International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02132-5

CG-FAS: Cross-label Generative Augmentation for Face Anti-Spoofing

Xing Liu[2] · Anyang Su[1,4] · Minghui Wu[1] · Zitong Yu[3] · Kangle Wu[4] · Da An[4] · Jie Hao[1] · Mengzhen Xu[5] · Chenxu Zhao[1,4] · Zhen Lei[6,7,8]

Received: 31 July 2023 / Accepted: 22 May 2024

Abstract
Face Anti-Spoofing (FAS) is essential to secure face recognition systems from various physical attacks. A sufficient and diverse training set helps to build robust FAS models. To exploit the potential of FAS datasets, we propose to generate high-quality data including live and diverse presentation attacks (PAs) faces, for data augmentation during the model training stage. Our method is called **C**ross-label **G**enerative augmentation for **F**ace **A**nti-**S**poofing (CG-FAS), which could convert a live face into a 3D high-fidelity mask, replay, print, or other extra physical PAs. Correspondingly, CG-FAS can also restore a specific physical presentation attack into a live face. This function is realized by innovatively building an Interchange Bridge matrix, which stores disentangled spoof clues between PAs and live faces. To verify the effects of these generated data, we utilize them as augmentation data and conduct experiments on several typical FAS benchmarks. Extensive experimental results demonstrate the superior performance gain with CG-FAS for off-the-shelf data-driven FAS models. We hope the CG-FAS can shine a light on the deep FAS community to alleviate the data-hungry issue. The code will be released soon at: https://github.com/liuxingwt/CG-FAS.

Keywords Face Anti-Spoofing · Data augmentation · Generative model · Face editing

1 Introduction

Face recognition technique plays a crucial role in modern applications like access control system and electronic payment. Meanwhile, existing face recognition systems are exposed to diverse presentation attacks (PAs), such as the printed face (print attack), face replay on digital devices (replay attack), face covered by a mask (3D high-fidelity mask), etc. As a result, the Face Anti-Spoofing (FAS) (Yu

Communicated by Segio Escalera.

Xing Liu, Anyang Su and Minghui Wu have contributed equally.

✉ Chenxu Zhao
zhaochenxu@mininglamp.com

✉ Zhen Lei
zlei@nlpr.ia.ac.cn

Xing Liu
liuxing9406@gmail.com

Anyang Su
suanyang@mininglamp.com

Minghui Wu
1901120009@pku.edu.cn

Zitong Yu
yuzitong@gbu.edu.cn

Kangle Wu
wukangle@gmail.com

Da An
anda93456@gmail.com

Jie Hao
haojie@mininglamp.com

Mengzhen Xu
mzxu@mail.tsinghua.edu.cn

[1] Mininglamp Technology, Beijing, China

[2] Zelos Technology, Beijing, China

[3] Great Bay University, Dongguan, China

[4] Shanghai Artificial Intelligence Laboratory, Shanghai, China

[5] Department of Hydraulic Engineering, Tsinghua University, Beijing, China

[6] Institute of Automation, Chinese Academy of Sciences, Beijing, China

[7] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing, China

[8] Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong, China

Published online: 10 June 2024

et al., 2022) technique, which detects whether the presented face is live or not, becomes indispensable to defend face recognition systems.

In recent decade, researchers have proposed lots of data-driven deep-learning-based methods (Lucena et al., 2017; Xu et al., 2015; Shao et al., 2017; Liu et al., 2018a) to distinguish spoof faces from live ones. Most of them train the live/spoof detector via learning from pre-collected dataset (Yu et al., 2020). Despite satisfactory performance on pre-defined testing set, these well-trained FAS models usually encounter generalization challenges when deployed in real-world scenarios. This overfitting phenomenon is mainly contributed to the limited scale and diversity of datasets. Frequently used datasets (Boulkenafet et al., 2017b; Liu et al., 2018b, 2022) contain less than 100 identities and partial PAs due to expensive cost. Further, data augmentation study shows limited promotion in this field (Wang et al., 2023).

Motivated by the rapid development of generative models (Goodfellow et al., 2014; Karras et al., 2019; Rombach et al., 2022), we intend to generate face images to extend the diversity of public datasets, and we conjecture that these generated samples could help to promote FAS models. Generating data as augmentation for FAS models has been studied in Liu et al. (2020), Wu et al. (2021a), Jourabloo et al. (2018). However, These existing methods (Wang et al., 2023; Liu et al., 2020; Ruiz et al., 2023) appear limited effects as shown in Fig. 1: **(1) Low quality:** unsatisfactory generation quality with eye-perceived artifacts, which is easy to be seen in EPCR (Wang et al., 2023), DSDG (Wu et al., 2021a), STDN (Liu et al., 2020) and other generative methods. **(2) Rare diversity:** unable to generate arbitrary PAs with any input face, as DSDG (Wu et al., 2021a) like methods are not able to control the generation attributes precisely. **(3) Inconsistent generation:** hard to disentangle face characteristics from spoof features. Since Stable Diffusion (Rombach et al., 2022) based method DreamBooth (Ruiz et al., 2023) is able to use prompt to generate spoof faces, we can see that these generate faces indeed own spoof trait like 3D mask margin. But the face appearance is not consistent with inputs and changed apparently throughout generation.

To overcome these challenges, we propose a novel framework named **C**ross-label **G**enerative augmentation for **F**ace **A**nti-**S**poofing (CG-FAS). Using any FAS public dataset as input, CG-FAS could generate samples whose spoof labels are contrary to input images, while other characteristics are consistent as shown in the last column of Fig. 1. To disentangle spoof and spoof-irrelevant features, we execute the face editing in a highly disentangled latent space $\mathcal{W}+$ (Abdal et al., 2019), which ensures that the face identity information will not be exterminated throughout generation. What's more, an encoder and generator are trained to connect the RGB space and latent space. We utilize the prevalent StyleGAN (Karras et al., 2019) as generator, which is able to produce natural FAS faces superior to previous researches (Wu et al., 2021a; Liu et al., 2020). To exploit the advantages of the linear $\mathcal{W}+$ space, we organize each presentation attack's discriminative feature into an Interchange Bridge (IB) matrix, which can be used to generate images between arbitrary PA and live labels even for unseen face identities.

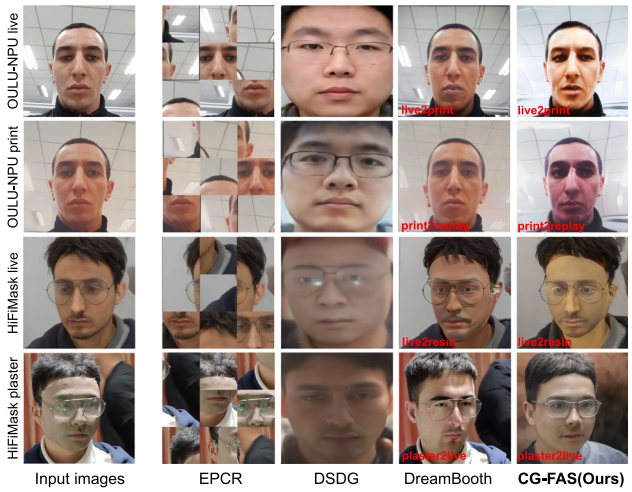

Fig. 1 Comparison of different FAS augmentation manners applied on HiFiMask (Liu et al., 2022) and OULU-NPU (Boulkenafet et al., 2017b) datasets. The first column lists live, print attack and plaster mask face images as input. The second column shows the results of Patch Shuffle Augmentation proposed in EPCR (Wang et al., 2023). Another typical GAN-based augmentation namely DSDG (Wu et al., 2021a) is compared in the third column. Further we conduct image manipulation with Stable Diffusion (Rombach et al., 2022) based method DreamBooth (Ruiz et al., 2023). The last column shows the result of our proposed CG-FAS, which is able to convert the input images' spoofing label and keep other face attributes consistent with input images

Given any face images from a public FAS dataset, CG-FAS firstly encodes images into low-dimensional latent codes in the $\mathcal{W}+$ space. In this latent space, it is flexible to index the IB matrix to obtain a residual vector, which represents an editing direction, such as live to print attack. The editing process is executed by adding the latent codes with the residual vector, controlled by an editing coefficient scalar. After that, the resultant vectors are eventually fed into the pre-trained generator to produce target PA face images. Adding these generated images into existing FAS datasets as augmentation, the proposed CG-FAS is demonstrated to obtain better FAS models.

The main contributions of this study are summarized below:

- We propose the Interchange Bridge(IB) matrix, which could be used to generate arbitrary live/PA faces while keeping spoof-irrelevant attributes consistent with input images.
- Applying the IB matrix as augmentation during FAS model training, we introduce a novel framework called

CG-FAS, which significantly enhances the performance of the FAS model.
- Evaluated on single-domain and cross-domain experiments, our proposed CG-FAS achieves competitive performance on several FAS benchmarks.

