# Vec2Face: Unveil Human Faces from their Blackbox Features in Face Recognition

Chi Nhan Duong[1], Thanh-Dat Truong[2], Khoa Luu[2], Kha Gia Quach[1], Hung Bui[3], Kaushik Roy[4]
[1] Concordia University, Canada [2] University of Arkansas, USA [3] VinAI Research
[4] North Carolina A&T State University, USA
[1]{dcnhan, kquach}@ieee.org, [2]{tt032, khoaluu}@uark.edu, [3]v.hungbh1@vinai.io, [4]kroy@ncat.edu

## Abstract

*Unveiling face images of a subject given his/her high-level representations extracted from a blackbox Face Recognition engine is extremely challenging. It is because the limitations of accessible information from that engine including its structure and uninterpretable extracted features. This paper presents a novel generative structure with Bijective Metric Learning, namely Bijective Generative Adversarial Networks in a Distillation framework (DiBiGAN), for synthesizing faces of an identity given that person's features. In order to effectively address this problem, this work firstly introduces a bijective metric so that the distance measurement and metric learning process can be directly adopted in image domain for an image reconstruction task. Secondly, a distillation process is introduced to maximize the information exploited from the blackbox face recognition engine. Then a Feature-Conditional Generator Structure with Exponential Weighting Strategy is presented for a more robust generator that can synthesize realistic faces with ID preservation. Results on several benchmarking datasets including CelebA, LFW, AgeDB, CFP-FP against matching engines have demonstrated the effectiveness of DiBiGAN on both image realism and ID preservation properties.*

## 1. Introduction

Face recognition has recently matured and achieved high accuracy against millions of identities [47, 48]. A face recognition system is often designed in two main stages, i.e. feature extraction and feature comparison. The role of feature extraction is more important since it directly determines the robustness of the engine. This operator defines an embedding process mapping input facial images into a higher-level latent space where embedded features extracted from photos of the same subject distribute within a small margin [32]. Moreover, since most face recognition engines are set into a blackbox mode to protect the technologies [25], there is no apparent technique to inverse that embedding process to reconstruct the faces of a subject given his/her extracted features from those engines.

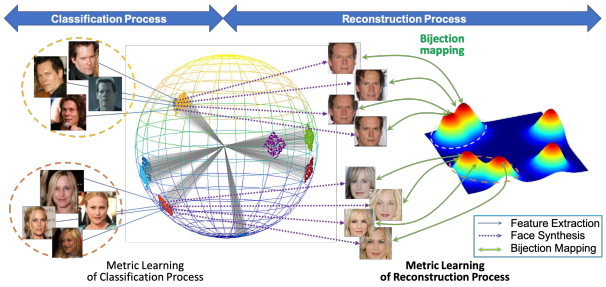


Figure 1. **Metric Learning for Image Reconstruction.** By maintaining the one-to-one mapping via a bijection, the distance between images can be directly and intuitively measured and enhances the metric learning process for image reconstruction task.

Some Blackbox Adversarial Attack approaches [20, 21, 44] have partially addressed this task by analyzing the gradients of the classifier's outputs to generate adversarial examples that mislead the behaviour of that classifier. However, they only focus on a *closed-set* problem where the output classes are predefined. Moreover, their goal is to generate imperceptible pertubations added to the given input signal. Other methods [1, 4, 7, 43, 53] are also introduced in literature but still require the access to the classifier structure, i.e. whitebox setting. Meanwhile, our goal focuses on a more challenging reconstruction task with a *blackbox* face recognition. Firstly, this process *reconstructs faces from scratch* without any hint from input images. Secondly, in a blackbox setting, there is *no information about the engine's structure*, and, therefore, it is unable to directly exploit knowledge from the inverse mapping process (i.e. back-propagation). Thirdly, the embedded features from a face recognition engine are for *open-set problem* where no label information is available. More importantly, the subjects to be reconstructed may have never been seen during training process of the face recognition engine. In the *scope of this work*, we assume that the face recognition engines are primarily developed by Convolutional Neural Networks

Table 1. Comparisons of our DibiGAN and other unrestricted synthesis methods. Image Reconstruction (Img_Recon), Feature Representation (Feat), Guided Image ($\text{Img}_G$), Feature Conditional (Feat_Cond), Neighborly Deconvolution (NB_Deconv), Optimization (Opt).

| | **Ours** | NBNet [34] | SynNormFace [4] | IFaceRec [53] | INVREP [33] |
|---|---|---|---|---|---|
| **Input** | Feat | Feat | Feat | Feat + $\text{Img}_G$ | Feat |
| **Generator Structure** | **Feat_Cond** | NB_Deconv | MLP + CNN | DeConvNet | Opt |
| **Blackbox Support** | ✓ | ✓ | ✗ | ✗ | ✗ |
| **Img_Recon Metric** | **Bijective** | ✗ | ✗ | ✗ | ✗ |
| **Exploited Knowledge from Classifier** | **Fully (Distillation)** | Partially | Fully (Whitebox) | Fully (Whitebox) | Fully (Whitebox) |

(CNN) that dominate recent state-of-the-art results in face recognition [5, 10, 12, 28, 29, 40, 45, 46, 52]. We also assume that there is no further post-processing after the step of CNN feature extraction. We then develop a theory to guarantee the reconstruction robustness of the proposed method.

**Contributions.** This paper presents a novel generative structure, namely Bijective Generative Adversarial Networks in a Distillation framework (DiBiGAN), with Bijective Metric Learning for the image reconstruction task. The contributions of this work are four-fold. (1) Although many metric learning techniques have been introduced in the literature, they are mainly adopted for classification rather than *reconstruction* process. By addressing limitations of classifier-based metrics for image reconstruction, we propose a novel ***Bijective Metric Learning*** with bijection (one-to-one mapping) property so that the distances in latent features are equivalent to those between images (see Fig. 1). It, therefore, provides a more effective and natural metric learning approach to the image reconstruction task. (2) We exploit different aspects of the *distillation process* for the image reconstruction task in a blackbox mode. They include *distilled knowledge* from the blackbox face matcher and *ID knowledge extracted* from a real face structure. (3) We introduce a Feature-Conditional Generator Structure with Exponential Weighting Strategy for Generative Adversarial Network (GAN)-based framework to learn a more robust generator to synthesize realistic faces with ID preservation. (4) Evaluations on benchmarks against various face recognition engines have illustrated the improvements of DiBiGAN in both image realism and ID preservation. To the best of our knowledge, this is one of the first metric learning methods for image reconstruction (Table 1).

## 2. Related Work

Synthesizing images [4, 7, 8, 9, 34, 53] has brought several interests from the community. We divided into two groups, i.e. unrestricted and adversarial synthesis.

