ↓	AUC ↑	BPCER@ APCER=1% ↓
DeePixBiS [9]	14.18	94.49	54.90	26.61	81.94	78.60	15.52	89.01	43.00	16.60	91.74	69.20
LMFD-PAD [7]	11.19	95.98	41.40	18.58	89.34	76.90	17.91	90.08	64.30	13.68	94.42	57.70
CF-PAD [8]	10.24	95.44	61.40	15.84	91.76	62.20	17.26	83.54	53.50	14.82	92.59	72.60
Zero-shot-CLIP (ViT-B-16) [77]	16.95	90.11	57.14	13.64	93.63	51.25	29.50	76.98	94.55	24.92	83.22	85.99
Zero-shot-DINOv2 (ViT-B-14) [77]	15.28	91.99	54.68	15.23	93.36	58.47	13.48	94.03	75.14	16.62	91.16	71.98
EfficientNetV2 (L) [78]	18.49	89.17	72.54	25.48	82.38	74.53	26.59	78.70	78.10	22.99	84.66	87.29

TABLE 7. Average Cross-domain performance evaluation. Metrics are reported in percent (%).

Method	Avg. without FaceSpoofLDM			Avg. with FaceSpoofLDM			Relative Improvement		
	HTER ↓	AUC ↑	BPCER@ APCER=1% ↓	HTER ↓	AUC ↑	BPCER@ APCER=1% ↓	HTER ↓	AUC ↑	BPCER@ APCER=1% ↓
DeePixBiS [9]	18.77	87.04	73.70	**18.23**	**89.30**	**61.40**	2.88	2.60	16.69
LMFD-PAD [7]	15.63	92.08	75.60	**15.34**	**92.46**	**60.00**	1.86	0.41	20.63
CF-PAD [8]	**14.32**	90.83	73.90	14.54	90.83	**62.40**	-1.54	-	15.56
Zero-shot-CLIP (ViT-B-16) [77]	21.8	85.20	74.83	**21.25**	**85.99**	**72.23**	2.52	0.93	3.47
Zero-shot-DINOv2 (ViT-B-14) [77]	16.45	91.33	67.29	**15.15**	**92.63**	**65.07**	7.90	1.42	3.30
EfficientNetV2 (L) [78]	23.42	83.36	78.38	**23.39**	**83.73**	**78.12**	0.13	0.44	0.33

replication of training data. We have found, that there are some cases, where the soft-biometric characteristics described in the text-prompt do not align with those in the resulting image. This is particularly noticeable in T-shirt attacks. Examples of failed text-to-image synthesis can be seen in Figure 9. We want to note, that such failure-cases (soft-biometric characteristics provided via text-prompt that do not fully align with those in the resulting synthetic image) are primarily caused by the restricted subject diversity in the training dataset (see Table 2, TFPA). In these scenarios, the model lacks sufficient data diversity (e.g. certain demographic combinations such as a 5-year-old subject for the T-Shirt PAI species are missing in the training database), which prevents the model from learning and thus limits it ability to incorporate such demographic combinations into the image generation process. To address these limitations, various mitigation strategies can be considered in future work. Increasing training diversity for underrepresented PAI species and rare demographic combinations could (1) reduce text-to-image synthesis errors and (2) decrease the risk of unintentional replication of training data. Particularly with regard to future model training using authentic data, differential privacy (DP) techniques such as DP-RDM [79] could further limit potential training identity leakage. Integrating post-hoc filtering pipelines (e.g., semantic consistency checks between the generated image and the associated text prompt) can prevent erroneous PAD images from being used into subsequent PAD tasks. Furthermore, a systematic comparison of different conditioning mechanisms for the controllable demographic synthesis can be considered as a subject for future work.

