Controllable Inversion of Black-Box Face Recognition Models via Diffusion

Manuel Kansy[1,2] *, Anton Raël[1], Graziana Mignone[2], Jacek Naruniec[2], Christopher Schroers[2], Markus Gross[1,2], and Romann M. Weber[2]

[1]ETH Zurich, Switzerland, [2]DisneyResearch|Studios, Switzerland

{mkansy, grossm}@inf.ethz.ch, anrael@student.ethz.ch, {<first>.<last>}@disneyresearch.com

Abstract

*Face recognition models embed a face image into a low-dimensional identity vector containing abstract encodings of identity-specific facial features that allow individuals to be distinguished from one another. We tackle the challenging task of inverting the latent space of pre-trained face recognition models without full model access (*i.e. black-box *setting). A variety of methods have been proposed in literature for this task, but they have serious shortcomings such as a lack of realistic outputs and strong requirements for the data set and accessibility of the face recognition model. By analyzing the black-box inversion problem, we show that the conditional diffusion model loss naturally emerges and that we can effectively sample from the inverse distribution even without an identity-specific loss. Our method, named identity denoising diffusion probabilistic model (ID3PM), leverages the stochastic nature of the denoising diffusion process to produce high-quality, identity-preserving face images with various backgrounds, lighting, poses, and expressions. We demonstrate state-of-the-art performance in terms of identity preservation and diversity both qualitatively and quantitatively, and our method is the first black-box face recognition model inversion method that offers intuitive control over the generation process.*

1. Introduction

Face recognition (FR) systems are omnipresent. Their applications range from classical use cases such as access control to newer ones such as tagging a picture by identity or controlling the output of generative models [31, 2, 47]. The goal of an FR method f is to obtain embeddings $\mathbf{y}$ of face images $\mathbf{x}$ such that the embeddings of images of the same person are closer to each other than those of images of other people. We refer to this embedding $\mathbf{y}$ as the *identity vector* or *ID vector*. In this paper, we propose a technique to sample from $p(\mathbf{x}|\mathbf{y})$, *i.e.* to produce realistic face images from an ID vector.

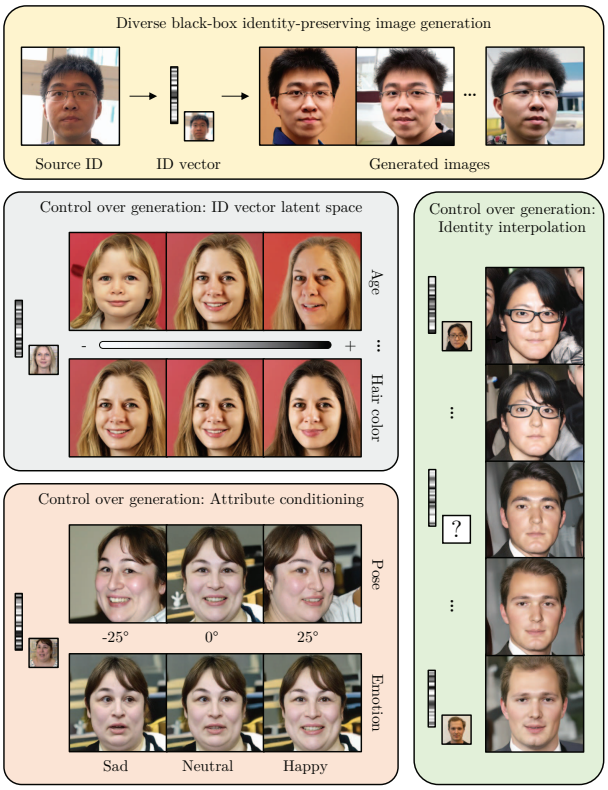

Figure 1: Overview. Our method inverts a pre-trained face recognition model (here InsightFace [13]) to produce high-quality identity-preserving images. It also provides intuitive control over the image generation process.

By design, the many-to-one mapping of FR methods assigns multiple images of a given identity to the same ID vector. The inverse one-to-many problem, *i.e.* producing a high-dimensional image from a low-dimensional ID vector, is extremely challenging. Previous methods often rely on the gradient of FR models either directly [74] or use it during training in the form of a loss function [9, 47]. This

*Corresponding author.

gradient or information about the model's architecture and weights is often not available, *e.g*. if using an API of a proprietary model. We therefore focus on the more generally applicable *black-box* setting, where only the resulting ID vectors are available. In addition to being more general, the black-box setting simplifies the analysis of different FR models as explored in the supplementary material. Another benefit is that we can easily extend our conditioning mechanism to include information from different, even non-differentiable, sources (*e.g*. labels, biological signals).

We propose the identity denoising diffusion probabilistic model (ID3PM), the first method that uses a diffusion model (DM) to invert the latent space of an FR model, *i.e*. to generate identity-preserving face images conditioned solely on black-box ID vectors as seen in Fig. 1. We show mathematically that we can effectively invert a model f even without access to its gradients by using a conditional DM. This allows us to train our method with an easy-to-obtain data set of pairs of images and corresponding ID vectors (easily extracting from images) without an identity-specific loss term.