**Unrestricted synthesis.** The approaches focus on reconstructing an image from scratch given its high-level representation. Since the mapping is from a low-dimensional latent space to a highly nonlinear image space, several regularizations have to be applied, e.g. Gaussian Blur [51] for high-frequency samples or Total Variation [11, 33] for maintaining piece-wise constant patches. These optimization-based techniques are limited with high computation or unrealistic reconstructions. Later, Dosovitskiy et al. [7] proposed to reconstruct the image from its shallow (i.e. HOG, SIFT) and deep features using a Neural Network (NN). Zhmoginov et al. [53] presented an iterative method to invert Facenet [40] feature with feed-forward NN. Cole et al. [4] proposed an autoencoder structure to map the features to frontal neutral face of the subject. Yang et al. [49] adopted autoencoder for model inversion task. Generally, to produce better synthesized quality, these approaches require full access to the deep structure to exploit the gradient information from the embedding process. Mai et al. [34] developed a neighborly deconvolutional network to support the blackbox mode. However, with only pixel and perceptual [23] losses, there are limitations of ID preservation when synthesizing different features of the same subject. In this work, we address this issue with Bijective Metric Learning and Distillation Knowledge for reconstruction task.

**Adversarial synthesis.** Adversarial approaches aim at generating unnoticable perturbations from input images for adversarial examples to mislead the behaviour of a deep structure. Either directly accessing or indirectly approximating gradients, adversarial examples are created by maximizing corresponding loss which can fool a classifier [1, 2, 3, 20, 21, 30, 35, 42, 44]. Ilyas et. al. [21] proposed bandit optimization to exploit prior information about the gradient of deep learning models. Later, Ilyas et. al. [20] introduced Natural Evolutionary Strategies to enable query-efficient generation of black-box adversarial examples. Other knowledge from the blackbox classifier are also exploited for this task [44, 1, 43]. Generally, although the approaches in this direction tried to extract the gradient information from a blackbox classifier, their goal are mainly to mislead the behaviours of the classifier with respect to a pre-defined set of classes. Therefore, they are closed-set approaches. Meanwhile, in our work, the proposed framework can reconstruct faces of subjects that have not been seen in the training process of the classifier.

## 3. Our Proposed Method

Let $F : \mathcal{I} \mapsto \mathcal{F}$ be a function that maps an input image $I$ from image domain $\mathcal{I} \in \mathbb{R}^{W \times H \times C}$ to its high-level embedding feature $F(I)$ in latent domain $\mathcal{F} \in \mathbb{R}^M$. In addition, a function $C : \mathcal{F} \mapsto \mathcal{Y}$ takes $\mathcal{F}(I)$ as its input and gives the identity (ID) prediction of the subject in space $\mathcal{Y} \in \mathbb{R}^N$ where each dimension represents a predefined subject class.

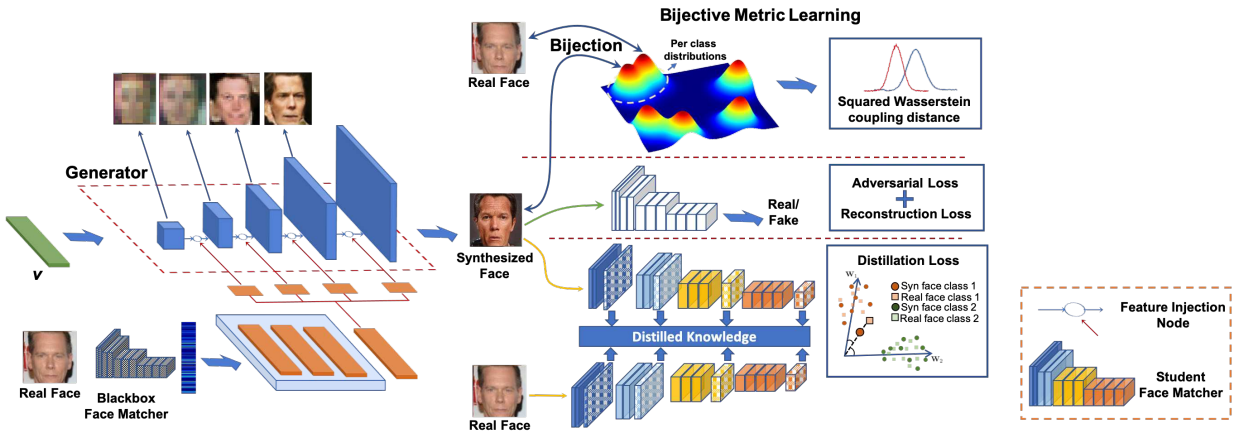


Figure 2. **Proposed Framework.** Given a high-level embedding representation, a *Feature-Conditional Generator* injects that representation through-out its structure as the conditional information for all scales. The cost functions are designed with *Bijective Metric* to directly exploit ID distributions in image domain, and *Distillation Loss* to maximize the knowledge could be extracted from the blackbox matcher.

**Definition 1** (Model Inversion). *Given blackbox functions $F$ and $C$; and a prediction score vector $s = [F \circ C](I)$ extracted from an unknown image $I$, the goal of model inversion is to recover $I$ from $s$ such that $\tilde{I}^* = \arg\min_{\tilde{I}} \mathcal{L}([F \circ C](\tilde{I}), s)$ where $\mathcal{L}$ denotes some types of distance metrics.*

The approaches solving this problem usually exploit the relationship between an input image and its class label for the reconstruction process. Moreover, since the output score $s$ is fixed according to predefined $N$ classes, the reconstruction is limited on images of training subject IDs.

**Definition 2** (Feature Reconstruction). *Given a blackbox functions $F$; and its embedding feature $f = F(I)$ of an unknown image $I$, feature reconstruction is to recover $I$ from $f$ by optimizing $\tilde{I}^* = \arg\min_{\tilde{I}} \mathcal{L}(F(\tilde{I}), f)$.*

Compared to the model inversion problem, Feature Reconstruction is more challenging since the constraints on predefined classes are removed. Therefore, the solution for this problem turns into an ***open-set mode*** where it can reconstruct faces other than the ones used for learning $F$, i.e. face recognition engine. Moreover, since the parameters of $F$ are inaccessible due to its blackbox setting, directly recovering $I$ based on its gradient is impossible. Therefore, the feature reconstruction task can be reformulated via a function (generator) $G : \mathcal{F} \mapsto \mathcal{I}$ as the reverse mapping of $F$.