VI. CONCLUSION

In this work, we proposed FaceSpoofLDM a latent diffusion model for the language-guided generation of synthetic face PAs and non-attacks across various scenarios and demographic groups. By leveraging our approach, we verified the effectiveness of incorporating our synthetic identities for PAD by extensive cross-domain detection performance evaluation. Our experiments highlighted that, on average, this incorporation improves security against PA. In addition to the privacy benefits and improved PAD detection performance, we demonstrated our model quality based on several experiments. Our proposed approach offers great cost and resource conservation advantages while minimizing the need for manually crafting physical attack types. By automating the generation of synthetic identities across various demographic groups, our method enhances scalability and adaptability across different attack scenarios. Furthermore, the face PA synthesis approach presented in this paper is not limited to this particular biometric modality. Given enough training data, this approach could be applied to other biometric modalities such as fingerprints or iris. This flexibility underlines the potential for broader biometric applications and could lead to more robust systems against PA across multiple biometric modalities.

REFERENCES

[1] S. Li and A. Jain, *Handbook of Face Recognition*, 2nd ed., Cham, Switzerland: Springer, 2011.

[2] W.-Y. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.

[3] *Information Technology–Biometric Presentation Attack Detection—Part 3: Testing and Reporting*, Standard ISO/IEC 30107-3, 2023.

[4] *Information Technology–Biometric Presentation Attack Detection—Part 1: Framework*, Standard ISO/IEC 30107-1, 2023.

[5] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Comput. Surv.*, vol. 50, no. 1, pp. 1–37, Jan. 2018.

[6] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, "Deep learning for face anti-spoofing: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5609–5631, May 2022.

[7] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper, "Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1131–1140.

[8] M. Fang and N. Damer, "Face presentation attack detection by excavating causal clues and adapting embedding statistics," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 6257–6267.

[9] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2019, pp. 1–8.

[10] G. Ozgur, E. Caldeira, T. Chettaoui, F. Boutros, R. Ramachandra, and N. Damer, "FoundPAD: Foundation models reloaded for face presentation attack detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Feb. 2025, pp. 697–707.

[11] G. Zhang, K. Wang, H. Yue, A. Liu, G. Zhang, K. Yao, E. Ding, and J. Wang, "Interpretable face anti-spoofing: Enhancing generalization with multimodal large language models," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2025, vol. 39, no. 9, pp. 9896–9904.

[12] S. Muhammad Ibrahim, M. Sohail Ibrahim, S. Khan, Y.-W. Ko, and J.-G. Lee, "Improving face presentation attack detection through deformable convolution and transfer learning," *IEEE Access*, vol. 13, pp. 31228–31238, 2025.

[13] A. Antil and C. Dhiman, "Unmasking deception: A comprehensive survey on the evolution of face anti-spoofing methods," *Neurocomputing*, vol. 617, Feb. 2025, Art. no. 128992.

[14] F. Abdullakutty, E. Elyan, and P. Johnston, "A review of state-of-the-art in face presentation attack detection: From early development to advanced deep learning and multi-modal fusion methods," *Inf. Fusion*, vol. 75, pp. 55–69, Nov. 2021.

[15] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao, and Z. Liu, "CelebA-spoof: Large-scale face anti-spoofing dataset with rich annotations," in *Proc. ECCV*, 2020, pp. 70–85.

[16] X. Guo, Y. Liu, A. K. Jain, and X. Liu, "Multi-domain learning for updating face anti-spoofing models," in *Proc. ECCV*, 2022, pp. 230–249.

[17] D. Wang, J. Guo, Q. Shao, H. He, Z. Chen, C. Xiao, A. Liu, S. Escalera, H. J. Escalante, Z. Lei, J. Wan, and J. Deng, "Wild face anti-spoofing challenge 2023: Benchmark and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 6380–6391.

[18] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, "NAS-FAS: Static-dynamic central difference network search for face anti-spoofing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 9, pp. 3005–3023, Sep. 2021.

[19] R. Ramachandra, S. Venkatesh, K. B. Raja, S. Bhattacharjee, P. Wasnik, S. Marcel, and C. Busch, "Custom silicone face masks: Vulnerability of commercial face recognition systems & presentation attack detection," in *Proc. 7th Int. Workshop Biometrics Forensics (IWBF)*, May 2019, pp. 1–6.

[20] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*, document 32016R0679, 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[21] *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, 2024.

[22] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, A. Morales, D. Lawatsch, F. Domin, and M. Schaubert, "Synthetic data for the mitigation of demographic biases in face recognition," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2023, pp. 1–9.