2 Related Work

2.1 Face Anti-Spoofing

Before the deep learning era, FAS researchers were keen on extracting handcrafted local features to distinguish live and PAs face images. The most commonly used features are LBP (Tiago et al., 2013; Boulkenafet et al., 2015), HOG (Komulainen et al., 2013), SIFT (Patel et al., 2016), and DoG (Boulkenafet et al., 2017), which show limited performance. In recent decades, FAS methods have indeed benefited from the huge breakthrough of deep neural networks (He et al., 2016a; Ronneberger et al., 2015) and large-scale datasets (Liu et al., 2018a; Zhang et al., 2020; Liu et al., 2021a, 2022; Fang et al., 2024a, b). A lot of deep learning based FAS methods (Liu et al., 2019; Menotti et al., 2015; Nagpal & Dubey, 2019; Jourabloo et al., 2018) have emerged.

The significant Central Difference Convolution Network (CDCN) (Yu et al., 2020) is proposed to improve the representation capacity of detailed textures via leveraging local gradient features. After that, dual-cross central difference networks (Yu et al., 2021) are proposed to exploit the difference of the center and surrounding sparse local features to alleviate the information redundancy and sub-optimal problem in the training stage. PatchNet (Wang et al., 2022a) utilize fine-grained face patch to enhance model's discriminative ability. Other works (Wang et al., 2022b; Sun et al., 2023) pay attention to the domain adaption problem in FAS task.

Some generated-based methods show impressive results by augmenting training data like STDN (Liu et al., 2020) and DSDG (Wu et al., 2021a). However, previous generated-based methods (Liu et al., 2020; Wu et al., 2021a) are limited to intra-dataset generation scenarios and the generated images do not seem as realistic as natural samples. In contrast, our proposed CG-FAS is able to flexibly generate vivid samples whose spoof label is different from the inputs and can be easily applied on unseen dataset.

2.2 Image Generation and Editing

2.2.1 Generative Methods

Generative methods are broadly studied and applied for image editing (Ling et al., 2021; Ruiz et al., 2023). We first introduce the recently prevailing generative methods, and image editing related advances later. Many popular generative paradigms are put forward like Auto-regressive models (Van Oord et al., 2016), Variational Autoencoder (VAE) (Diederik & Max, 2014), Generative Adversarial Network (GAN) (Goodfellow et al., 2014) and diffusion models (Sohl-Dickstein et al., 2015). Among all, diffusion models are popular but not easy to precisely control generation details with text prompt (Rombach et al., 2022). Since GAN-based methods are particularly concerned for generating high-quality and realistic samples (Arjovsky et al., 2017; Karras et al., 2018; Miyato et al., 2018), and generally applied in tasks like image-to-image translation (Isola et al., 2017), semantic image editing (Ling et al., 2021). We choose the distinguished StyleGAN (Karras et al., 2019, 2020b, a, 2021) network as our image generator (Wu et al., 2021b).

GAN inversion aims to invert real-world images into latent codes in the low-dimensional latent space (Xia et al., 2021), which is the reverse function of a GAN generator (Goodfellow et al., 2014). The latent space of GAN is generally studied and recognized as a Riemannian manifold (Shen et al., 2020a). The $\mathcal{Z}$ space is utilized by randomly samples a normal distribution vectors (Radford et al., 2016). StyleGAN utilizes a non-linear mapping network to convert a $\mathcal{Z}$ space latent code into $\mathcal{W}$ space, enabling better interpolation and disentangles (Karras et al., 2019, 2020b). Some researchers employ $\mathcal{W}+$ space, which extends $\mathcal{W}$ space to a better representation (Abdal et al., 2019, 2020). In this study, the cutting-edge e4e (Tov et al., 2021) method and $\mathcal{W}+$ space is chosen as our encoder module for high efficiency.

2.2.2 Face Image Editing

Face image editing technique is attractive for its versatility and beyond imagination results (Patashnik et al., 2021). A typical editing manner of StyleGAN-based researches obeys a paradigm of "invert first, edit later" (Richardson et al., 2021), which is conducted by firstly converting the given image into latent space, manipulating the latent code, and lastly generating the desired image by generator (Härkönen et al., 2020; Shen et al., 2020a). For instance, InterFaceGAN (Shen et al., 2020b) uses the SVM method to find semantic directions in $\mathcal{W}+$ space to revise face attributes like age, gender, and expression. GANSpace (Härkönen et al., 2020) applies PCA(Principal Component Analysis) to find meaningful direction and execute an interpolation manipulation in a BigGAN (Brock et al., 2018) or StyleGAN (Karras et al., 2019) latent space. StyleCLIP (Patashnik et al., 2021) enables natural language to edit input images, relying on a large-scale visual-language model CLIP (Radford et al., 2021). While these methods are rarely applied in FAS area, we are seeking to use them for improving typical FAS methods.

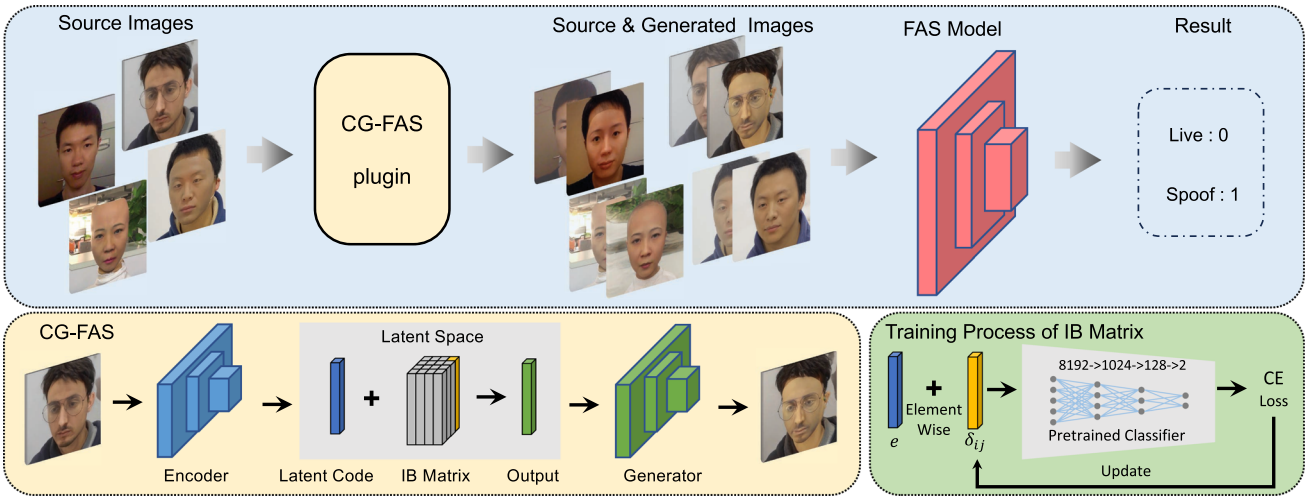

Fig. 2 An overview of the CG-FAS pipeline. The upper subpicture shows that our CG-FAS could serve as a plugin for any existing FAS methods. Fed with source images, CG-FAS could generate new samples as augmentation, which helps to improve the training of the FAS model. The lower-left subpicture illustrates the generation process of CG-FAS. Initially, a well-trained encoder maps the source face images into latent codes in the latent space. Subsequently, face editing is performed by adding the latent code with the IB matrix element. Finally, a StyleGAN generator is utilized to generate target images. The lower-right subpicture shows the training process of IB matrix. By adding latent codes with IB matrix element, the resultant vector is sent into a pre-trained classifier. Thereafter, the cross entropy loss is calculated to update the IB matrix element merely. **(Best viewed in color)**

3 Methodology

3.1 Overview

Since image generation techniques are rarely incorporated in contemporary FAS methods, we aim to utilize generative models to promote FAS model's performance. By elaborately designing the latent space and editing approach, we propose CG-FAS to generate new images whoes spoof labels are reversed while keeping other face attributes reserved. These generated samples are subsequently used as augmentation to train a more robust FAS model, which provably and practically performs better.

Our CG-FAS consists of three main stages: (1) Determining the latent space. The $\mathcal{W}+$ space is selected as our latent space for its convenient semantic editing capability. Consequently we train an encoder and a generator to connect RGB space and this latent space shown in Sect. 3.2. (2) Bridging live and PAs in latent space. After mapping RGB images into the semantic disentangled $\mathcal{W}+$ space, we are able to train and obtain each PA's unique spoof characteristics vector, and gather them into a matrix named Interchange Bridge (IB) matrix. The IB matrix can be used to transfer any face image's spoof label with a zero-shot ability introduced in Sect. 3.3. (3) Augmentation for FAS models. The IB matrix serves as a plug-in for training a FAS model, and its effectiveness is conceptionally proved in Sect. 3.4. When executing face editing on batch images, we found a dilemma of balancing FAS score and identification similarity score. We propose a strategy to reach a trade-off in Sect. 3.5. The overall pipeline of CG-FAS is illustrated in Fig. 2.