$$\begin{aligned} \tilde{I} &= G(f; \theta_g) \\ \theta_g &= \arg\min_{\theta} \mathbb{E}_{\mathbf{x}\sim p_I} \left[\mathcal{L}_G^x \left([G \circ F](\mathbf{x}; \theta), \mathbf{x}\right)\right] \\ &= \arg\min_{\theta} \int \mathcal{L}_G^x \left(\tilde{\mathbf{x}}, \mathbf{x}\right) p_I(\mathbf{x}) d\mathbf{x} \end{aligned} \tag{1}$$

where $\tilde{\mathbf{x}} = [G \circ F](\mathbf{x}; \theta)$, $\theta_g$ denotes the parameters of $G$, and $p_I(\mathbf{x})$ is the probability density function of $\mathbf{x}$. In other words, $p_I(\mathbf{x})$ indicates the distribution that image $I$ belonged to (i.e. *the distribution of training data of $F$*). Intuitively, function $G$ can be seen as a function that maps images from embedding space $\mathcal{F}$ back to image space such that all reconstructed images $[G \circ F](\mathbf{x}; \theta_g)$ are maintained to be close to its real $\mathbf{x}$ with respect to the distance metric $\mathcal{L}_G^x$. To produce "good quality" synthesis (i.e. *realistic images with ID preservation*), different choices for $\mathcal{L}_G^x$ have been commonly exploited [22, 23, 34] such as pixel difference in image domain via $L_1/L_2$ distance; Probability Distribution Divergence (i.e. Adversarial loss defined via an additional Discriminator) for realistic penalization; or Perceptual distance that penalizes the image structure in high-level latent space. Among these metrics, except the pixel difference that is computed directly in image domain, the others are indirect metrics where another mapping function (i.e. classifier) from image space to latent space is required.

### 3.1. Limitations of Classifier-based Metrics

Although these indirect metrics have shown their advantages in several tasks, there are limitations when only the blackbox function $F$ and its embedded features are given.

***Limitation 1.*** As shown in several adversarial attack works [15, 39], since the function $F$ is not a one-to-one mapping function from $\mathcal{I}$ to $\mathcal{F}$, it is straightforward to find two images of similar latent representation that are drastically different in image content. Therefore, with no prior knowledge about the subject ID of image $I$, starting to reconstruct it from scratch may easily fall into the case where the reconstructed image $\tilde{I}$ is totally different to $I$ but has similar embedding features. The current Probability Distribution Divergence with Adversarial Loss or Perceptual Distance is limited in maintaining the constrain "*the reconstructions of features of the same subject ID should be similar*".

***Limitation 2.*** Since the access to the structure and intermediate features of $F$ is unavailable in the blackbox mode, the function $G$ is unable to directly exploit valuable information from the gradient of $F$ and the intermediate representation during embedding. As a result, the distance metrics defined via $F$, i.e. perceptual distance, is less effective as in white-box setting. Next sections will introduce two loss functions to tackle these problems to learn a robust function $G$.

### 3.2. Bijective Metrics for Image Reconstruction

Many metric learning proposals for face recognition [5, 28, 29, 45, 46, 52] have been used to improve both intra-class compactness and inter-class separability with a large margin. However, for feature reconstruction, directly adopting these metrics, e.g. angular distance for reconstructed images to cluster images of the same ID is infeasible.

Therefore, we propose a bijection metric for feature reconstruction task such that the mapping function from image to latent space is one-to-one. The distance between their latent features is equivalent to the distance between images. By this way, these metrics are more aligned to image domain and can be effectively adopted for reconstruction task. Moreover, since two different images cannot be mapped to the same latent features, the metric learning process is more reliable. The optimization of $G$ in Eqn. (1) is rewritten as:

$$\theta_g \approx \arg\min_{\theta} \int \mathcal{L}_G^x(\tilde{\mathbf{x}}, \mathbf{x}) p_x(\mathbf{x}) d\mathbf{x} \tag{2}$$

where $\tilde{\mathbf{x}} = [G \circ F](\mathbf{x}; \theta)$; and $p_x(\mathbf{x})$ denotes a density function estimated from an alternative large-scale face dataset. Notice that although the access to $p_I(\mathbf{x})$ is not available, this approximation can be practically adopted due to a prior knowledge about $p_I(\mathbf{x})$ that images drawn from $p_I(\mathbf{x})$ are facial images. Let $H : \mathcal{I} \mapsto \mathcal{Z}$ define a bijection mapping from $\mathbf{x}$ to a latent variable $\mathbf{z} = H(\mathbf{x})$. With the bijective property, the optimization in Eqn. (2) is equivalent to.

$$\begin{aligned} &\arg\min_{\theta} \int \mathcal{L}_G^z(H(\tilde{\mathbf{x}}), H(\mathbf{x})) p_x(\mathbf{x}) d\mathbf{x} \\ =&\arg\min_{\theta} \int \mathcal{L}_G^z(H(\tilde{\mathbf{x}}), H(\mathbf{x})) p_z(\mathbf{z}) |\det(\mathbf{J}_{\mathbf{x}}^\top \mathbf{J}_{\mathbf{x}})|^{1/2} d\mathbf{z} \\ =&\arg\min_{\theta} \int \mathcal{L}_G^z(\tilde{\mathbf{z}}, \mathbf{z}) p_z(\mathbf{z}) |\det(\mathbf{J}_{\mathbf{x}}^\top \mathbf{J}_{\mathbf{x}})|^{1/2} d\mathbf{z} \end{aligned} \tag{3}$$

where $\mathbf{z} = H(\tilde{\mathbf{x}})$; $p_x(\mathbf{x}) = p_z(\mathbf{z}) |\det(\mathbf{J}_{\mathbf{x}}^\top \mathbf{J}_{\mathbf{x}})|^{1/2}$ by the change of variable formula; $\mathbf{J}_{\mathbf{x}}$ is the Jacobian of $H$ with respect to $\mathbf{x}$; and $\mathcal{L}_G^z$ is the distance metric in $\mathcal{Z}$. Intuitively, Eqn. (3) indicates that instead of computing the distance $\mathcal{L}_G^x$ and estimating $p_x(\mathbf{x})$ directly in image domain, the optimization process can be equivalently accomplished via the distance $\mathcal{L}_G^z$ and density $p_z(\mathbf{z})$ in $\mathcal{Z}$ according to the bijective property of $H$.

***The prior distributions*** $p_z$**.** In general, there are various choices for the prior distribution $p_z$ and the ideal one should have two properties: (1) *simplicity in density estimation*, and (2) *easily sampling*. Motivated from these properties, we choose Gaussian distribution for $p_z$. Notice that other distribution types are still applicable in our framework.