[23] A. Dörsch, C. Rathgeb, M. Grimmer, and C. Busch, "Detection and mitigation of bias in under exposure estimation for face image quality assessment," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2024, pp. 1–5.

[24] P. Melzi et al., "FRCSyn-onGoing: Benchmarking and comprehensive evaluation of real and synthetic data to improve face recognition systems," *Inf. Fusion*, vol. 107, Jul. 2024, Art. no. 102322.

[25] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10850–10869, Sep. 2023.

[26] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[27] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–20.

[28] F. Boutros, M. Huber, P. Siebke, T. Rieber, and N. Damer, "SFace: Privacy-friendly and accurate face recognition using synthetic data," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2022, pp. 1–11.

[29] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.

[30] I. Joshi, M. Grimmer, C. Rathgeb, C. Busch, F. Bremond, and A. Dantcheva, "Synthetic data in human analysis: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 4957–4976, Jul. 2024.

[31] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, T. J. Cashman, and J. Shotton, "Fake it till you make it: Face analysis in the wild using synthetic data alone," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3661–3671.

[32] E. Wood, T. Baltrušaitis, C. Hewitt, M. Johnson, J. Shen, N. Milosavljević, D. Wilde, S. J. Garbin, T. Sharp, I. Stojiljkovic, T. Cashman, and J. Valentin, "3D face reconstruction with dense landmarks," in *Proc. ECCV*, 2022, pp. 160–177.

[33] L. Colbois, T. D. F. Pereira, and S. Marcel, "On the use of automatically generated synthetic image datasets for benchmarking face recognition," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Aug. 2021, pp. 1–8.

[34] J. J. Engelsma, S. Grosz, and A. K. Jain, "PrintsGAN: Synthetic fingerprint generator," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6111–6124, May 2023.

[35] J. Priesnitz, C. Rathgeb, N. Buchmann, and C. Busch, "SynCoLFinGer: Synthetic contactless fingerprint generator," *Pattern Recognit. Lett.*, vol. 157, pp. 127–134, May 2022.

[36] P. Zhang, H. Dou, W. Zhang, Y. Zhao, Z. Qin, D. Hu, Y. Fang, and X. Li, "A large-scale synthetic gait dataset towards in-the-wild simulation and comparison study," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 1, pp. 1–23, Jan. 2023.

[37] A. Kordas, E. Bartuzi-Trokielewicz, M. Ołowski, and M. Trokielewicz, "Synthetic iris images: A comparative analysis between Cartesian and polar representation," *Sensors*, vol. 24, no. 7, p. 2269, Apr. 2024.

[38] P. Drozdowski, C. Rathgeb, and C. Busch, "Sic-gen: A synthetic iris-code generator," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2017, pp. 1–6.

[39] M. Falkenberg, A. Bensen Ottsen, M. Ibsen, and C. Rathgeb, "Child face recognition at scale: Synthetic data generation and performance benchmark," *Frontiers Signal Process.*, vol. 4, pp. 1–17, May 2024.

[40] F. V. S. Kouam, C. Rathgeb, M. Ibsen, and C. Busch, "SynChildFace: Fine-tuning face recognition for children with synthetic data," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, Dec. 2024, pp. 1–6.

[41] C. Rathgeb, M. Ibsen, D. Hartmann, S. Hradetzky, and B. Ólafsdóttir, "Testing the performance of face recognition for people with down syndrome," in *Proc. IEEE 18th Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2024, pp. 1–5.

[42] P. Tinsley, A. Czajka, and P. Flynn, "This face does not exist…But it might be yours! Identity leakage in generative models," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1319–1327.

[43] H. Zohny, J. McMillan, and M. King, "Ethics of generative AI," *J. Med. Ethics*, vol. 49, no. 2, pp. 79–80, 2023.