3.2 The Latent Space

Determining a proper latent space is vital for image editing tasks. While the RGB space is high-dimensional and unsuitable for image editing, regular GAN-based methods generate images from a low-dimensional latent space, which is convenient for editing. Among all, StyleGAN (Karras et al., 2020a) is popular for generating vivid images, and some typical latent spaces (Radford et al., 2016; Karras et al., 2019; Abdal et al., 2019) of StyleGAN are therefore put forward. The $\mathcal{W}+$ space (Abdal et al., 2019) is advanced and specialized in human face manipulation, we confidently choose the linear $\mathcal{W}+$ space as latent space. In this study, our $\mathcal{W}+$ space is a concatenation of 16 different 512-dimensional vectors, which could be used to generate 512×512 resolution face images.

To transfer the RGB space images into $\mathcal{W}+$ space, an indispensable component is training a generator as connection. In this study, the official StyleGAN2-ada (Karras et al., 2020a) is utilized as its generator could produce images obeying a similar distribution with input images. The Fréchet Inception Distances (FID) (Heusel et al., 2017) is used as supervision. During training, we add up the FFHQ dataset (Karras et al., 2019) and some FAS datasets as the complete training set, which is sufficient to produce faces with various characteristics. In summary, we define a generator f_G to

Algorithm 1 Training IB Matrix.

Require: Training set $\mathcal{S}$: image and label pair $(x, y) \in \mathcal{S}$
initialize IB matrix: $\mathcal{D}_{n+1}$ = zeros(n+1, n+1, d)
while not end of training **do**
  choose PA type i: i = randint[1, n]
  build dataset: $\mathcal{S}_i = \{(x, y) \mid y = 0 \ or \ i\}$
  get mini-batch data: $(X, Y) \subseteq \mathcal{S}_i$
  get residual vector: $\delta_{0i} = \mathcal{D}[0, \ i]$
  encode X into latent codes:
    $e = encoder(X)$
  edit latent codes with residual vector:
    $e' = e + \beta \cdot \delta_{0i} \cdot \ (2 \cdot \mathbb{1}_{Y \neq i} - 1)$
  compute target labels:
    $Y' = i - Y$
  compute loss:
    $\mathcal{L}_{bridge} = classifier(e', Y')$
  compute gradient $\Delta\mathcal{D}$:
    $\Delta\mathcal{D} = backward(\mathcal{L}_{bridge})$
  update Interchange Bridge Matrix:
    $\mathcal{D} \leftarrow \mathcal{D} - learning\_rate \cdot \Delta\mathcal{D}$
end while
$\mathcal{D}[i, \ j] = \mathcal{D}[0, \ j] - \mathcal{D}[0, i], \ for \ 1 \leq i < j \leq n$
$\mathcal{D}[i, \ j] = -\mathcal{D}[j, i], \ for \ 0 \leq i \neq j \leq n$
Indicator function $\mathbb{1}_c \in \{0, 1\}$ returns 1 if c is true.

map any latent code $e \in \mathbb{R}^d$ into RGB space image x by the following equation:

$$x = f_G(e), \tag{1}$$

where d is equal to 8192 in this study.

Reversely, to obtain the latent code e for any given image x, we train a deep neural network based encoder to map images into latent space. In this study, the e4e (Tov et al., 2021) encoder network is utilized to execute this mapping operation. The encoding procedure could be expressed by the formulation: $e = f_E(x)$. Generally, f_G and f_E are approximately inverse functions of each other and can be conveyed as follow:

$$e = f_E(x) = f_E \circ f_G(e), \tag{2}$$

where the notation $\circ$ is a link in composite function. When training the encoder, we choose the LPIPS loss (Zhang et al., 2018) and ArcFace (Deng et al., 2019) loss to compute total loss: $\mathcal{L}_{encoder} = \mathcal{L}_{LPIPS} + \lambda_{ID} \cdot \mathcal{L}_{ArcFace}$.

3.3 The Interchange Bridge Matrix

After the encoder and generator are well trained, we are able to manipulate images in latent space. For any given latent code e, commonly used semantic face editing (such as expression, age, and gender) approaches are finding a residual vector (Härkönen et al., 2020; Shen et al., 2020b), which represents a specific semantic edition in the disentangled $\mathcal{W}+$ space. In this study, we use a Multi Layer Perceptron named $classifier$ as supervision and calculate cross entropy loss to train each residual vector. For better usage, we organize a series of residual vectors into a matrix, namely the Interchange Bridge matrix described below:

$$\mathcal{D}_{n+1} = \begin{bmatrix} \delta_{00} & \delta_{01} & \cdots & \delta_{0n} \\ \delta_{10} & \delta_{11} & \cdots & \delta_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n0} & \delta_{n1} & \cdots & \delta_{nn} \end{bmatrix}, \tag{3}$$

Here n represents presentation attack categories. Any arbitrary element $\delta_{ij} \in \mathcal{D}_{n+1}$ denotes a residual vector that could convert live/PA type i into type j. To be mentioned, the IB matrix's diagonal elements are zero vectors, meaning no transformation within one live/PA type. Once we got the IB matrix, the editing operation can be expressed as adding latent code e with δ_{ij} by the following formulation:

$$e' = e + \beta \cdot \delta_{ij}, \tag{4}$$

where β is editing coefficient, and $\delta_{ij} \in \mathbb{R}^d$.

Since transforming live face into PA face and transforming PA face into live face are reverse manipulation in linear space $\mathcal{W}+$, naturally we get $\delta_{ij} = -\delta_{ji}$. Therefore the IB matrix $\mathcal{D}_{n+1}$ is skew-symmetric, which means that we only need to obtain the upper triangular part of the IB matrix. What's more, transforming live/PA from type i to type j can be considered as two separate steps: first transforming live/PA from type i into live face, then transforming live face into live/PA type j. Therefore, we have $\delta_{ij} = \delta_{i0} + \delta_{0j}$.

Depending on the relationships described above, we only need to train the first line elements of $\mathcal{D}_{n+1}$, while other elements could be obtained by these relationships. The whole training procedure is explicitly described in Alg.1. The matrix $\mathcal{D}_{n+1}$ can be utilized to generate arbitrary PA or live faces, thus we named it Interchange Bridge matrix, which shows zero-shot generation capability even for unseen face identities.

3.4 Effect Analysis of CG-FAS

Problem Definition Arranging the IB matrix as a plugin when training a FAS model, we conjecture this augmentation could help to promote the FAS model's performance, which is called CG-FAS in this study. In this subsection, we will demonstrate how CG-FAS could assist training FAS models by prevent overfitting. Firstly, we mathematically define a FAS task described in Eq. (5). For any input image $x \in \mathbb{X}$, there exists a corresponding label $y \in \{0, 1\}$ representing live or PA face respectively. Researchers aim to find an optimal mapping relation between input x and label y. Generally we use a deep neural network f_S to approximate this relationship, and the objective is minimizing the cross entropy (CE)

loss of model output and ground truth:

$$\min_{f_S} L(x, y; f_S) = \sum_{x,y} CE(f_S(x), y) \quad s.t.\ x \in \mathbb{X},\ y \in \{0, 1\}, \tag{5}$$

In the physical world, it's observed that different types of presentation attack, such as print attack and 3D mask attack, exhibit distinct spoofing characteristics. Base on this observation, we have

Assumption 1 Any two PAs' normalized residual vectors in $\mathbf{E}_s = \{s_i \mid s_i = \frac{\delta_{0i}}{\|\delta_{0i}\|},\ i = 1, \ldots, n\}$ are orthonormal, namely $s_i \cdot s_j^T = 0, \forall\, s_i, s_j \in \mathbf{E}_s$.

Since the linear $\mathcal{W}+$ space is proved to be highly disentangled (Abdal et al., 2019, 2020), showing that any attribute (e.g., glasses, hat) in live face can be encoded into a distinct embedding. Consequently, it is concluded that any live face can be expressed by a linear combination of such orthogonal embeddings, each corresponding to a specific attribute. Thus we have

Assumption 2 Any spoof-irrelevant face features (e.g., glasses, hat) could be conveyed by a set of orthonormal vectors $\mathbf{E}_b = \{b_i \mid i = 1, \ldots, m\}$ in $\mathcal{W}+$.