Our method obtains state-of-the-art performance for the inversion task and is, to the best of our knowledge, the first black-box FR model inversion method with control over the generation process as seen in Fig. 1. Specifically, we can control (1) the diversity among samples generated from the same ID vector via the classifier-free guidance scale, (2) identity-specific features (*e.g*. age) via smooth transitions in the ID vector latent space, and (3) identity-agnostic features (*e.g*. pose) via explicit attribute conditioning.

To summarize, our main contributions are:

1. Showing that the conditional diffusion model loss naturally emerges from an analysis of the black-box inversion problem.
2. Applying the resulting framework to invert face recognition models without identity-specific loss functions.
3. Demonstrating state-of-the-art performance in generating diverse, identity-preserving face images from black-box ID vectors.
4. Providing control mechanisms for the face recognition model inversion task.

2. Related work

2.1. Face recognition

While early deep learning works such as DeepFace [64] and VGG-Face [48] treated FR as a classification problem (one class per identity), FaceNet [59] introduced the triplet loss, a distance-based loss function. The trend then shifted towards margin-based softmax methods [37, 67, 66, 14] that incorporate a margin penalty and perform sample-to-class rather than sample-to-sample comparisons. More recently, some FR methods tackle specific challenges such as robustness to different quality levels [33] and occlusions [36, 49].

2.2. Inversion of face recognition models

Similar to gradient-based feature visualization techniques [61, 38, 72, 44], Zhmoginov and Sandler [74] perform gradient ascent steps using the gradient of a pre-trained FR model to generate images that approach the same ID vector as a target image. To avoid generating adversarial examples, strong image priors such as a total-variation loss and a guiding image are necessary. Cole *et al*. [9] transform the one-to-many task into a one-to-one task by mapping features of an FR model to frontal, neutral-expression images, which requires a difficult-to-obtain data set. Nitzan *et al*. [47] map the identity features and attributes of images into the style space of a pre-trained StyleGAN [28] to produce compelling results. However, their method struggles to encode real images since it is trained exclusively with images generated by StyleGAN. Furthermore, all of the above methods require white-box access to (the gradient of) an FR model, which is not always available in practice.

Many black-box methods view the problem from a security lens, focusing on generating images that deceive an FR model rather than appearing realistic. Early attempts using linear [43] or radial basis function models [42] lacked generative capacity to produce realistic images. NbNet [40] introduces a neighborly de-convolutional neural network that can generate images with a reasonable resemblance to a given image, but it has line artifacts and relies on a huge data set augmented with a GAN. On the contrary, Razzhigaev *et al*. [52] propose a data-set-free method using Gaussian blobs (which we call "Gaussian sampling" for simplicity), but they need thousands of FR model queries (10-15 minutes) per image, and their results lack realism. Yang *et al*. [70] rely on background knowledge to invert a model and only produce blurry images in the black-box setting. Vec2Face [16] uses a bijection metric and knowledge distillation from a black-box FR model to produce realistic identity-preserving faces; however, it requires a large data set with multiple images per identity (Casia-WebFace [71]) during training. The method by Vendrow and Vendrow [65] (which we call "StyleGAN search") searches the latent space of a pre-trained StyleGAN2 [29] to find images with an ID vector close to the target. While their search strategy generates highly realistic images, it needs hundreds of FR model queries (5-10 minutes) per image and often lands in local minima, resulting in completely different identities.

Table 1 compares attributes of state-of-the-art FR model inversion methods. Ours is the only one that generates diverse, realistic, identity-preserving images in the black-box setting, can be trained with easy-to-obtain data, and only requires one FR model query during inference.

2.3. Diffusion models for inverse problems

A number of approaches for solving inverse problems in a more general setting using conditional [57, 55] and uncon-

Method	Black-box	FR model queries (inference)	Training data set	Realistic [1]	Mapping
Zhmoginov and Sandler [74]	No	~ 1000 [2] 1 [2]	Any images	No	One-to-one
Cole *et al*. [9]	No	1	Frontalized images	Yes	One-to-one
Nitzan *et al*. [47]	No	1	Any images	Yes	One-to-many
NbNet [40]	Yes	1	Huge data set	No	One-to-one
Gaussian sampling [52]	Yes	240000	Data-set-free	No	One-to-many
Yang *et al*. [70]	Yes	1	Any images	No	One-to-one
Vec2Face [16]	Yes	1	Multiple images per identity	Yes	One-to-many
StyleGAN search [65]	Yes	400	Data-set-free	Yes	One-to-many
ID3PM (Ours)	Yes	1	Any images	Yes	One-to-many

Table 1: Comparison of state-of-the-art face recognition (FR) model inversion methods. Our method does not have any of the common shortcomings, producing diverse, realistic images from black-box ID vectors with few requirements for the training data set or accessibility of the FR model during inference. [1] By visual inspection of the results of the respective papers. [2] The authors propose two methods: one taking hundreds or thousands of queries and the second one doing it in one shot.

ditional [26, 30, 62, 6, 7, 20, 5, 8, 3, 41, 63] exist; however, they mostly focus on image-to-image tasks such as inpainting and super-resolution whereas we focus on a vector-to-image task. The method by Graikos *et al*. [20] can generate images from low-dimensional, nonlinear constraints such as attributes, but it requires the gradient of the attribute classifier during inference whereas ours does not. Thus, conditional diffusion models with vectors as additional input [15, 51, 53, 56], while not directly geared towards inversion, are conceptually more similar to our approach.