[44] P. Esser, R. Rombach, and B. Ommer, "A note on data biases in generative models," in *Proc. NeurIPS Workshop Mach. Learn. Creativity Design*, 2020, pp. 1–8.
[45] M. Fang, M. Huber, and N. Damer, "SynthASpoof: Developing face presentation attack detection based on privacy-friendly synthetic data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 1061–1070.
[46] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 12104–12114.
[47] M. Ibsen, C. Rathgeb, F. Brechtel, R. Klepp, K. Pöppelmann, A. George, S. Marcel, and C. Busch, "Attacking face recognition with T-shirts: Database, vulnerability assessment, and detection," *IEEE Access*, vol. 11, pp. 57867–57879, 2023.
[48] Y. Ma, C. Lyu, L. Li, Y. Wei, and Y. Xu, "Algorithm of face anti-spoofing based on pseudo-negative features generation," *Frontiers Neurosci.*, vol. 18, pp. 1–13, Apr. 2024.
[49] J. E. Tapia, F. Stockhardt, L. J. González-Soler, and C. Busch, "SynID: Passport synthetic dataset for presentation attack detection," 2025, *arXiv:2505.07540*.
[50] S. A. Grosz and A. K. Jain, "SpoofGAN: Synthetic fingerprint spoof images," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 730–743, 2023.
[51] N. Kohli, D. Yadav, M. Vatsa, R. Singh, and A. Noore, "Synthetic iris presentation attack using iDCGAN," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 674–680.
[52] S. Yadav, C. Chen, and A. Ross, "Synthesizing iris images using RaSGAN with application in presentation attack detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2422–2430.
[53] E. Wenger, M. Bronckers, C. Cianfarani, J. Cryan, A. Sha, H. Zheng, and B. Y. Zhao, "Hello, It's me: Deep learning-based speech synthesis attacks in the real world," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2021, pp. 235–251.
[54] D. T. Nguyen, T. D. Pham, G. Batchuluun, K. J. Noh, and K. R. Park, "Presentation attack face image generation based on a deep generative adversarial network," *Sensors*, vol. 20, no. 7, p. 1810, Mar. 2020.
[55] X. Ge, X. Liu, Z. Yu, J. Shi, C. Qi, J. Li, and H. Kälviäinen, "DiffFAS: Face anti-spoofing via generative diffusion models," in *Proc. ECCV*, 2024, pp. 144–161.
[56] B. Zhang, X. Zhu, X. Zhang, and Z. Lei, "Modeling spoof noise by de-spoofing diffusion and its application in face anti-spoofing," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2023, pp. 1–10.
[57] N. Ko, Y. Jeong, and J. C. Ye, "Text-to-image synthesis for domain generalization in face anti-spoofing," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Feb. 2025, pp. 1850–1860.
[58] J. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–13.
[59] T. Zhang, "Diffusion based data augmentation for face anti-spoofing," in *Proc. 2nd Int. Conf. Mach. Learn. Autom., CONF-MLA*, Mar. 2025, pp. 1–11.
[60] M. Fang, W. Yang, A. Kuijper, V. Štruc, and N. Damer, "Fairness in face presentation attack detection," *Pattern Recognit.*, vol. 147, Mar. 2024, Art. no. 110002.
[61] M. Grimmer and C. Busch, "AgeDiff: Latent diffusion-based face age editing with dual cross-attention," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, Dec. 2024, pp. 1–6.
[62] Y. Wang, Z. Zhang, C. Chen, and F. Yang, "Facial privacy protection via attention-guided latent diffusion model for adversarial sample generation," *Signal, Image Video Process.*, vol. 19, no. 9, p. 749, Sep. 2025.
[63] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12868–12878.
[64] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.*, 2015, pp. 234–241.
[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2025, pp. 5998–6008.
[66] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[67] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8026–8037.
[68] Q. Feng, C. Guo, F. Benitez-Quiroz, and A. Martinez, "When do GANs replicate? On the choice of dataset size," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6681–6690.
[69] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A universal representation for face recognition and quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14220–14229.
[70] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 947–954.
[71] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "OULU-NPU: A mobile face presentation attack database with real-world variations," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 612–618.
[72] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, Mar. 2012, pp. 26–31.
[73] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2012, pp. 1–7.
[74] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 746–761, Apr. 2015.
[75] Z. Yu, X. Li, J. Shi, Z. Xia, and G. Zhao, "Revisiting pixel-wise supervision for face anti-spoofing," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 3, pp. 285–295, Jul. 2021.
[76] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[77] L. J. Gonzalez-Soler, J. E. Tapia, and C. Busch, "Are foundation models all you need for zero-shot face presentation attack detection?" in *Proc. IEEE 19th Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2025, pp. 1–10.
[78] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
[79] J. Lebensold, M. Sanjabi, P. Astolfi, A. Romero-Soriano, K. Chaudhuri, M. Rabbat, and C. Guo, "DP-RDM: Adapting diffusion models to private domains without fine-tuning," 2024, *arXiv:2403.14421*.