As illustrated in Fig. 3, CG-FAS is able to modify any face image's spoofing attribute without altering other spoof-irrelevant attributes. This leads us to believe that the spoofing features described in assumption 1 and features described in assumption 2 are orthogonal to each other. We have

Assumption 3 Any two elements in $\mathbf{E} = \mathbf{E}_s \cup \mathbf{E}_b$ are orthogonal, we get $e_i \cdot e_j^T = 0, \forall\, e_i, e_j \in \mathbf{E}$. Therefore $\mathbf{E} = [s_1, .., s_n, \ldots, b_1, \ldots, b_m]$ forms a standard orthogonal basis of latent space $\mathcal{W}+$, and arbitrary latent code e is equivalent to coordinate vector $\alpha = [\alpha_1, \ldots, \alpha_n, \ldots, \alpha_{n+m}]^T$ under this basis:

$$\begin{aligned} e &= \mathbf{E} \cdot \alpha \\ &= [s_1, .., s_n, b_1, \ldots, b_m] \cdot [\alpha_1, \ldots, \alpha_n, \ldots, \alpha_{n+m}]^T, \end{aligned} \tag{6}$$

In this study, the subset $\mathbf{E}_s = \{s_1, \ldots, s_n\}$ stands for different kinds of spoof clues like replay-attack texture, 3D mask margin features and so on. And $\mathbf{E}_b = \{b_1, \ldots, b_m\}$ stands for spoof-irrelevant information like lighting condition, scene, camera device which are independent of spoof clues. Based on the assumptions above, it's natural to define a valid FAS model to discriminate PA faces from live ones below:

Fig. 3 An illustration of the editing process on the HiFiMask dataset. By increasing the editing coefficient β progressively, the intermediate images are exhibited in detail. Most spoof-irrelevant attributes like hats, glasses and lighting conditions are reserved during the process

$$\begin{aligned} f_S(x) &= f_S \circ f_G(e) \\ &= [\underbrace{1, \ldots, 1}_{n}, \underbrace{1, \ldots, 1}_{t}, \underbrace{0, \ldots, 0}_{m-t}] \cdot \mathbf{E}^T \cdot e \\ &= \alpha_1 + \ldots + \alpha_n + \underbrace{\alpha_{n+1} + \ldots + \alpha_{n+t}}_{overfitting\ items}, \end{aligned} \tag{7}$$

In this FAS model, the coefficient from α_1 to α_{n+t} will determine the final discriminative result. Since the previous n items represent level of spoof features, the latter t spoof-irrelevant features are mistakenly considered as spoof clues which are overfitting items. Live faces with these features may be misidentified as PA faces. Thus, a better FAS model ought to own fewer overfitting items. In the following part, we are going to prove how our proposed CG-FAS could eliminate overfitting.

For a typical FAS dataset, we suppose that $b_1 \in \mathbf{E}_b$ represents a spoof-irrelevant but overfitted feature, which mostly occurs in PA samples but rarely occurs in live samples. In such circumstance, researchers tend to train a FAS model like Eq. (7), where the overfitting item α_{n+1} exists.

For any input images x, it is equal to a latent code e by Eq. (1), while e is equal to α under the basis $\mathbf{E}$ by Eq. (6). To simplify, we set $\{\alpha_i = 0\ or\ 1 \mid i \in \alpha\}$. Thus, x could be categorized into the following two types:

$$x \triangleq e \triangleq \alpha = \begin{cases} [\underbrace{0, \ldots, 0}_{n}, 0, \underbrace{0, \ldots, 0}_{m-1}]^T\ if\ x\ is\ live \\ [\underbrace{1, \ldots, 1}_{n}, 1, \underbrace{0, \ldots, 0}_{m-1}]^T\ if\ x\ is\ PA, \end{cases} \tag{8}$$

Once we conduct CG-FAS on x, m items of spoof-irrelevant features keep consistent and n items of spoof clues are converted. The generated sample x' could be expressed as:

$$x' = \text{CG-FAS}(x) \triangleq e' \triangleq \alpha' = \begin{cases} [\underbrace{1,\ldots,1}_{n}, 0, \underbrace{0,\ldots,0}_{m-1}]^T \ while\ x'\ is\ PA \\ [\underbrace{0,\ldots,0}_{n}, 1, \underbrace{0,\ldots,0}_{m-1}]^T \ while\ x'\ is\ live, \end{cases} \tag{9}$$

by adding up x and x' into training set, we can see that the $n+1\ th$ feature b_1 in Eq. (6) exists in both live and PA examples. Using the added images as training set, we tend to obtain a better FAS model where the overfitting item α_{n+1} in Eq. (6) no longer exists, thus relieving overfitting.

3.5 Batch Image Editing

Since we can use Eq. (9) to realize our desired editing task, as shown in Fig. 3, by progressively increasing the editing coefficient β in Eq. (4), input live faces are smoothly converted into 3D plaster mask faces as output. If we set β as low value, the face attributes similarity between input and output is high but the output's spoof degree is low. If we increase β, the output image's FAS score become higher but attributes similarity would get lower. Thus, determining the value of β becomes a vital problem.

To be mentioned, β is easy to be determined for a single image by delicate manual adjustment, but infeasible for large batch images which would exhaust too much human efforts. In this study, we will use a face recognition model f_R to evaluate spoof-irrelevant features' consistency and a FAS model f_S to evaluate spoof confidence score. When editing batches of live/spoof faces, we hope the f_R score between input and output keep close and f_S score be reversed after edition.

To measure our objective quantitatively, we use f_R and f_S score on the original validation set as standards, and we hope CG-FAS generated samples obey a similar distribution with validation set on the two scores. Thus, we calculate the average value of f_R and f_S score on the validation set, noted as $\tilde{t} = [t_R, t_S]$. The optimization problem can be conveyed as: finding an optimal value β^*, where generated images' f_R and f_S score point should be close to $\tilde{t}$, which is described below:

$$\beta^* = \arg\min_{\beta} \{ \mid \frac{1}{k}\sum_{i=1}^{k} f_R(f_G(e_i + \beta\delta),\ f_G(e_i)) - t_R \mid + \mid \frac{1}{k}\sum_{i=1}^{k} f_S(f_G(e_i + \beta\delta)) - t_S \mid \}, \tag{10}$$

Here k is batch size, e_i is the $i\ th$ latent code in this batch and δ is a residual vector in IB matrix $\mathcal{D}_{n+1}$. By trying different value of β in the equation above, we could figure out the approximately optimal value β^*.

4 Experiments

In this section, firstly we introduce the experimental settings, and then quantitatively evaluate the editing result by some expert models. Next, we compare our proposed CG-FAS with other contemporary methods on two intra-testing datasets, and conduct two cross-testing experiments. We execute ablation studies on four key factors throughout our research. Finally, we show more visualization results of IB matrix applied on four typical FAS datasets.

4.1 Experimental Settings

4.1.1 Datasets & Preprocessing

Four high resolution datasets namely OULU-NPU (Boulkenafet et al., 2017b), SiW (Liu et al., 2018a), HKBU MARsV2 (Liu et al., 2016) and HiFiMask (Liu et al., 2022, 2021b) are chosen as FAS datasets as shown in Table 1. Both OULU-NPU and SiW contain two categories of 2D PA: print and replay attack. MarsV2 is a 3D mask presentation attack dataset which includes live images two types of mask: ThatsMyFace and RealF masks. HiFiMask is a newly proposed 3D high fidelity dataset which contains live, transparent, plaster and resin masks. Besides, the FFHQ (Karras et al., 2019) dataset including 70000 identities face images is used while training StyleGAN. After executing the face detection and alignment operation, all images are cropped into 512×512 resolution as preprocssing.

4.1.2 Implementation Details

We choose StyleGAN2-ada (Karras et al., 2020a) configuration with a pre-trained model as the generator. While fine-tuning, we freeze the preceding ten layers and train other parameters with FFHQ and the four FAS datasets. For the encoder, we follow the implementation of e4e (Tov et al., 2021) network as our encoder and λ_{ID} set as 0.5. We select CDCN (Yu et al., 2020) as our FAS model backbone. During batch image edition, we set β as 0.22 in HiFiMask and 0.25 in OULU-NPU. Moreover, we set the ratio of generated images to original images as 1.0 when applying IB matrix on FAS tasks. Eight NVIDIA RTX-2080 GPUs are employed during training.

4.1.3 Performance Metrics

For intra testings, we strictly follow the protocols and evaluation metrics of OULU-NPU and HiFiMask. APCER (Attack Presentation Classification Error Rate) and BPCER (Bona Fide Presentation Classification Error Rate) are computed first, and their mean value ACER (Average Classification Error Rate) is used as the evaluation metric. During cross

Table 1 Four FAS datasets used in our experiments

Dataset	Live	2D PA	3D PA
OULU-NPU (Boulkenafet et al., 2017b)	✓	Print, replay	✗
SiW (Liu et al., 2018a)	✓	Print, replay	✗
HiFiMask (Liu et al., 2022)	✓	✗	Transparent, plaster, resin
MARsV2 (Liu et al., 2016)	✓	✗	RealF, ThatsMyFace

'PA' is short for presentation attack

testings, HTER (Half Total Error Rate) value and AUC (Area Under Curve) value are calculated as the evaluation metric.

4.2 Analyzing Editing Result

As shown in Fig. 3, we demonstrate the complete editing process on images of HiFiMask. By changing the editing coefficient β progressively, live faces turn into 3D high-fidelity masks smoothly. Attributes like glasses, expression, skin color, hat, and light condition are perfectly preserved after generation, which shows the huge advantages of our proposed CG-FAS.