***The distance metric*** $\mathcal{L}_G^z$**.** With the choice of $p_z$ as a Gaussian, the distance between images in $\mathcal{I}$ is equivalent to the deviation between Gaussians in latent space. Therefore, we can effectively define $\mathcal{L}_G^z$ as the squared Wasserstein coupling distance between two Gaussian distributions.

$$\begin{aligned} \mathcal{L}_G^z(\tilde{\mathbf{z}}, \mathbf{z}) =& d(\tilde{\mathbf{z}}, \mathbf{z}) = \inf \mathbb{E}(||\tilde{\mathbf{z}} - \mathbf{z}||_2^2) \\ =& ||\tilde{\mu} - \mu||_2^2 + \mathrm{Tr}(\tilde{\Sigma} + \Sigma - 2(\tilde{\Sigma}^{1/2} \Sigma \tilde{\Sigma}^{1/2})^{1/2}) \end{aligned} \tag{4}$$

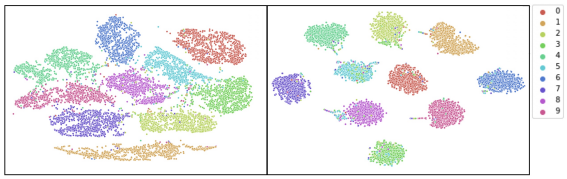


Figure 3. **The distributions of synthesized MNIST samples** on testing set (a) without, and (b) with adopting Bijective Metric.

where $\{\tilde{\mu}, \tilde{\Sigma}\}$ and $\{\mu, \Sigma\}$ are the means and covariances of $\tilde{\mathbf{z}}$ and $\mathbf{z}$, respectively. The metric $\mathcal{L}_G^z$ then can be extended with image labels to reduce the distance between images of the same ID and enhance the margin between different IDs.

$$\mathcal{L}_G^{z_{id}}(\tilde{\mathbf{z}}_\mathbf{1}, \tilde{\mathbf{z}}_\mathbf{2}) = \begin{cases} d(\tilde{\mathbf{z}}_\mathbf{1}, \tilde{\mathbf{z}}_\mathbf{2}) & \text{if } l_{\tilde{\mathbf{z}}_\mathbf{1}} = l_{\tilde{\mathbf{z}}_\mathbf{2}} \\ \max(0, m - d(\tilde{\mathbf{z}}_\mathbf{1}, \tilde{\mathbf{z}}_\mathbf{2})) & \text{if } l_{\tilde{\mathbf{z}}_\mathbf{1}} \neq l_{\tilde{\mathbf{z}}_\mathbf{2}} \end{cases} \tag{5}$$

where $m$ defines parameter controlling the margin between classes; and $\{l_{\tilde{\mathbf{z}}_\mathbf{1}}, l_{\tilde{\mathbf{z}}_\mathbf{2}}\}$ denote the subject ID of $\{\tilde{\mathbf{z}}_\mathbf{1}, \tilde{\mathbf{z}}_\mathbf{2}\}$.

***Learning the Bijection*** $H$**.** In order to effectively learn the bijection $H$, we adopt the structure of mapping function from [6, 13, 14] as the backbone for the tractable log-det computation with the log-likelihood loss for training process. Moreover, to further improve the discriminative property of $H$ in latent space $\mathcal{Z}$, we propose to exploit the ID label in training process of $H$. Particularly, given $K$ classes (i.e. ID) of the training set, we choose $K$ Gaussian distributions with different means $\{\mu_1, \mu_2, .., \mu_K\}$ and covariances $\{\Sigma_1, \Sigma_2, ..., \Sigma_K\}$ and enforce samples of each class distributed on its own prior distribution, i.e. $\mathbf{z}_k \sim \mathcal{N}(\mu_k, \Sigma_k)$. Formally, the log-likelihood loss function to learn $H$ is formulated as $\theta_H^* = \arg\max_{\theta_H} \log p_x(\mathbf{x}, k; \theta_H) = \arg\max_{\theta_H} \log p_z(\mathbf{z}, k; \theta_H) + \frac{1}{2} \log |\det(\mathbf{J}_{\mathbf{x}}^\top \mathbf{J}_{\mathbf{x}})|$.

### 3.3. Reconstruction from Distillation Knowledge

In the simplest approach, the generator $G$ can still learn to reconstruct image by adopting the Perceptual Distance as in previous works to compare $F(\tilde{I})$ and $F(I)$. However, as mentioned in Sec. 3.1, due to limited information that can be accessed from $F$, "key" information (i.e. the gradients of $F$ as well as its intermediate representations) making the perceptual loss effective is lacking. Therefore, we propose to first distill the knowledge from the blackbox $F$ to a "student" function $F^S$ and then take advantages of these knowledge via $F^S$ for training the generator. On one hand, via the distillation process, $F^S$ can mimic $F$ by aligning its feature space to that of $F$ and keeping the semantics of the extracted features for reconstruction. On the other hand, with $F^S$, the knowledge about the embedding process of $F$ (i.e. gradient, and intermediate representation) becomes more transparent; and, therefore, maximize the information which can be exploited from $F$. Particularly, let $F^S : \mathcal{I} \mapsto \mathcal{F}$ and $F^S = F_1^S \circ F_2^S \cdots \circ F_n^S$ be the composition of $n$-sub components. The knowledge from $F$ can be

distilled to $F^S$ by aligning their extracted features as.

$$\begin{aligned}\theta_S &= \arg\min_{\theta_S} \mathcal{L}_S = \mathbb{E}_{\mathbf{x}\sim p_x} d_{distill}\left(F(\mathbf{x}), F_S(\mathbf{x};\theta_S)\right) \\ &= \arg\min_{\theta_S} \mathbb{E}_{\mathbf{x}\sim p_x} \left\| 1 - \frac{F(\mathbf{x})}{\parallel F(\mathbf{x}) \parallel} * \frac{F^S(\mathbf{x};\theta_S)}{\parallel F^S(\mathbf{x};\theta_S) \parallel} \right\|_2^2 \end{aligned} \tag{6}$$

Then $G$ is enhanced via the distilled knowledge of both final embedding features and intermediate representation by.

$$\begin{aligned}\mathcal{L}_G^{distill}(\tilde{\mathbf{x}}, \mathbf{x}) = &\sum_{j=1}^{n} \lambda_j \frac{\left\| F_j^S(\tilde{\mathbf{x}};\theta_S) - F_j^S(\mathbf{x};\theta_S) \right\|}{W_j H_j C_j} \\ &+ \lambda_a \left\| 1 - \frac{F^S(\tilde{\mathbf{x}};\theta_S)}{\parallel F^S(\tilde{\mathbf{x}};\theta_S) \parallel} * \frac{F^S(\mathbf{x};\theta_S)}{\parallel F^S(\mathbf{x};\theta_S) \parallel} \right\|_2^2 \end{aligned} \tag{7}$$

where $\{\lambda_j\}_1^n$ and $\lambda_a$ denote the hyper-parameters controlling the balance between terms. The first component of $\mathcal{L}_G^{distill}(\tilde{\mathbf{x}}, \mathbf{x})$ aims to penalize the differences between the intermediate structure of the desired and reconstructed facial images while the second component validates the similarity of their final features.

### 3.4. Learning the Generator

Fig. 2 illustrates our proposed framework with Bijective Metric and Distillation Process to learn the generator $G$.