3. Motivation

3.1. Inverse problems

In a system under study, we often have a *forward problem* or function f that corresponds to a set of observations $\mathbf{y} \sim \mathcal{Y}$. The function f has input arguments $\mathbf{x}$ and a set of parameters $\boldsymbol{\theta}$, such that $f(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{y}$. An *inverse problem* seeks to reverse this process and make inferences about the values of $\mathbf{x}$ or $\boldsymbol{\theta}$ given the observations $\mathbf{y}$. For the application explored in this work, f is a face recognition model that takes an image $\mathbf{x}$ as input and produces an ID vector $\mathbf{y}$.

When the function f is not bijective, no inverse exists in the traditional mathematical sense. However, it is possible to generalize our concept of what an inverse is to accommodate the problem of model inversion, namely by considering an inverse to be the set of pre-images of the function f that map ϵ-close to the target $\mathbf{y}$. For bijective f, this corresponds to the traditional inverse for $\epsilon = 0$.

3.1.1 Model inversion with model access

One way to handle the model-inversion problem when f is not bijective is to treat it pointwise, defining a loss, such as

$$\mathcal{L} = \frac{1}{2}\|\mathbf{y} - f(\mathbf{x})\|^2, \tag{1}$$

and minimizing it via gradient descent on $\mathbf{x}$ from some starting point $\mathbf{x}_0$ according to

$$\Delta\mathbf{x}_t = -\nabla_{\mathbf{x}}\mathcal{L} = \left(\frac{\partial f}{\partial \mathbf{x}}\right)^{\top} (\mathbf{y} - f(\mathbf{x}))\,. \tag{2}$$

In common cases where the inverse problem is one-to-many, we can take a statistical approach. Here we want to sample from $p(\mathbf{x}|\mathbf{y})$, which is equivalent to drawing from the pre-image set that defines the inverse $f^{-1}(\mathbf{y})$.

However, if we assume a Gaussian observation model

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; f(\mathbf{x}), \sigma^2\mathbf{I}) \propto \exp\left(-\frac{\mathcal{L}}{\sigma^2}\right), \tag{3}$$

where the last term follows from (1), then we can rewrite equation (2) as $\Delta\mathbf{x}_t \propto \sigma^2 \nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x}_t)$.

This shows that traditional model inversion via gradient descent performs a type of deterministic sampling from $p(\mathbf{y}|\mathbf{x})$—and not the distribution we want, $p(\mathbf{x}|\mathbf{y})$—by pushing toward modes of $p(\mathbf{y}|\mathbf{x})$ close to the initialization point $\mathbf{x}_0$, regardless of whether it possesses the desired characteristics of the data $p(\mathbf{x})$. This can lead to results, such as adversarial examples [19], that, while technically satisfying the mathematical criteria of inversion, do not appear to come from $p(\mathbf{x})$.

Various types of regularization exist to attempt to avoid this issue, which are most often *ad hoc* methods geared toward the specific problem at hand [74, 39, 10, 69]. A more general approach is to introduce regularization terms proportional to the *(Stein) score*, $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, since

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}) = \nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x}) + \nabla_{\mathbf{x}} \log p(\mathbf{x})$$

provides the *conditional* score needed to sample from $p(\mathbf{x}|\mathbf{y})$, the distribution we are actually interested in.

Previous work has shown that diffusion models (DMs) effectively learn the score $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, which allows them to be used alongside model gradients to guide sampling [63,

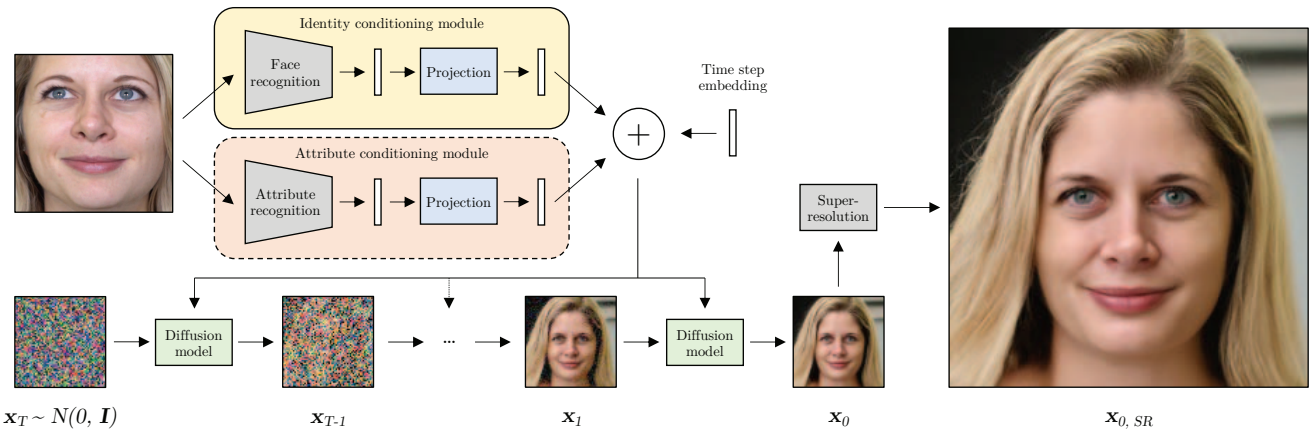

Figure 2: Method architecture. Given an image of a source identity, the identity conditioning module extracts the ID vector with a black-box, pre-trained face recognition network. This is projected with a fully connected layer and added to the time step embedding which is injected into the residual blocks of a diffusion model (DM). Starting with Gaussian noise $\mathbf{x}_T$, the DM iteratively denoises the image to finally obtain the output image $\mathbf{x}_0$ in 64×64 resolution. Lastly, the image is upsampled to a resolution of 256×256 using an unconditional super-resolution DM. The optional attribute conditioning module helps disentangle features and allows intuitive control over attributes such as the pose. Gray components are frozen during training.