Furthermore, we conducted a group of experiments to evaluate generated images quantitatively. As shown in Table 2, the first column lists the original testing set and CG-FAS generated sets which remain to be evaluated, while the last column lists two training sets on which we train two expert models. The middle columns display the comparison results on HiFiMask and OULU-NPU protocol 1. Evaluated on HiFiMask trained expert model, the testing ACER value on the original testing set is 1.3, while the generated testing set is 3.1. These two values are extremely close and near zero, meaning that our generated data owns the same spoof clues as the original HiFiMask. Additionally, when using the model trained on OULU-NPU dataset as expert, ACER on generated sets is 0.9, near the original testing set result as well.

4.3 Intra Testing

4.3.1 Result on OULU-NPU

OULU-NPU is a widely used evaluation dataset designed for 2D presentation attacks. There are four protocols on OULU-NPU by allocating different identities, PA types, devices and sessions. As shown in Table 3, we apply our proposed CG-FAS framework on the four protocols' training tasks, achieving the best performance on each protocol. Compared with other contemporary methods, CG-FAS shows evident superior performance on protocol three and protocol four, which verifies its strong generalization ability on hard examples. This experiment demonstrates the superiority of our proposed CG-FAS on 2D presentation attacks.

4.3.2 Result on HiFiMask

We further conduct an intra testing experiment on a 3D mask dataset called HiFiMask. It is a newly released 3D high resolution mask dataset which contains three representative masks of transparent, plaster and resin materials. HiFiMask dataset gathered various identities, lighting conditions, scenes and devices, while three protocols were raised by these rules. Within its training set, we utilize our CG-FAS framework to generate mask faces from live ones, and live faces from mask ones. After adding these generated images into the training set, our FAS model acquires state-of-the-art ACER by considerable advantage on all three protocols, which is shown in Table 4.

Table 2 Evaluation of the original and CG-FAS generated testing set (marked as ✓) on two FAS models, which are well-trained on HiFiMask and OULU-NPU training set respectively

Evaluated set	APCER (%)	BPCER (%)	ACER (%)	Training set
HiFiMask	0.8	1.9	1.3	HiFiMask
HiFiMask ✓	0.5	5.6	3.1	
HiFiMask	39.5	24.2	21.5	OULU-NPU
HiFiMask ✓	18.8	5.6	3.1	
OULU-NPU	35.8	36.7	36.3	HiFiMask
OULU-NPU ✓	0.4	54.6	27.5	
OULU-NPU	0.0	0.0	0.0	OULU-NPU
OULU-NPU ✓	1.7	0.2	0.9	

All models use CDCN as backbone

Table 3 The intra-dataset testing results on OULU-NPU

Prot	Method	APCER (%)	BPCER (%)	ACER (%)
1	FAS-SGTD (Wang et al., 2020)	2.0	0.0	1.0
	STDN (Liu et al., 2020)	0.8	1.3	1.1
	CDCN (Yu et al., 2020)	0.4	1.7	1.0
	BCN (Yu et al., 2020)	0.0	1.6	0.8
	LMFD-PAD (Fang et al., 2022)	1.4	1.6	1.5
	CDCN++ (Yu et al., 2020)	0.4	0.0	0.2
	NAS-FAS (Yu et al., 2020)	0.4	0.0	0.2
	DCN (Zhang et al., 2021)	1.3	0.0	0.6
	DSDG (Wu et al., 2021a)	0.6	0.0	0.3
	PatchNet (Wang et al., 2022a)	0.0	0.0	**0.0**
	CG-FAS (Ours)	0.0	0.0	**0.0**
2	STASN (Yang et al., 2019)	4.2	0.3	2.2
	FAS-SGTD (Wang et al., 2020)	2.5	1.3	1.9
	STDN (Liu et al., 2020)	2.3	1.6	1.9
	BCN (Yu et al., 2020)	2.6	0.8	1.7
	CDCN (Yu et al., 2020)	1.5	1.4	1.5
	LMFD-PAD (Fang et al., 2022)	3.1	0.8	2.0
	CDCN++ (Yu et al., 2020)	1.8	0.8	1.3
	NAS-FAS (Yu et al., 2020)	1.5	0.8	**1.2**
	DCN (Zhang et al., 2021)	2.2	2.2	2.2
	DSDG (Wu et al., 2021a)	1.5	0.8	**1.2**
	PatchNet (Wang et al., 2022a)	1.1	1.2	**1.2**
	CG-FAS (Ours)	0.7	1.7	**1.2**
3	STDN (Liu et al., 2020)	1.6±1.6	4.0±5.4	2.8±3.3
	FAS-SGTD (Wang et al., 2020)	3.2±2.0	2.2±1.4	2.7±0.6
	BCN (Yu et al., 2020)	2.8±2.4	2.3±2.8	2.5±1.1
	CDCN (Yu et al., 2020)	2.4±1.3	2.2±2.0	2.3±1.4
	LMFD-PAD (Fang et al., 2022)	3.5±3.2	3.3±3.2	3.4±3.1
	CDCN++ (Yu et al., 2020)	1.7±1.5	2.0±1.2	1.8±0.7
	NAS-FAS (Yu et al., 2020)	2.1±1.3	1.4±1.1	1.7±0.6
	DCN (Zhang et al., 2021)	2.3±2.7	1.4±2.6	1.9±1.6
	DSDG (Wu et al., 2021a)	1.2±0.8	1.7±3.3	1.4±1.5
	PatchNet (Wang et al., 2022a)	1.8±1.5	0.6±1.2	1.2±1.3
	CG-FAS (Ours)	1.3±1.2	0.8±1.4	**1.0±0.6**
4	FaceDs (Jourabloo et al., 2018)	1.2±6.3	6.1±5.1	5.6±5.7
	BCN (Yu et al., 2020)	2.9±4.0	7.5±6.9	5.2±3.7
	FAS-SGTD (Wang et al., 2020)	6.7±7.5	3.3±4.1	5.0±2.2
	STDN (Liu et al., 2020)	2.3±3.6	5.2±5.4	3.8±4.2
	LMFD-PAD (Fang et al., 2022)	4.5±5.3	2.5±4.1	3.3±3.1
	CDCN++ (Yu et al., 2020)	4.2±3.4	5.8±4.9	5.0±2.9
	NAS-FAS (Yu et al., 2020)	4.2±5.3	1.7±2.6	2.9±2.8
	DCN (Zhang et al., 2021)	6.7±6.8	0.0±0.0	3.3±3.4
	DSDG (Wu et al., 2021a)	2.1±1.0	2.5±4.2	2.3±2.3
	PatchNet (Wang et al., 2022a)	2.5±3.8	3.3±3.7	2.9±3.0
	CG-FAS (Ours)	3.8±5.2	0.0±0.0	**1.9±2.6**

Best results are marked in bold

Table 4 The intra-dataset testing results on HiFiMask

Prot	Method	APCER (%)	BPCER (%)	ACER (%)
1	ResNet50 (He et al., 2016b)	3.7	5.7	4.7
	Aux.(Depth) (Liu et al., 2018a)	4.9	1.8	3.4
	CDCN (Yu et al., 2020)	3.3	3.9	3.6
	CCL(ViT* (Dosovitskiy et al., 2021))	2.4	1.5	1.9
	CCL (Liu et al., 2022)	2.1	3.1	2.6
	DSDG (Wu et al., 2021a)	0.8	1.9	1.3
	CG-FAS (Ours)	0.7	1.4	**1.1**
2	ResNet50 (He et al., 2016b)	22.4±15.3	14.6±6.7	18.5±11.0
	Aux.(Depth) (Liu et al., 2018a)	11.1±9.4	11.2±9.8	11.2±9.0
	CDCN (Yu et al., 2020)	12.6±7.3	16.8±15.6	14.7±11.4
	CCL(ViT*) (Dosovitskiy et al., 2021)	12.2±10.3	12.9±11.2	12.5±10.7
	CCL (Liu et al., 2022)	10.7±7.5	10.7±9.4	10.7±8.4
	DSDG (Wu et al., 2021a)	9.5±8.6	6.2±5.0	7.8±6.8
	CG-FAS (Ours)	6.8±5.2	5.9±4.9	**6.3±5.0**
3	ResNet50 (He et al., 2016b)	13.5	28.3	20.9
	Aux.(Depth) (Liu et al., 2018a)	9.6	16.2	12.9
	CDCN (Yu et al., 2020)	20.8	12.5	16.7
	CCL(ViT*) (Dosovitskiy et al., 2021)	6.4	2.5	4.5
	CCL (Liu et al., 2022)	8.2	12.7	10.5
	DSDG (Wu et al., 2021a)	14.3	1.6	8.0
	CG-FAS (Ours)	12.0	3.3	**7.6**

Methods with '*' denote using pre-trained model
Best results are marked in bold

4.4 Cross Testing

We evaluate several state-of-the-art methods and our CG-FAS regarding the above protocols. In Table 6, all results from three protocols demonstrate a tendency that CG-FAS is superior to previous methods. This is mainly attributed to the ability of generating cross domain images by CG-FAS. Adding these generated data would change the original distribution of HiFiMask and OULU-NPU, which improves the FAS model's generalization ability.