**Network Architecture.** Given an input image $\mathbf{x}$, the generator $G$ takes $F(\mathbf{x})$ as its input and aims to synthesize an image $\tilde{\mathbf{x}}$ that is as similar to $\mathbf{x}$ as possible in terms of identity and appearance. We adopt the GAN-based generator structure for $G$ and optimize using different criteria.

$$\begin{aligned}\mathcal{L}_G =& \lambda_b \mathcal{L}^{biject} + \lambda_d \mathcal{L}^{distill} + \lambda_{adv} \mathcal{L}^{adv} + \lambda_r \mathcal{L}^{recon} \\ \mathcal{L}^{biject} =& \mathbb{E}_{\mathbf{x}\sim p_x} \left[ \mathcal{L}_G^x \left( [G \circ F](\mathbf{x};\theta), \mathbf{x} \right) \right] \\ &+ \mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2\sim p_x} \left[ \mathcal{L}_G^{x_{id}} \left( [G \circ F](\mathbf{x}_1;\theta), [G \circ F](\mathbf{x}_2;\theta) \right) \right] \\ =& \mathbb{E}_{\mathbf{z}\sim p_z} \left[ \mathcal{L}_G^z \left( \tilde{\mathbf{z}}, \mathbf{z} \right) \right] + \mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2\sim p_z} \left[ \mathcal{L}_G^{z_{id}} \left( \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2 \right) \right] \\ \mathcal{L}^{distill} =& \mathbb{E}_{\mathbf{x}\sim p_x} \left[ \mathcal{L}_G^{distill} \left( [G \circ F](\mathbf{x};\theta), \mathbf{x} \right) \right] \\ \mathcal{L}^{adv} =& \mathbb{E}_{\mathbf{x}\sim p_x} \left[ D\big([G \circ F](\mathbf{x};\theta)\big) \right] \\ \mathcal{L}^{recon} =& \mathbb{E}_{\mathbf{x}\sim p_x} \left[ \| [G \circ F](\mathbf{x}) - \mathbf{x} \|_1 \right] \end{aligned} \tag{8}$$

where $\{\mathcal{L}^{biject}, \mathcal{L}^{distill}, \mathcal{L}^{adv}, \mathcal{L}^{recon}\}$ denote the bijective, distillation, adversarial, and reconstruction losses, respectively. $\{\lambda_b, \lambda_d, \lambda_{adv}, \lambda_r\}$ are their parameters controlling their relative importance. $D$ is a discriminator distinguishing the real images from a synthesized one. There are three main critical components in our framework including the Bijective $H$, the student matcher $F^S$ for ID preservation; and the discriminator $D$ for realistic penalization. The Discriminator $D$ is updated with the objective function as.

$$\begin{aligned}\mathcal{L}_D =& \mathbb{E}_{\mathbf{x}\sim p_x} \left[ D\big([G \circ F](\mathbf{x};\theta)\big) \right] - \mathbb{E}_{\mathbf{x}\sim p_x} \left[ D(\mathbf{x}) \right] \\ &+ \lambda \mathbb{E}_{\hat{\mathbf{x}}\sim p_{\hat{x}}} \left[ \left( \|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}})\|_2 - 1 \right)^2 \right] \end{aligned} \tag{9}$$

where $p_{\hat{x}}$ is the random interpolation distribution between real and generated images [16]. Then, the whole framework is trained following GAN-based minimax strategy.

**Learning Strategies.** Besides the losses, we introduce a Feature-Conditional Structure for $G$ and a exponential Weighting Strategy to adaptively scheduling the importance factors between loss terms during training process.

***Feature-conditional Structure.*** A natural design for the structure of $G$ is to directly use $F(\mathbf{x})$ as the input for $G$. However, this structure limits the learning capability of $G$. Particularly, besides ID information, $F(\mathbf{x})$ may include other "background" conditions such as poses, illuminations, expressions. Therefore, setting $F(\mathbf{x})$ as the only input implicitly enforces $G$ to "strictly" model these factors as well. This makes the training process of $G$ less effective. To relax this constraint, we introduce a Feature-Conditional structure (i.e. the generator structure in Fig. 2) where a random variable $\mathbf{v}$ is adopted as an additional input so that these background factors can be modeled through $\mathbf{v}$. Moreover, we propose to use $\mathbf{v}$ as the direct input to $G$ and inject the information from $F(\mathbf{x})$ through out the structure. By this way, $F(\mathbf{x})$ can act as the conditional ID-related information for all reconstruction scales and gives the better synthesis.

***Exponential Weighting Strategy.*** As the progressive growing training strategy [24] initializes its learning process on synthesizing low-resolution images and then gradually increasing their levels of details, it is quite effective for enhancing the details of generated images in general. However, this strategy has limited capability in preserving the subject ID. In particular, in the early stages at low scales with blurry synthesis, it is difficult to control the subject ID of faces to be synthesized while in the later stages at higher scales when the generator becomes more mature and learns to add more details, the IDs of those faces have already been constructed and become hard to be changed. Therefore, we propose to adopt a exponential weighting scheme for (1) emphasizing on ID preservation in early stages; and (2) enhancing the realism in later stages. Particularly, the parameter set $\{\lambda_b, \lambda_d, \lambda_{adv}, \lambda_r\}$ is set to $\lambda_b = \alpha e^{R_M - R(i)}$, $\lambda_d = e^{R_M - R(i)}$, $\lambda_{adv} = \beta e^{R(i)}$, $\lambda_r = e^{R_M - R(i)}$ where $R(i)$ denotes the current scales of stage $i$ and $R_M$ is the maximum scales to be learned by $G$.

## 4. Experimental Results

We qualitatively and quantitatively validate our proposed method in both reconstructed quality and ID preservation in several in-the-wild modes such as poses, expressions, and occlusions. Both image-quality and face-matching datasets are used for evaluations. Different face recognition engines are adopted to demonstrate the robustness of our model.