23, 46, 15, 27]. When those models are classifiers, the procedure is known as *classifier guidance* [15]. However, this imposes an additional computational burden on sampling and also requires that the model f be differentiable.

3.1.2 Model inversion without full model access

In the case we focus on in this work, we assume to have access only to the values of the function f via some oracle or a lookup table of $(\mathbf{x}, \mathbf{y})$ pairs but not its gradient ∇f. In this case, also referred to as *black-box* setting, we may wish to train a function g_ψ to learn the inverse by minimizing

$$\mathcal{J} = \|\mathbf{x} - g_\psi(\mathbf{y})\|^2 \tag{4}$$

across all observed $\{(\mathbf{x}, \mathbf{y})\}$. Recalling that $\mathbf{y} = f(\mathbf{x})$, we have essentially described an encoder-decoder setup with the encoder frozen and only the decoder being trained, which requires no gradients from the "encoder" f.

If we consider perturbed data $\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_t^2 \mathbf{I})$. Then (4) is equivalent to

$$\begin{aligned} \tilde{\mathcal{J}} &= \|(\tilde{\mathbf{x}} - \mathbf{x}) - (\tilde{\mathbf{x}} - g_\psi(\mathbf{y}))\|^2 \\ &= \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, \mathbf{y}, t)\|^2, \end{aligned} \tag{5}$$

and we are now training a conditional model $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ to learn the noise added to $\mathbf{x}$ instead of a model g to reconstruct $\mathbf{x}$. This new task is exactly the one facing conditional diffusion models (Section 4.1).

Although we cannot *force* the model to leverage the conditioning on $\mathbf{y}$ or t, if it is to successfully minimize the loss $\tilde{\mathcal{J}}$, it should learn a function proportional to the conditional score. That is because, by Tweedie's formula [17, 32],

$$\begin{aligned} \mathbb{E}[\mathbf{x}|\tilde{\mathbf{x}}, \mathbf{y}] &= \tilde{\mathbf{x}} + \sigma_t^2 \nabla_{\tilde{\mathbf{x}}} \log p(\tilde{\mathbf{x}}|\mathbf{y}) \\ &\approx \tilde{\mathbf{x}} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, \mathbf{y}, t). \end{aligned} \tag{6}$$

As a result, we can effectively sample from the "inverse distribution" $p(\tilde{\mathbf{x}}|\mathbf{y})$ via $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, \mathbf{y}, t)$ using Langevin dynamics [1, 68] without having access to the gradients of the model f or any other model-specific loss terms.

Intuitively, during training, especially in early denoising steps, it is difficult for the DM to both denoise an image to get a realistic face and match the specific training image. The ID vector contains information (*e.g.* face shape) that the DM is incentivized to use ($\rightarrow$ lower loss) to get closer to the training image. During inference, the random seed determines identity-agnostic features ($\rightarrow$ many results), and the ID conditioning pushes the DM to generate an image that resembles the target identity.

4. Method

Motivated by the results from Sec. 3, we adopt a conditional diffusion model (DM) for the task of inverting a face recognition (FR) model. Since conditional DMs have inherent advantages for one-to-many and inversion tasks, this results in a minimal problem formulation compared to task-specific methods that require complicated supervision [16] or regularization [74] signals. The overall architecture of our method is visualized in Fig. 2.

4.1. Diffusion model formulation

We build up on the diffusion model proposed by Dhariwal and Nichol [15]. Given a sample $\mathbf{x}_0$ from the image dis-

tribution, a sequence $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T$ of noisy images is produced by progressively adding Gaussian noise according to a variance schedule. At the final time step, $\mathbf{x}_T$ is assumed to be pure Gaussian noise: $\mathcal{N}(0, \mathbf{I})$. A neural network is then trained to reverse this diffusion process in order to predict $\mathbf{x}_{t-1}$ from the noisy image $\mathbf{x}_t$ and the time step t. To sample a new image, we sample $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoise it, producing a sequence $\mathbf{x}_T, \mathbf{x}_{T-1}, \ldots, \mathbf{x}_1, \mathbf{x}_0$. The final image, $\mathbf{x}_0$, should resemble the training data.

As [15], we assume that we can model $p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t)$ as a Gaussian $\mathcal{N}(\mathbf{x}_{t-1}; \mu_{\boldsymbol{\theta}}(\mathbf{x}_t, t), \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathbf{x}_t, t))$ whose mean $\mu_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ can be calculated as a function of $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$, the (unscaled) noise component of $\mathbf{x}_t$. We extend this by conditioning on the ID vector $\mathbf{y}$ and thus predict $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{y}, t)$. Extending [46] to the conditional case, we predict the noise $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{y}, t)$ and the variance $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{y}, t)$ from the image $\mathbf{x}_t$, the ID vector $\mathbf{y}$, and the time step t, using the objective

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t, \mathbf{x}_0, \mathbf{y}, \boldsymbol{\epsilon}}[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{y}, t)||^2]. \tag{7}$$

For more details, refer to the diffusion model works [23, 46, 15]. Note that this objective is identical to the one theoretically derived in (5). While some recent work has considered the application of diffusion models to inverse problems, they typically assume $p(\mathbf{y}|\mathbf{x})$ is known [26, 30, 62, 6, 7, 20, 5, 8, 3, 41, 63], while we make no such assumption.