4.4.1 HiFiMask & MARsV2

Since each FAS dataset owns its unique spoofing characteristics, cross dataset experiments could prove a method's generalization ability. Following the predecessor's paradigm (Liu et al., 2022), we choose two 3D mask datasets namely HiFiMask and MARsV2 to demonstrate our CG-FAS framework. Using HiFiMask dataset as training set, CG-FAS shows a significant metric improvement by a large margin when testing on MARsV2 dataset. On the contrary, we further use MARsV2 as the training set and HiFiMask as the testing set. The computed HTER and AUC value result also outperforms previous works shown in Table 5. This performance strongly proves the generalization ability of our proposed CG-FAS framework.

4.4.2 Cross-domain Attack Benchmark

To further demonstrate that our method is more effective in generating faces of unseen domains. In this part, we build specified settings as: (1) We combine the training set of OULU-NPU, the training set of HiFiMask, and the generated label transformation results of the above dataset as the actual training set. (2) We set the testing set of SiW as the actual testing set of protocol 1. This protocol ensures that the training and testing sets have the same 2D presentation attacks but different distributions. (3) We set the whole MARsV2 as the actual testing set of protocol 2. This protocol ensures that the training and testing sets have the same 3D presentation attacks but different distributions. Considering that MARsV2's overall data magnitude is relatively small, we choose all of its data as the testing set. (4) Finally, we combine the two above protocols as protocol 3.

4.5 Ablation Study

In this part, we execute four groups of experiments on four import factors in CG-FAS: FAS model backbone, editing coefficient β, the generated image number and generative methods. These experiments are conducted on OULU-NPU and HiFiMask datasets.

Table 5 Cross-testing results on two 3D presentation attack datasets

Method	HiFiMask to MARsV2		MARsV2 to HiFiMask	
	HTER (%)↓	AUC (%)↑	HTER (%)↓	AUC (%)↑
ResNet50 (He et al., 2016b)	20.61	86.87	38.96	67.05
CDCN (Yu et al., 2020)	16.56	90.81	45.20	56.13
Aux.(Depth) (Liu et al., 2018a)	9.31	96.31	44.24	57.05
CCL(ViT*) (Dosovitskiy et al., 2021)	9.82	96.72	38.03	63.07
CG-FAS (Ours)	**4.86**	**99.11**	**34.64**	**71.54**

Methods with '*' denote using pre-trained model
Best results are marked in bold

Table 6 Cross-domain results on our proposed cross-domain attack benchmark

Prot	Testing set	Method	APCER (%)	BPCER (%)	ACER (%)
1	SiW	CDCN (Yu et al., 2020)	28.34	29.16	28.75
		DSDG (Wu et al., 2021a)	26.41	26.07	26.24
		CG-FAS (Ours)	24.62	23.84	**24.23**
2	MARsV2	CDCN (Yu et al., 2020)	8.53	9.13	8.83
		DSDG (Wu et al., 2021a)	6.55	6.55	6.55
		CG-FAS (Ours)	6.15	6.55	**6.35**
3	SiW + MARsV2	CDCN (Yu et al., 2020)	22.11	22.17	22.14
		DSDG (Wu et al., 2021a)	19.46	19.50	19.48
		CG-FAS (Ours)	18.13	18.22	**18.17**

Best results are marked in bold

Table 7 Ablation study of CG-FAS with different backbones on OULU-NPU protocol 1

Method	ACER (%)	Method	ACER (%)	Method	ACER (%)
ResNet	6.8	Aux	1.6	CDCN	1.0
ResNet(DSDG)	4.8	Aux.(DSDG)	1.3	CDCN(DSDG)	0.3
ResNet(CG-FAS)	**3.5**	Aux.(CG-FAS)	**1.1**	CDCN(CG-FAS)	**0.0**

Best results are marked in bold

4.5.1 Impact of FAS Backbones

To verify the effectiveness of our proposed CG-FAS under any FAS model backbones, we select three prevalent networks: ResNet50 (He et al., 2016b), Aux. (Liu et al., 2018a) and CDCN (Yu et al., 2020) as backbones. The compared results are shown in Table 7. In the first row, there baseline results were performed by using these backbones. In the second row, we equipped the training set with DSDG generated data, the ACER value on three backbones all improved to some extent. In the last row, CG-FAS achieved the most competitive ACER value, which illustrates that our method does not rely on any specific backbone.

4.5.2 Impact of Editing Coefficient

As shown in Table 8, we conduct five group experiments by setting different values of β, ranging from 0.20 to 0.35. What's more, we further draw the similarity score vs. FAS score curve shown in Fig. 4 All experiments are conducted

Table 8 Ablation study of editing coefficient β on HiFiMask protocol 1

β	Avg. Similarity (%)	Avg. AUC (%)	Avg. Accuracy (%)
Val. set	39.77	99.73	98.34
0.20	44.72	89.97	85.22
0.22	**39.52**	**98.36**	**95.27**
0.25	36.07	99.16	96.89
0.30	31.01	99.67	98.24
0.35	26.72	99.85	98.69

Best results are marked in bold

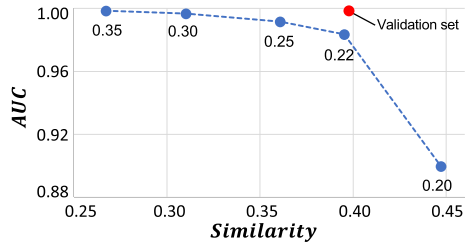

Fig. 4 The face recognition **similarity** score vs. FAS **AUC** score under different editing coefficient β. The red point is the result on HiFiMask validation set

on the HiFiMask protocol one's training set. When β is set as 0.35, the face recognition score between original training set and generated samples are low, which means that faces are over edited. And the face anti-spoofing score is high, meaning that we indeed obtain desired PA or live face images. When β is set as 0.20, we notice that the face recognition score is high while face anti-spoofing score is low. Since the curve is fitted with unavoidable error, we set the optimal β^* as 0.22 approximately, which is highly close to target point $\tilde{t}$. Thus $\beta^* = 0.22$ an approximately optimal solution of Eq. (10).

4.5.3 Impact of Generated Image Numbers

An intuitive question is how many generated samples are suitable as data augmentation. By adding different amount of generated images into training set, we use CDCN (Yu et al., 2020) as backbone and train models on HiFiMask protocol one. As shown in Table 9, adding 0.1 times generated images could promote the FAS model performance. And the ACER keeps improving until 1.0 times generated images are added to the training set. After that, the performance saturates when the ratio is set to 2.0 on HiFiMask. While 3.0 times of generated images are added into the training set, the ACER value degrades. We conjecture that too much generated images would alter the distribution of the original dataset by introducing model bias. Thus we set 1.0 as the best ratio throughout all experiment settings.

4.5.4 Impact of Generative Methods

As shown in Fig. 1, the Stable Diffusion (Rombach et al., 2022) based method DreamBooth (Ruiz et al., 2023) could also generate faces with PA trait, while other face features are not enough consistent with input images. Therefore, we also utilize the DreamBooth generated images as augmentation and test on HiFiMask dataset. Table 10 demonstrates that DreamBooth augmented model get better ACER value than the baseline CDCN model, but inferior to our proposed CG-FAS. We believe that the diffusion model based generation also works, but less convenient to use than our proposed IB matrix. Besides, it's hard to find proper prompt to align desired face editing tasks.

4.6 More Visualization Results

Aiming to show the superiority of our proposed IB matrix, we conduct more visualization experiments on the mentioned four FAS datasets. According to the difference of source and target images in domain and presentation attack, our generation results could be categorized into three types: intra-dataset generation, cross-domain generation and expanding-PA generation shown in Fig. 5.