ANDRÉ DÖRSCH is currently pursuing the Ph.D. degree with the da/sec Research Group, National Research Center for Applied Cybersecurity (ATHENE), Hochschule Darmstadt, Germany. His research activities focus on synthetic face images and their potential for improving biometric security and fairness in biometric systems.

MARCEL GRIMMER received the Ph.D. degree from the Norwegian Biometrics Laboratory (NBL), Norwegian University of Science and Technology (NTNU), in 2025. His active research was dedicated to the generation of synthetic images in the context of face recognition, with a particular focus on face image quality assessment.

LAZARO JANIER GONZALEZ-SOLER received the B.Sc. degree in mathematics and computer science from the University of Havana, in 2014, and the Ph.D. degree in applied computer science from Hochschule Darmstadt, Germany, in 2022. He is currently a Postdoctoral Researcher with the da/sec Research Group, National Research Center for Applied Cybersecurity (ATHENE), Hochschule Darmstadt. His work focuses on biometric system security, particularly presentation attack detection (PAD) and morphing attack detection (MAD). He has received multiple awards, including the Best Ph.D. Thesis Award (German Universities of Applied Sciences), the Best Paper Award (WIFS 2021), and recognition for the best-performing algorithm in LivDet 2019 and 2021. He contributes to projects, such as Einstein, Bio4ensics, and RESPECT.

ROBERTO CASULA (Member, IEEE) received the Ph.D. degree in electronic and computer engineering from the University of Cagliari, Italy, in 2023, discussing a thesis called ''The Art of Fingerprint Spoofing.'' Since November 2015, he has been collaborating with the Pattern Recognition and Applications Laboratory (PRA Lab) in the field of fingerprint spoofing and fingerprint liveness detection. He is currently an Assistant Professor with PRA Lab, Department of Electrical and Electronic Engineering (DIEE). His research interests include fingerprint spoofing, fingerprint liveness detection, deepfake detection and analysis, and crowd detection and analysis.

GIAN LUCA MARCIALIS (Senior Member, IEEE) received the Ph.D. degree in electronic engineering and computer science from the University of Cagliari, Italy, in 2004. He is currently an Associate Professor with the University of Cagliari and the Research Director of the Biometric Unit of the Pattern Recognition and Applications Laboratory (PRA Lab), Department of Electrical and Electronic Engineering. His research interests are in the fields of biometrics, namely, fingerprint presentation attack detection, fingerprint classification, multiple classifiers for biometric identification, self-update-based biometric systems, facial deepfake detection, and anomaly group behavior in crowds.

CHRISTOPH BUSCH (Fellow, IEEE) is a member of NTNU-Gjøvik, Norway, and holds a joint appointment with Hochschule Darmstadt, Germany. Further, he has lectured with Denmark's DTU, since 2007. He was the initiator and participated in multiple projects on biometrics (e.g., 3D-Face, FIDELITY, and iMARS). He is also a PI in ATHENE. He is the Co-Founder of the European Association for Biometrics (EAB). He co-authored more than 600 articles. Furthermore, he is a convenor of WG3 in ISO/IEC JTC1 SC37.

CHRISTIAN RATHGEB is currently a Professor with the Faculty of Computer Science, Hochschule Darmstadt (HDA), Germany. He is also a Principal Investigator with the National Research Center for Applied Cybersecurity (ATHENE). His research interests include pattern recognition, iris and face recognition, the security aspects of biometric systems, secure process design, and privacy-enhancing technologies for biometric systems.