Following Ramesh *et al*. [51], we adapt classifier-free guidance [24] by setting the ID vector to the $\mathbf{0}$-vector with 10% probability during training (unconditional setting). During inference, we sample from both settings, and the model prediction $\hat{\boldsymbol{\epsilon}}_\theta$ becomes

$$\hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{y}, t) = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{0}, t) + s[\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{y}, t) - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{0}, t)], \tag{8}$$

where $s \geq 1$ is the guidance scale. Higher guidance scales cause the generation process to consider the identity conditioning more.

4.2. Architecture

The model is a U-net [54] that takes the image $\mathbf{x}_t$, the ID vector $\mathbf{y}$, and the time step t as input. The U-net architecture is adapted from [15] and is described in detail in the supplementary material. To condition the diffusion model on the identity, we add an identity embedding to the residual connections of the ResNet blocks, as commonly done for class embeddings [15] and the CLIP [50] embedding in text-to-image generation methods [51, 56]. The identity embedding is obtained by projecting the ID vector through a learnable fully connected layer such that it has the same size as the time step embedding and can be added to it.

4.3. Controllability

Due to its robustness and ability to pick a mode by setting the random seed during image generation, our method permits smooth interpolations and analyses in the ID vector latent space unlike other works that invert FR models. For example, we can smoothly interpolate between different identities as visualized in Fig. 1. Furthermore, we can find meaningful directions in the latent spaces. Since the directions extracted automatically using principal component analysis (PCA) are generally difficult to interpret beyond the first dimension (see supplementary material), we calculate custom directions using publicly available metadata [11] for the FFHQ data set. For binary features (*e.g*. glasses), we define the custom direction vector as the difference between the mean ID vectors of the two groups. For continuous features (*e.g*. age), we map to the binary case by considering ID vectors with feature values below the 10th percentile and values above the 90th percentile for the two groups respectively. Examples of traveling along meaningful ID vector directions can be seen in Fig. 1.

To better disentangle identity-specific and identity-agnostic information and obtain additional interpretable control, we can optionally extend our method by also conditioning the DM on an attribute vector as done for the ID vector. In practice, we recommend using only identity-agnostic attributes (referred to as set 1) along with the identity. In the supplementary material, we also show attribute sets that overlap more with identity (sets 2 & 3) for completeness.

4.4. Implementation details

As data set, we use FFHQ [28] and split it into 69000 images for training and 1000 for testing. As we can only show images of individuals with written consent (see Sec. 7), we use a proprietary data set of faces for the qualitative results in this paper. To condition our model, we use ID vectors from a PyTorch FaceNet implementation [59, 18] or the default InsightFace method [13]. To evaluate the generated images and thereby match the verification accuracy on real images shown in Vec2Face [16] as closely as possible, we use the official PyTorch ArcFace implementation [14, 12] and a TensorFlow FaceNet implementation [59, 58]. A detailed description of the remaining implementation details and ID vectors is in the supplementary material.

5. Experiments and results

5.1. Comparison to state-of-the-art methods

We mainly compare our model with the three methods that generate faces from black-box features whose code is available online: NbNet [40] ("vgg-percept-nbnetb" parameters), Gaussian sampling [52], and StyleGAN search [65].

Figure 3 compares the outputs of our method with those of current state-of-the-art methods. While capturing the identity of the input face well in some cases, NbNet [40] and Gaussian sampling [52] both fail to produce realistic faces. In contrast, StyleGAN search [65] always generates

high-quality images, but they are not always faithful to the original identity, sometimes failing completely as seen in the last row. Our method is the only method that produces high-quality, realistic images that consistently resemble the original identity. Our observations are supported by the user study in the supplementary material.

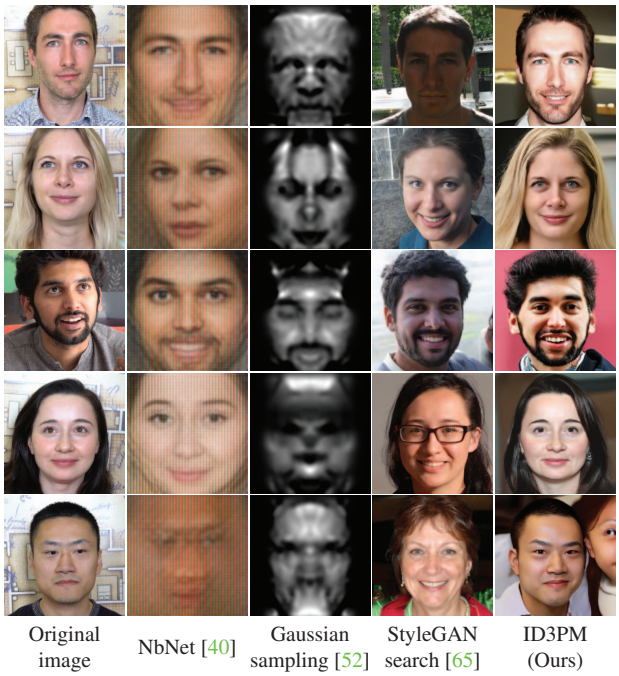

Figure 3: Qualitative evaluation with state-of-the-art methods. The generated images of our method (with InsightFace [13] ID vectors) look realistic and resemble the identity of the original image more closely than other methods. Note that the second-best performing method, StyleGAN search [65], often fails completely as seen in the last row.