Table 9 Ablation study of the ratio of CG-FAS generated images to original images during training

Generated/Original	0.0/1.0	0.1/1.0	0.5/1.0	1.0/1.0	2.0/1.0	3.0/1.0	1.0/0.1
ACER (%)	3.6	1.13	1.08	**1.05**	1.06	1.26	2.20

Experiments are conducted on HiFiMask protocol 1 dataset, and 1.0 times is equal to 32514 samples

Table 10 Ablation study of generative methods on HiFiMask dataset

Prot	Method	APCER (%)	BPCER (%)	ACER (%)
1	CDCN (Yu et al., 2020)	3.3	3.9	3.6
	DreamBooth (Ruiz et al., 2023)	1.1	1.6	1.3
	CG-FAS (Ours)	0.7	1.4	**1.1**
2	CDCN (Yu et al., 2020)	12.6±7.3	16.8±15.6	14.7±11.4
	DreamBooth (Ruiz et al., 2023)	9.8±8.3	6.9±4.5	8.3±6.4
	CG-FAS (Ours)	6.8±5.2	5.9±4.9	**6.3±5.0**
3	CDCN (Yu et al., 2020)	20.8	12.5	16.7
	DreamBooth (Ruiz et al., 2023)	13.6	3.8	8.7
	CG-FAS (Ours)	12.0	3.3	**7.6**

Best results are marked in bold

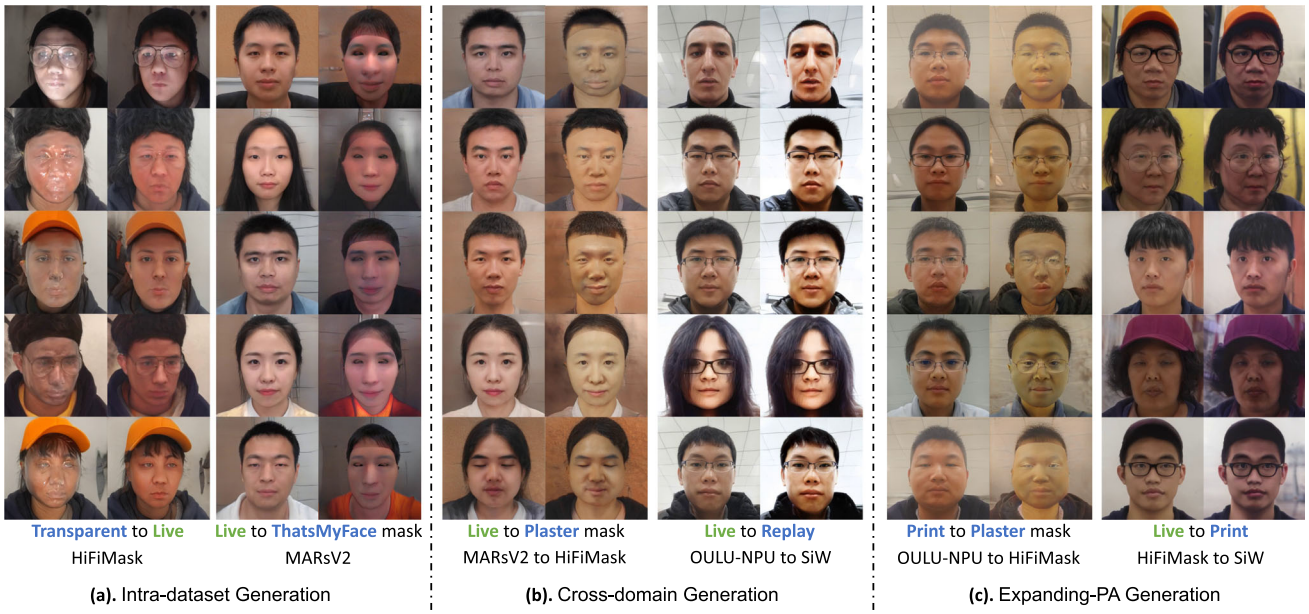

Fig. 5 An exhibition of applying the IB matrix on four FAS datasets. The figure is split into three parts by dotted line. **a** Intra-dataset Generation: The left two columns compare input transparent mask images from HiFiMask dataset and our generated live faces, while the right two columns show live faces from MARsV2 dataset and our generated ThatsMyFace mask images. **b** Cross-domain Generation: Both MARsV2 and HiFiMask are 3D mask datasets. Using MARsV2 live faces as input, we could generate HiFiMask plaster mask style faces. What's more, the OULU-NPU live faces can be used to generate SiW replay attack style faces. **c** Expanding-PA Generation: Here we use 2D PA dataset as input to generate 3D PA faces and reversely. OULU-NPU print attack faces can be used to generate HiFiMask plaster 3D mask faces. Besides, HiFiMask live faces are used to generate SiW print attack faces. **(Best viewed in color)**

4.6.1 Intra-dataset Generation

Firstly, we utilize our proposed CG-FAS to generate samples within a FAS dataset. Here we conduct experiments on HiFiMask dataset, which contains live, transparent and plaster and resin mask images. We select some transparent 3D mask faces from HiFiMask and convert them into their corresponding live ones. This cross-label generation results are shown in the first column of Fig. 5a. We also conduct the intra-dataset generation on another 3D mask FAS dataset, namely MARsV2 in the second column. It is clear that some selected live face images are converted into high-fidelity ThatsMyFace mask style images, which could alleviate the high expense of the ThatsMyFace mask data collection issue and facilitates data diversity during training.

4.6.2 Cross-domain Generation

In this study, we assume that cross domain datasets indicate the FAS datasets have the same PA types but different distribution. For instance, both OULU-NPU and SiW datasets contain print and replay attack images but with serious domain shifts. Figure 5b shows the process we transform the live face images from OULU-NPU into replay attack style images from SiW. Besides, we conduct the cross-domain experiments on two 3D mask FAS datasets MARsV2 and HiFiMask. Live faces from MARsV2 are transformed into plaster mask style images of HiFiMask with strong perceptive quality.

4.6.3 Expanding-PA Generation

Here we utilize CG-FAS to conduct generation across two datasets which contain different presentation attack types. For example, we are able to convert the live face images from OULU-NPU into the plaster mask style images of HiFiMask. As shown in Fig. 5c, such generated 3D mask attacks are visually high-quality. Besides, we can also convert live face images from HiFiMask into print attack style of SiW dataset. These generated face images contain obvious print features (e.g., color distortion) and show the ability of expanding PA types by our proposed CG-FAS.

4.6.4 T-SNE Visualization

Fig. 6 shows t-SNE visualization result on the protocol 1 training set of OULU-NPU dataset. As illustrated, live and generated live images obey similar distribution, as do PA and generated PA images. This similarity suggests that the generated images effectively extend the boundary of the original dataset, thereby serving as beneficial augmentations for training FAS models.

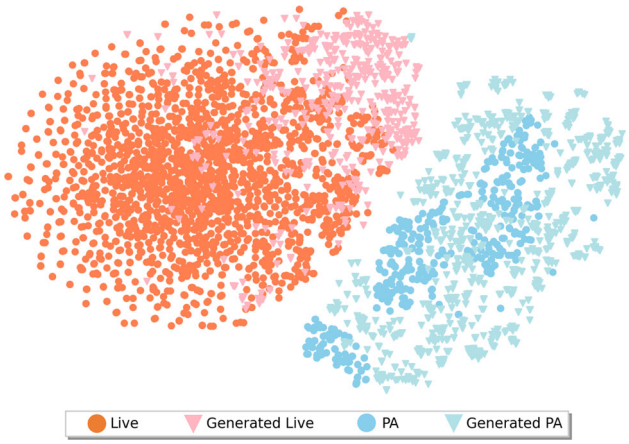

Fig. 6 Result of t-SNE visualization on OULU-NPU dataset. Circle markers represent samples from the original dataset, while triangle markers represent CG-FAS generated images

5 Conclusion and Future Work

In this study, we propose a novel Interchange Bridge matrix, which could convert a live face into a 3D high-fidelity mask, replay, print, or other extra physical presentation attacks. Correspondingly, it can also restore a specific physical presentation attack to a live face. Served as an augmentation manner, we put forward CG-FAS to promote the training of FAS models. To validate CG-FAS, we conduct experiments on both existing FAS benchmarks and a proposed cross-domain attack benchmark. Experimental results show that CG-FAS outperforms existing generation methods with a clear margin.

Vision Foundation Models (VFMs) like Stable Diffusion model also show impressive result in this study. Fine-tuning VFM on downstream tasks is widely studied but less used in face anti-spoofing domain. In the future, we seek for effective adaption of VFMs on FAS research.

Acknowledgements This work was supported by the Brain-like General Vision Model and Applications project (Grant No. 2022ZD0160402), Chinese National Natural Science Foundation Projects 62276254, U23B2054, and InnoHK program, Frontier Interdiscipline Project of Tsinghua University (20221080082), National Natural Science Foundation of China under Grant 62306061, and Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515140037).

Data Availability The data from four public face anti-spoofing datasets and one human face dataset (i.e., OULU-NPU (Boulkenafet et al., 2017b), SiW (Liu et al., 2018a), HiFiMask (Liu et al., 2022), HKBU MARsV2 (Liu et al., 2016) and FFHQ (Karras et al., 2019)) that support the findings of this study are available from the third party institutions (including University of Oulu, Michigan State University, Institute of Automation, Chinese Academy of Sciences, Hong Kong Baptist University and NVIDIA) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the above-mentioned third-party institutions.

References

Abdal, R., Qin, Y., & Wonka, P. (2019). Image2StyleGAN: How to embed images into the StyleGAN latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4432–4441.

Abdal, R., Qin, Y., & Wonka, P. A. (2020). Image2StyleGAN++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8296–8305.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, PMLR. pp. 214–223.

Boulkenafet, Z., Komulainen, J., & Hadid, A. (2015). Face anti-spoofing based on color texture analysis. In *International conference on image processing (ICIP)*.