**Data Setting.** Our training data includes the publicly available Casia-WebFace [50] with 490K labeled facial images of over 10K subjects. The duplicated subjects between training and testing sets are removed to ensure no overlapping between them. For validation, as commonly used for attribute learning and image quality evaluation, we adopt the testing split of 10K images from CelebA [31] to validate the reconstruction quality. For ID preservation, we explore

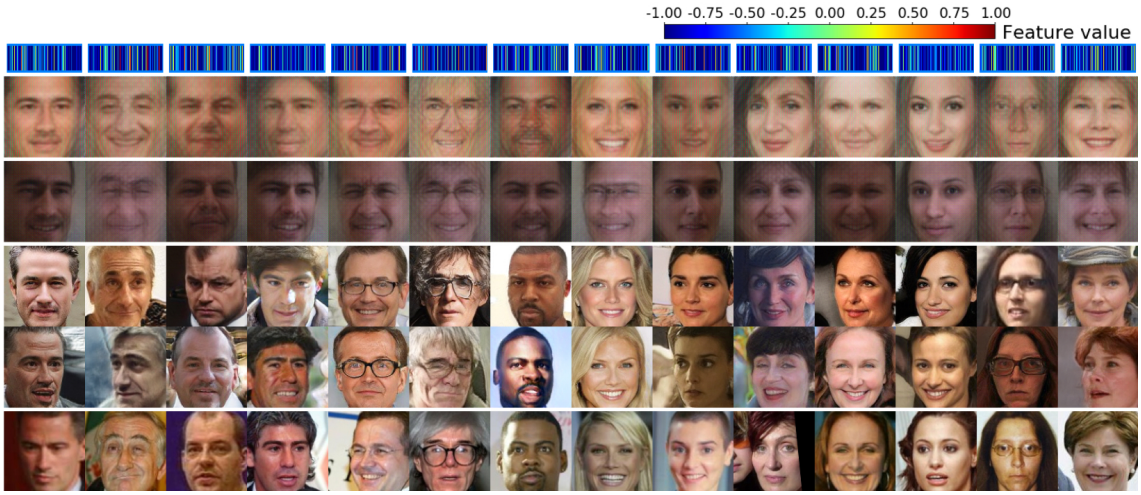


Figure 4. **Feature Reconstruction against in-the-wild facial variations.** For each subject, given an input feature (1st row), while VGG-NBNet [34] and MPIE-NBNet [34] (2nd and 3rd rows) reconstruct faces with limited quality, DibiGAN in whitebox (4th row) and blackbox (5th row) modes are able to produce realistic faces with better ID preservation comparable to real faces (6th row).

LFW [18], AgeDB [36], and CFP-FP [41] which provide face verification protocols against different in-the-wild face variations. Since each face matcher engine requires different preprocessing process, the training and testing data are aligned to the required template accordingly.

**Network Architectures.** We exploited the Generator structure of PO-GAN [24] with 5 convolutional blocks for $G$ while the Feature Conditional branch consists of 8 fully connected layers. The discriminator $D$ includes five consecutive blocks of two convolution and one downsampling operators. In the last block of $D$, the minibatch-stddev operator followed by convolution and fully connected are also adopted. AdaIN operator [19] is applied for feature injection node. For the bijection $H$, we set a configuration of 5 sub-mapping functions where each of them is presented with two 32-feature-map residual blocks. This structure is trained using the log-likelihood objective function on Casia-WebFace. Resnet-50 [17] is adopted for $F^S$.

**Model Configurations.** Our framework is implemented in TensorFlow and all the models are trained on a machine with four NVIDIA P6000 GPUs. The batch size is set based on the resolution of output images, for the very first resolution of output images ($4 \times 4$), the batch size is set to 128, the batch size will be divided by two when the resolution of images is doubled. We use Adam Optimizer with the started learning rate of $0.0015$. We experimentally set $\{\alpha = 0.001, \beta = 1.0, \lambda_j = 1, \lambda_a = 10.0\}$ to maintain the balanced values between loss terms.

**Ablation Study.** To study the effectiveness of the proposed bijective metric for image reconstruction task, we employ an ablation study on MNIST [26] with LeNet [27] as the function $F$. We also set to whitebox mode where $F$ is directly used in $\mathcal{L}^{distill}$ to remove the effects of other factors. Then 50K training images from MNIST and their $1 \times 1024$ feature vectors are used to train $G$. Notice that since the image size is $32 \times 32$, $G$ and $D$ structures are configured with three convolutional blocks. The resulting distributions of synthesized testing images of all classes without and with $\mathcal{L}^{biject}$ are plotted in Fig. 3. Compared to $G$ learned with only classifier-based metrics (Fig. 3(a)), the one with bijective metric learning (Fig. 3(b)) is supervised with more direct metric learning mechanism in image domain, and, therefore, shows the advantages with enhanced intra- and inter-class distributions.

### 4.1. Face Reconstruction Results

This section demonstrates the capability of our proposed methods in terms of effectively synthesizing faces from subject's features. To train DibiGAN, we adopt the ArcFace-Resnet100 [5] trained on 5.8M images of 85K subjects for function $F$ and extract the $1 \times 512$ feature vectors for all training images. These features together with the training images are then used to train the whole framework. We divided the experiments in two settings, i.e. *whitebox* and *blackbox*, where the main difference is the visibility of the matcher structure during training process. In the whitebox mode $F$ is directly used in Eqn. (7) to evaluate $\mathcal{L}_G^{distill}$ while in the blackbox mode, $F^S$ is learned from $F$ through a distillation process as in Eqn. (6) and used for $\mathcal{L}_G^{distill}$. The first row of Table 2 shows the matching accuracy of $F$ and $F^S$ using real faces on benchmarking datasets.

**Face Reconstruction from features of frontal faces.** After training, given only the deep features extracted from $F$ on testing images, the generator $G$ is applied to synthesize the subjects' faces. Qualitative examples of our synthesized faces in comparison with other methods are illustrated in Fig. 4. As can be seen, our generator $G$ is able to reconstruct realistic faces even when their embedding features are extracted from faces with a wide range of in-the-wild

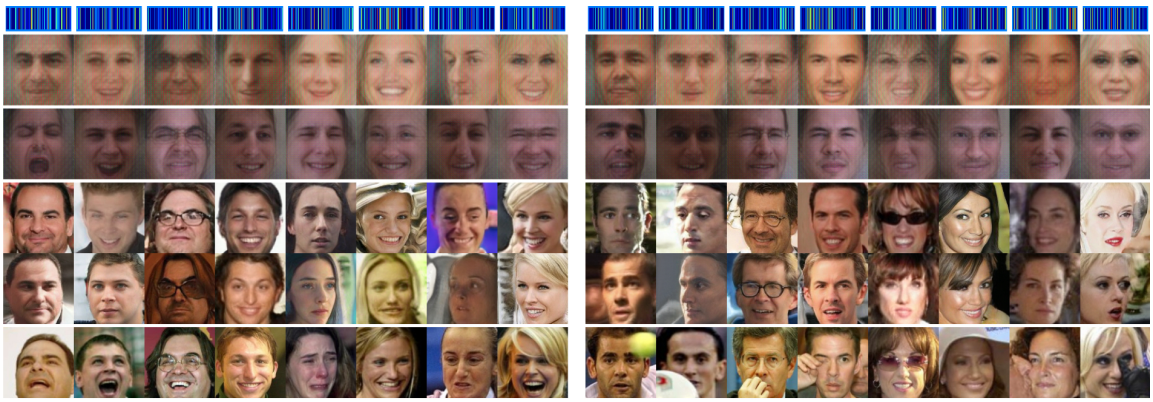


Figure 5. **Feature Reconstruction against expressions (a) and occlusions (b).** For each subject, the 1st row shows the input feature. The next five rows are VGG-NBNet [34], MPIE-NBNet [34], Our DibiGAN in whitebox and blackbox settings, and Real Faces, respectively.