For the quantitative evaluation of the identity preservation, we generate one image from each ID vector of all 1000 images of the FFHQ [28] test set for each method. We then calculate the distances according to the ArcFace [14, 12] and FaceNet [59, 58] face recognition methods for the 1000 respective pairs. The resulting distance distributions are plotted in Fig. 4. Note that StyleGAN search [65] optimizes the FaceNet distance during the image generation and thus performs well when evaluated with FaceNet but poorly when evaluated with ArcFace. The opposite effect can be seen for Gaussian sampling, which optimizes ArcFace during image generation. Despite not optimizing the ID vector distance directly (neither during training nor inference), our method outperforms all other methods, producing images that are closer to the original images' identities.

To further evaluate the identity preservation and to compare to Vec2Face [16] despite their code not being available online, we follow the procedure used in Vec2Face [16].

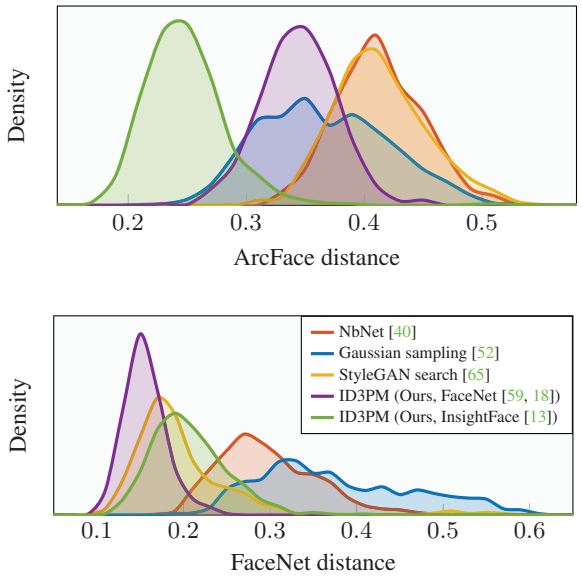

Figure 4: Probability density functions of the ArcFace [14, 12] and FaceNet [59, 58] distances (lower is better) of 1000 FFHQ [28] test images and their respective reconstructions.

Specifically, we use the official validation protocols of the LFW [25], AgeDB-30 [45], and CFP-FP [60] data sets and replace the first image in each positive pair with the image reconstructed from its ID vector, while keeping the second image as the real reference face. The face matching accuracies for ArcFace [14, 12] and FaceNet [59, 58] are reported in Tab. 2. Our method outperforms NbNet [40], Gaussian sampling [52], and StyleGAN search [65] in almost all tested configurations and performs on-par with or better than Vec2Face [16]. Note that our method has fewer requirements for the training data set (70000 images vs. 490000 images grouped into 10000 classes) and produces visually superior results compared to Vec2Face [16], as confirmed in the user study in the supplementary material.

To evaluate the diversity of the generated results, we generate 100 images for the first 50 identities of the FFHQ [28] test set. Motivated by the diversity evaluation common in unpaired image-to-image translation literature [4, 35], we calculate the mean pairwise LPIPS [73] distances among all images of the same identity. We further calculate the mean pairwise pose and expressions extracted using 3DDFA_V2 [21]. We additionally calculate the mean identity vector distances according to ArcFace [14, 12] and FaceNet [59, 58] to measure the identity preservation. We report these values in Tab. 3.

Since NbNet [40] is a one-to-one method and Gaussian sampling [52] produces faces that often fail to be detected by 3DDFA_V2 [21], we only compare with StyleGAN search [65]. In our default configuration (marked with * in Tab. 3), we obtain similar diversity scores as Style-

Method	LFW		AgeDB-30		CFP-FP	
	ArcFace ↑	FaceNet ↑	ArcFace ↑	FaceNet ↑	ArcFace ↑	FaceNet ↑
Real images	99.83%	99.65%	98.23%	91.33%	98.86%	96.43%
NbNet [40]	87.32%	92.48%	81.83%	82.25%	87.36%	89.89%
Gaussian sampling [52]	89.10%	75.07%	80.43%	63.42%	61.39%	55.26%
StyleGAN search [65]	82.43%	95.45%	72.70%	85.22%	80.83%	92.54%
Vec2Face [16] [1]	99.13%	98.05%	93.53%	**89.80%**	89.03%	87.19%
ID3PM (Ours, FaceNet [59, 18])	97.65%	**98.98%**	88.22%	88.00%	94.47%	**95.23%**
ID3PM (Ours, InsightFace [13])	**99.20%**	96.02%	**94.53%**	79.15%	**96.13%**	87.43%

Table 2: Quantitative evaluation of the identity preservation with state-of-the-art methods. The scores depict the matching accuracy when replacing one image of each positive pair with the image generated from its ID vector for the protocols of the LFW [25], AgeDB-30 [45], and CFP-FP [60] data sets. The best performing method per column is marked in bold. [1] Values taken from their paper.