Boulkenafet, Z., Komulainen, J., & Hadid, A. (2017). Face antispoofing using speeded-up robust features and fisher vector encoding. *IEEE Signal Processing Letters, 24*(2), 141–145.

Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., & Hadid, A. (2017b). OULU-NPU: A mobile face presentation attack database with real-world variations. In *FGR*, pp. 612–618.

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. In *International conference on learning representations*.

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699.

Diederik, P, K. & Max, W. (2014). Auto-encoding variational bayes. *International conference on learning representation*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Fang, H., Liu, A., Wan, J., Escalera, S., Zhao, C., Zhang, X., Li, S. Z., & Lei, Z. (2024a). Surveillance face anti-spoofing. *IEEE Transactions on Information Forensics and Security, 19*, 1535–1546.

Fang, H., Liu, A., Yuan, H., Zheng, J., Zeng, D., Liu, Y., Deng, J., Escalera, S., Liu, X., Wan, J., & Lei, Z. (2024b). *Unified physical-digital face attack detection*.

Fang, M., Damer, N., Kirchbuchner, F., & Kuijper, A. (2022). Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Härkönen, E., Hertzmann, A., Lehtinen, J., & Paris, S. (2020). GANSpace: Discovering interpretable GAN controls. *Advances in Neural Information Processing Systems, 33*, 9841–9850.

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, K., Zhang, X., Ren, S., & Sun, J. (2016b). Deep residual learning for image recognition. In *CVPR*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings*

of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.

Jourabloo, A., Liu, Y., & Liu, X. (2018). Face de-spoofing: Anti-spoofing via noise modeling. In *ECCV*.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *International conference on learning representations*.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems, 33*, 12104–12114.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020b). Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119.

Komulainen, J., Hadid, A., & Pietikäinen, M. (2013). Context based face anti-spoofing. In *2013 IEEE sixth international conference on biometrics: Theory, applications and systems (BTAS)*, pp. 1–8.

Ling, H., Kreis, K., Li, D., Kim, S. W., Torralba, A., & Fidler, S. (2021). EditGAN: High-precision semantic image editing. *Advances in Neural Information Processing Systems*, 34.

Liu, A., Tan, Z., Escalera, S., Guo, G., & Li, S. Z. (2021a). CASIA-SURF CeFA : A Benchmark for Multi-modal Cross-ethnicity Face Anti-spoofing. *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV),*, pp. 1179–1187.

Liu, A., Zhao, C., Yu, Z., Su, A., Liu, X., Kong, Z., Wan, J., Escalera, S., Escalante, H. J., Lei, Z., et al. (2021b). 3D high-fidelity mask face presentation attack detection challenge. In *Proceedings of the IEEE/CVF international conference on computer vision(ICCV) workshops*, pp. 814–823.

Liu, A., Zhao, C., Yu, Z., Wan, J., Su, A., Liu, X., Tan, Z., Escalera, S., Xing, J., Liang, Y., et al. (2022). Contrastive context-aware learning for 3D high-fidelity mask face presentation attack detection. *IEEE Transactions on Information Forensics and Security, 17*, 2497–2507.

Liu, S., Yang, B., Yuen, P. C., & Zhao, G. (2016). A 3D mask face anti-spoofing database with real world variations. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 100–106.

Liu, Y., Jourabloo, A., & Liu, X. (2018a). Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*.

Liu, Y., Jourabloo, A., & Liu, X. (2018b). Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*.

Liu, Y., Stehouwer, J., Jourabloo, A., & Liu, X. (2019). Deep tree learning for zero-shot face anti-spoofing. In *CVPR*.

Liu, Y., Stehouwer, J., & Liu, X. (2020). On disentangling spoof trace for generic face anti-spoofing. In *ECCV*.

Lucena, O., Junior, A., Moia, V., Souza, R., Valle, E., & Lotufo, R. (2017). Transfer learning using convolutional neural networks for face anti-spoofing. In *International conference image analysis and recognition*, pp. 27–34.

Menotti, D., Chiachia, G., Pinto, A., Schwartz, W. R., Pedrini, H., Falcao, A. X., & Rocha, A. (2015). Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Transactions on Information Forensics and Security, 10*(4), 864–879.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. In *International conference on learning representations*.

Nagpal, C. & Dubey, S. R. (2019). A performance evaluation of convolutional neural networks for face anti spoofing. In *2019 international joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., & Lischinski, D. (2021). StyleCLIP: Text-driven manipulation of StyleGAN imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2085–2094.

Patel, K., Han, H., & Jain, A. K. (2016). Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics & Security, 11*(10), 2268–2283.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, PMLR. pp. 8748–8763.

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. In *International conference on learning representations*.

Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., & Cohen-Or, D. (2021). Encoding in style: a StyleGAN encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2287–2296.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, Springer. pp. 234–241.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510.

Shao, R., Lan, X., & Yuen, P. C. (2017). Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3D mask face anti-spoofing. In *IJCB*, pp. 748–755.

Shen, Y., Gu, J., Tang, X., & Zhou, B. (2020a). Interpreting the latent space of GANs for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9243–9252.

Shen, Y., Yang, C., Tang, X., & Zhou, B. (2020b). InterFaceGAN: Interpreting the disentangled face representation learned by GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence., 44*(4), 2004–2018.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, PMLR. pp. 2256–2265.

Sun, Y., Liu, Y., Liu, X., Li, Y., & Chu, W.-S. (2023). Rethinking domain generalization for face anti-spoofing: Separability and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24563–24574.

Tiago, D., Anjos, A., Martino, J. D., & Marcel, S. (2013). Can face anti-spoofing countermeasures work in a real world scenario? In *International conference on biometrics*, pp. 1–8.

Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., & Cohen-Or, D. (2021). Designing an encoder for StyleGAN image manipulation. *ACM Transactions on Graphics (TOG), 40*(4), 1–14.

Van Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel recurrent neural networks. In *International conference on machine learning*, PMLR. pp. 1747–1756.

Wang, C.-Y., Lu, Y.-D., Yang, S.-T., & Lai, S.-H. (2022a). PatchNet: A simple face anti-spoofing framework via fine-grained patch recognition. In *CVPR*.

Wang, Z., Wang, Z., Yu, Z., Deng, W., Li, J., Gao, T., & Wang, Z. (2022b). Domain generalization via shuffled style assembly for face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4123–4133.

Wang, Z., Yu, Z., Wang, X., Qin, Y., Li, J., Zhao, C., Liu, X., & Lei, Z. (2023). Consistency regularization for deep face anti-spoofing. *IEEE Transactions on Information Forensics and Security, 18*, 1127–1140.

Wang, Z., Yu, Z., Zhao, C., Zhu, X., Qin, Y., Zhou, Q., Zhou, F., & Lei, Z. (2020). Deep spatial gradient and temporal depth learning for face anti-spoofing. In *CVPR*.

Wu, H., Zeng, D., Hu, Y., Shi, H., & Mei, T. (2021). Dual spoof disentanglement generation for face anti-spoofing with depth uncertainty learning. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(7), 4626–4638.

Wu, Z., Lischinski, D., & Shechtman, E. (2021b). StyleSpace analysis: Disentangled controls for StyleGAN image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12863–12872.

Xia, W., Zhang, Y., Yang, Y., Xue, J. -H., Zhou, B., & Yang, M. -H. (2021). GAN inversion: A survey. arXiv:2101.05278.

Xu, Z., Li, S., & Deng, W. (2015). Learning temporal features using LSTM-CNN architecture for face anti-spoofing. In *ACPR*, pp. 141–145.

Yang, X., Luo, W., Bao, L., Gao, Y., Gong, D., Zheng, S., Li, Z., & Liu, W. (2019). Face anti-spoofing: Model matters, so does data. In *CVPR*.

Yu, Z., Li, X., Niu, X., Shi, J., & Zhao, G. (2020). Face anti-spoofing with human material perception. In *ECCV*, Springer. pp. 557–575.

Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z., & Zhao, G. (2022). Deep learning for face anti-spoofing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(5), 5609–5631.

Yu, Z., Qin, Y., Zhao, H., Li, X., & Zhao, G. (2021). Dual-cross central difference network for face anti-spoofing. arXiv:2105.01290.

Yu, Z., Wan, J., Qin, Y., Li, X., Li, S. Z., & Zhao, G. (2020). NAS-FAS: Static-dynamic central difference network search for face anti-spoofing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(9), 3005–3023.

Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F., & Zhao, G. (2020). Searching central difference convolutional networks for face anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Zhang, K.-Y., Yao, T., Zhang, J., Liu, S., Yin, B., Ding, S., & Li, J. (2021). Structure destruction and content combination for face anti-spoofing. In *IJCB*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595.

Zhang, Y., Yin, Z., Li, Y., Yin, G., Yan, J., Shao, J., & Liu, Z. (2020). CelebA-Spoof: Large-scale face anti-spoofing dataset with rich annotations. In *European conference on computer vision*, Springer. pp. 70–85.