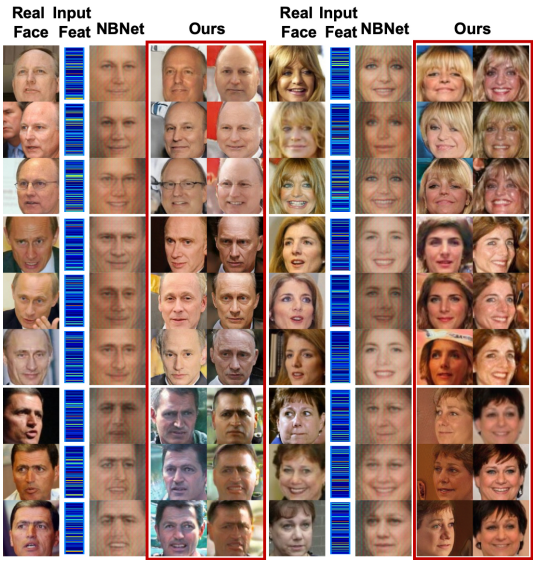


Figure 6. **Feature Reconstruction against features of the same subject.** For each subject, the first and second columns show different real faces and their features of a subject. Compared to VGG-NBNet [34] (third column), our DibiGAN in whitebox and blackbox modes can effectively preserve the ID of the subjects.

variations. More importantly, in both whitebox and blackbox settings, our proposed method successfully preserves the ID features of these subjects. In whitebox setting, since the structure of $F$ is accessible, the learning process can effectively exploit different aspects of embedding process from $F$ and produce a generator $G$ that depicts better facial features of the real faces. For example, together with ID information, poses, glasses, or hair style from the real faces can also be recovered. On the other hand, although the accessible information is very limited in blackbox setting, the learned $G$ can still be enjoyed from the distilled knowledge of $F^S$ and effectively fill the knowledge gap with whitebox setting. In comparison to different configurations of NBNet [34], better faces in terms of image quality and ID preservation can be obtained by our proposed model.

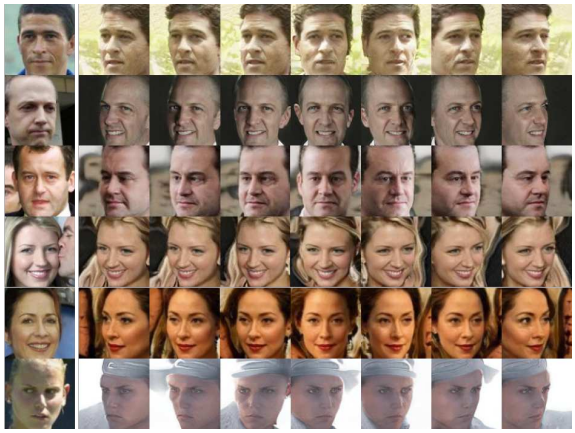

Figure 7. From the input features, our model can synthesize various conditions of a face by varying the "background" variable $\mathbf{v}$.

**Effect of expressions and occluded regions.** Fig. 5 illustrates our synthesis from features of faces that contain both expressions and occlusions. Similar to previous experiment, our model robustly depicts realistic faces with similar ID features as in the real faces. Those reconstructed faces' quality again outperforms NBNet in both realistic and ID terms. Notice that the success of robustly handling with those challenging factors comes from two properties: (1) The matcher $F$ was trained to ignore those facial variations in its embedding features; and (2) both bijective metric learning and distillation process can effectively exploit necessary knowledge from $F$ as well as real face distributions in image domain for synthesis process.

**Effect of different features from the same subject.** Fig. 6 illustrates the advantages of our method in synthesizing faces given different feature representations of the same subject. These results further show the advantages of the proposed bijective metric in enhancing the boundary between classes and constrain the similarity between reconstructed faces of the same subject in image domain. As a result, reconstructed faces from features of the same subject ID not only keep the features of that subject (i.e. similar to

Table 2. **Realism Quality and Matching Accuracy.** Comparison results in Multi-Scale Structural Similarity (MS-SSIM) (*the smaller value is better*); Inception score and Matching Accuracy (*the higher value is better*). For each configuration in (A)-(C) and (D)-(F), each loss function is cumulative enable on the top of the previous configuration. – denotes "not applicable".

| | **White-box Reconstruction** | | | | | **Black-box Reconstruction** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **CelebA** | | **LFW** | **AgeDB** | **CFP-FP** | **CelebA** | | **LFW** | **AgeDB** | **CFP-FP** |
| | MS-SSIM | IS | | | | MS-SSIM | IS | | | |
| Real Faces [1] | 0.305 | 3.008 | 99.78% | 98.40% | 97.10% | 0.305 | 3.008 | 99.70% | 96.80% | 93.10% |
| VGG-NBNet [34] | – | – | – | – | – | 0.661 | 1.387 | 91.42% | 80.42% | 74.63% |
| MPIE-NBNet [34] | – | – | – | – | – | 0.592 | 1.484 | 93.17% | 79.45% | 78.51% |
| (A) PO_GAN [24] | **0.331** | **2.226** | 68.20% | 63.42% | 68.89% | **0.315** | 2.227 | 66.63% | 62.37% | 65.59% |
| (B) + $\mathcal{L}^{distill}$ | 0.343 | 2.073 | 96.03% | 83.33% | 79.07% | 0.337 | **2.238** | 94.95% | 81.56% | 78.80% |
| (C) + $\mathcal{L}^{biject}$ | 0.358 | 2.052 | **98.1%** | **88.16%** | **88.01%** | 0.360 | 2.176 | **97.30%** | **85.71%** | **82.51%** |
| (D) **Ours** | 0.316 | 2.343 | 79.82% | 77.20% | 81.71% | 0.305 | **2.463** | 77.57% | 76.83% | 80.66% |
| (E) + $\mathcal{L}^{distill}$ | **0.306** | 2.349 | 97.76% | 92.33% | 89.20% | 0.305 | 2.423 | 97.06% | 91.70% | 84.83% |
| (F) + $\mathcal{L}^{biject}$ | 0.310 | **2.531** | **99.18%** | **94.18%** | **92.67%** | **0.303** | 2.422 | **99.13%** | **93.53%** | **89.03%** |

real faces) but also share similar features among each other.

**Effect of random variable v.** As mentioned in Sec. 3.4, the variable $\mathbf{v}$ is incorporated to model background factors so that $G$ can be more focused on modeling ID features. Therefore, by fixing the input feature and varying this variable values, different conditions of that face can be synthesized as shown in Fig. 7. These results further illustrate the advantages of our model structure in its capability of capturing various factors for the reconstruction process.