GAN search [65], while preserving the identity much better. Note that the diversity scores are slightly skewed in favor of methods whose generated images do not match the identity closely since higher variations in the identity also lead to more diversity in the LPIPS [73] features.

5.2. Controllability

5.2.1 Guidance scale

The classifier-free guidance scale offers control over the trade-off between the fidelity and diversity of the generated results. As seen in Fig. 5, by increasing the guidance, the generated faces converge to the same identity, resemble the original face more closely, and contain fewer artifacts. At the same time, higher guidance values reduce the diversity of identity-agnostic features such as the background and expressions and also increase contrast and saturation.

Figure 5: Effect of the guidance scale on the generated images. For the (InsightFace [13]) ID vector extracted from the image on the left, we generate images for four seeds at guidance scales s ranging from 1.0 to 5.0.

To measure this effect quantitatively, we perform the same evaluation as in the previous section and report the results in Tab. 3. As the guidance scale increases, the identity preservation improves as indicated by the decreasing identity distances, but the diversity in terms of poses, expressions, and LPIPS [73] features decreases. In practice, we choose a guidance scale of 2.0 for all experiments unless stated otherwise because that appears to be the best compromise between image quality and diversity. In the supplementary material, we further show FID [22] as well as precision and recall [34] values that measure how well the image distribution is preserved as the guidance scale varies.

5.2.2 Identity vector latent space

As described in Sec. 4, we can find custom directions in the ID vector latent space that enable us to smoothly interpolate identities as well as change features such as the age or hair color as seen in Fig. 1 and in the supplementary material. Note that we refer to these features as *identity-specific* because they exist in the ID vector latent space. In theory, this space should not contain any identity-agnostic information such as the pose. In practice, however, some FR methods inadvertently do extract this information. This is shown in great detail in the supplementary material, where we show an interesting application of our method to analyze pre-trained face recognition methods.

5.2.3 Attribute conditioning

By additionally conditioning our method on attributes, we can disentangle identity-specific and identity-agnostic features. As seen in Fig. 6, the additional attribute conditioning allows us to recover more of the original data distribution in terms of head poses and expressions whereas a model conditioned only on the ID vector is more likely to overfit and learn biases from the training data set. This is also shown in Tab. 3, where the diversity increases with attribute conditioning at the expense of worse identity preservation com-

Method	Setting	Diversity			Identity distance	
		Pose ↑	Expression ↑	LPIPS ↑	ArcFace ↓	FaceNet ↓
StyleGAN search [65]	-	12.57	**1.57**	**0.317**	0.417	0.215
ID3PM (Ours)	Guidance scale = 1.0	**17.36**	1.35	0.315	0.291	0.234
	1.5	16.69	1.18	0.301	0.260	0.211
	2.0 *	16.24	1.10	0.290	0.247	0.203
	2.5	15.88	1.05	0.282	0.242	0.201
	3.0	15.55	1.01	0.274	**0.239**	**0.200**
ID3PM (Ours)	Attribute conditioning	16.93	1.45	0.306	0.302	0.252

Table 3: Quantitative evaluation of the diversity and identity distances of 100 generated images for 50 identities with StyleGAN search [65], different guidance scales, and attribute conditioning (set 1). InsightFace [13] ID vectors are used for our methods in this experiment. The best performing method per column is marked in bold. * Indicates the default setting used in this paper and also for the run with attribute conditioning.

pared to the base configuration. The attribute conditioning also enables intuitive control over the generated images by simply selecting the desired attribute values as shown in Fig. 1 and in the supplementary material.

Figure 6: Attribute conditioning diversity. Through additional attribute conditioning, we can disentangle identity-specific and identity-agnostic features. As a result, we obtain more diverse results when using both (InsightFace [13]) ID vector and attribute vector conditioning (set 1) compared to when only using ID vector conditioning.

6. Limitations

Our method outputs images at a relatively low resolution of 64×64. While this can be upsampled using super-resolution models, some fine identity-specific details such as moles cannot be modeled currently (but this information might not even be stored in the ID vector). Our method also has relatively long inference times (15 seconds per image when using batches of 16 images on one NVIDIA RTX 3090 GPU) in the default setting, but this can be reduced to less than one second per image when using 10 respacing steps at a slight decrease in quality as shown in the supplementary material. Our method also occasionally has small image generation artifacts, but the above aspects are expected to improve with future advancements in diffusion models. Lastly, our model inherits the biases of both the face recognition model and the training data set. This can manifest as either accessorizing images corresponding to certain demographic factors (*e.g.* via make-up, clothing) or losing identity fidelity for underrepresented groups as shown in the supplementary material. This suggests an additional application of our work to the study of systematic biases in otherwise black-box systems.

7. Ethical concerns

All individuals portrayed in this paper provided informed consent to use their images as test images. This was not possible for the images from the FFHQ [28], LFW [25], AgeDB-30 [45], and CFP-FP [60] data sets. Therefore, we do not show them in the paper and cannot provide qualitative comparisons to Vec2Face [16] (code not available).

We recognize the potential for misuse of any method that creates realistic imagery of human beings, especially when the images are made to correspond to specific individuals. We condemn such misuse and support ongoing research into the identification of artificially manipulated data.

8. Conclusion

We propose a method to generate high-quality identity-preserving face images by injecting black-box, low-dimensional embeddings of a face into the residual blocks of a diffusion model. We mathematically reason and empirically show that our method produces images close to the target identity despite the absence of any identity-specific loss terms. Our method obtains state-of-the-art performance on identity preservation and output diversity, as demonstrated qualitatively and quantitatively. We further showcase advantages of our approach in providing control over the generation process. We thus provide a useful tool to create data sets with user-defined variations in identities and attributes as well as to analyze the latent spaces of face recognition methods, motivating more research in this direction.

References

[1] Giovanni Bussi and Michele Parrinello. Accurate sampling using langevin dynamics. *Physical Review E*, 75(5):056707, 2007. 4

[2] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2003–2011, 2020. 1

[3] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 3, 5

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 6

[5] Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. Parallel diffusion models of operator and image for blind inverse problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6059–6069, 2023. 3, 5

[6] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 3, 5

[7] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022. 3, 5

[8] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12413–12422, 2022. 3, 5

[9] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. Synthesizing normalized faces from facial identity features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3703–3712, 2017. 1, 2, 3

[10] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7):1967–1974, 2018. 3

[11] DCGM. Gender, age, and emotions extracted for flickr-faces-hq dataset (ffhq), 2020. 5

[12] Jinakang Deng, Jia Guo, Xiang An, Jack Yu, and Baris Gecer. Distributed arcface training in pytorch, 2021. 5, 6

[13] Jinakang Deng, Jia Guo, Xiang An, Jack Yu, and Baris Gecer. Insightface: 2d and 3d face analysis project, 2022. 1, 5, 6, 7, 8

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2, 5, 6

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3, 4, 5

[16] Chi Nhan Duong, Thanh-Dat Truong, Khoa Luu, Kha Gia Quach, Hung Bui, and Kaushik Roy. Vec2face: Unveil human faces from their blackbox features in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6132–6141, 2020. 2, 3, 4, 5, 6, 7, 8

[17] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 4

[18] Tim Esler. Face recognition using pytorch, 2021. 5, 6, 7

[19] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 3

[20] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 35:14715–14728, 2022. 3, 5

[21] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX*, pages 152–168. Springer, 2020. 6

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 4, 5

[24] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[25] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 6, 7, 8

[26] Zahra Kadkhodaie and Eero Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. *Advances in Neural Information Processing Systems*, 34:13242–13254, 2021. 3, 5

[27] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 4

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 5, 6, 8

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[30] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 3, 5

[31] Jiseob Kim, Jihoon Lee, and Byoung-Tak Zhang. Smooth-swap: a simple enhancement for face-swapping with smoothness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10779–10788, 2022. 1

[32] Kwanyoung Kim and Jong Chul Ye. Noise2score: tweedie's approach to self-supervised image denoising without clean images. *Advances in Neural Information Processing Systems*, 34:864–874, 2021. 4

[33] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18750–18759, 2022. 2

[34] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 7

[35] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128:2402–2417, 2020. 6

[36] Chenyu Li, Shiming Ge, Daichi Zhang, and Jia Li. Look through masks: Towards masked face recognition with de-occlusion distillation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3016–3024, 2020. 2

[37] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 2

[38] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015. 2

[39] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015. 3

[40] Guangcan Mai, Kai Cao, Pong C Yuen, and Anil K Jain. Face image reconstruction from deep templates. *arXiv preprint arXiv:1703.00832*, 2017. 2, 3, 5, 6, 7

[41] Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023. 3, 5

[42] Alexis Mignon and Frédéric Jurie. Reconstructing faces from their signatures using rbf regression. In *British Machine Vision Conference 2013*, pages 103–1, 2013. 2

[43] Pranab Mohanty, Sudeep Sarkar, and Rangachar Kasturi. From scores to face templates: A model-based approach. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2065–2078, 2007. 2

[44] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015. 2

[45] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017. 6, 7, 8

[46] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 4, 5

[47] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. *arXiv preprint arXiv:2005.07728*, 2020. 1, 2, 3

[48] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Association*, 2015. 2

[49] Haibo Qiu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. End2end occluded face recognition by masking corrupted features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3, 5

[52] Anton Razzhigaev, Klim Kireev, Edgar Kaziakhmedov, Nurislam Tursynbek, and Aleksandr Petiushko. Black-box face recovery from identity features. In *European Conference on Computer Vision*, pages 462–475. Springer, 2020. 2, 3, 5, 6, 7

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[55] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[56] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3, 5

[57] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022. 2

[58] David Sandberg. Face recognition using tensorflow, 2018. 5, 6

[59] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2, 5, 6, 7

[60] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016. 6, 7, 8

[61] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[62] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2022. 3, 5

[63] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3, 4, 5

[64] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 2

[65] Edward Vendrow and Joshua Vendrow. Realistic face reconstruction from deep embeddings. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021. 2, 3, 5, 6, 7, 8

[66] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 2

[67] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 2

[68] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011. 4

[69] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[70] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pages 225–240, New York, NY, USA, 2019. ACM. 2, 3

[71] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 2

[72] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015. 2

[73] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6, 7

[74] Andrey Zhmoginov and Mark Sandler. Inverting face embeddings with convolutional neural networks. *arXiv preprint arXiv:1606.04189*, 2016. 1, 2, 3, 4