## 4.2. Face Quality and Verification Accuracy

In order to quantitatively validate the realism of our reconstructed images and how well they can preserve the ID of the subjects, three metrics are adopted: (1) Multi-scale Structural similarity (MS-SSIM) [37]; (2) Inception Score (IS) [38]; and (3) face verification accuracy.

**Image quality.** To quantify the realism of the reconstructed faces, we synthesize testing images of CelebA in several training configurations as shown in Table 2, where each loss function in cumulatively enables on the top of the previous configuration. Then MS-SSIM and IS metrics are applied to measure their image quality. We also compare our model in both whitebox and blackbox settings with other baselines including PO_GAN structure [24] and NBNet [34]. Notice that we only adopt the adversarial and reconstruction losses for configs (A) and (D). For all configs (A), (B), and (C), PO_GAN baseline takes only the embedding features as its input. These results show that in all configurations, our method maintains comparative reconstruction quality as PO_GAN and very close to that of real faces. Moreover, our synthesis consistently outperforms NBNet in both metrics.

**ID Preservation.** Our model is experimented against LFW, AgeDB, and CFP-FP where an image in each positive pair is substituted by the reconstructed one while the remaining image of that pair is kept as the reference real face. The matching accuracy is reported in Table 2. These results further demonstrate the advantages and contributions of each component in our framework. Compared to PO_GAN structure, our Feature-Conditional Structure gives more flexibility in modeling ID features, and achieves better matching accuracy. In combination with distilled knowledge from $F^S$, the Generator produces a big jump in accuracy and close the gap to real faces to only 2.02% and 2.72% on LFW in whitebox and blackbox settings, respectively. By further incorporating the bijective metric, these gaps are further reduced to only 0.6% and 0.65% for the two settings.

[1] We report the accuracy of original matcher $F$ for whitebox setting and $F^S$ for blackbox setting.

Table 3. Accuracy against different blackbox face matchers.

| **Matcher** | **LFW** | **AgeDB** | **CFP-FP** |
|---|---|---|---|
| ArcFace[5]-Real | 99.78% | 98.40% | 97.1% |
| **ArcFace-Recon** | **99.13%** | **93.53%** | **89.03%** |
| FaceNet[40]-Real | 99.55% | 90.16% | 94.05% |
| **FaceNet-Recon** | **98.05%** | **89.80%** | **87.19%** |
| SphereFacePlus[28]-Real | 98.92% | 91.92% | 91.16% |
| **SphereFacePlus-Recon** | **97.21%** | **88.98%** | **86.86%** |

**Reconstructions against different Face Recognition Engines.** To illustrate the accuracy of our propose structure, we validate the its performance against different face recognition engines as shown in Table 3. All Generators are set to blackbox mode and only the final extracted features are accessible. Our reconstructed faces are able to maintain the ID information and achieve competitive accuracy as real faces. These performance again emphasizes the accuracy of our model in capturing behaviours of the feature extraction functions $F$ and provides high quality reconstructions.

# 5. Conclusions.

This work has presented a novel generative structure with Bijective Metric Learning for feature reconstruction problem to unveil the subjects' faces given their deep blackboxed features. Thanks to the introduced Bijective Metric and Distillation Knowledge, our DibiGAN effectively maximizes the information to be exploited from a given blackbox face matcher. Experiments on a wide range of in-the-wild face variations against different face matching engines demonstrated the advantages of our method on synthesizing realistic faces with subject's visual identity.

## References


[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *ICML*, 2018.

[2] Thomas Brunner, Frederik Diehl, and Alois Knoll. Copy and paste: A simple but effective initialization method for black-box adversarial attacks. *ArXiv*, abs/1906.06086, 2019.

[3] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *ArXiv*, abs/1906.06919, 2019.

[4] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T. Freeman. Synthesizing normalized faces from facial identity features. In *CVPR*, 2017.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019.

[6] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[7] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *CVPR*, 2016.

[8] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D Bui. Beyond principal components: Deep boltzmann machines for face modeling. In *CVPR*, pages 4786–4794, 2015.

[9] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D Bui. Deep appearance models: A deep boltzmann machine approach for face modeling. *IJCV*, 127(5):437–455, 2019.

[10] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Ngan Le. Shrinkteanet: Million-scale lightweight face recognition via shrinking teacher-student networks. *arXiv preprint arXiv:1905.10620*, 2019.

[11] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, Nghia Nguyen, Eric Patterson, Tien D Bui, and Ngan Le. Automatic face aging in videos via deep reinforcement learning. In *CVPR*, pages 10013–10022, 2019.

[12] Chi Nhan Duong, Kha Gia Quach, Ibsa Jalata, Ngan Le, and Khoa Luu. Mobiface: A lightweight deep learning face recognition on mobile devices. *BTAS*, 2019.

[13] Chi Nhan Duong, Kha Gia Quach, Khoa Luu, Ngan Le, and Marios Savvides. Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition. In *ICCV*, Oct 2017.

[14] Chi Nhan Duong, Kha Gia Quach, Khoa Luu, T Hoang Ngan Le, Marios Savvides, and Tien D Bui. Learning from longitudinal face demonstration—where tractable deep modeling meets inverse reinforcement learning. *IJCV*, 2019.

[15] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Learning perceptually-aligned representations via adversarial robustness. *arXiv preprint arXiv:1906.00945*, 2019.

[16] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NIPS*, pages 5767–5777, 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] 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.

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017.

[20] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, pages 2137–2146, 2018.

[21] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *ICLR*, 2019.

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[24] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[25] H. N. Le, K. Seshadri, K. Luu, and M. Savvides. Facial aging and asymmetry decomposition based approaches to identification of twins. *PR*, 48:3843–3856, 2015.

[26] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[27] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[28] Weiyang Liu, Rongmei Lin, Zhen Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song. Learning towards minimum hyperspherical energy. In *NIPS*, 2018.

[29] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pages 212–220, 2017.

[30] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Xiaodong Song. Delving into transferable adversarial examples and black-box attacks. *ArXiv*, abs/1611.02770, 2016.

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[32] K. Luu, K. Seshadri, M. Savvides, T. D. Bui, and C. Y. Suen. Contourlet appearance model for facial age estimation. In *IJCB*, pages 1–7. IEEE, 2011.

[33] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *CVPR*, pages 5188–5196, 2015.

[34] Guangcan Mai, Kai Cao, Pong C Yuen, and Anil K Jain. On the reconstruction of face images from deep face templates. *TPAMI*, 41(5):1188–1202, 2018.

[35] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. *CVPR*, pages 2574–2582, 2015.

[36] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *CVPRW*, pages 51–59, 2017.

[37] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, pages 2642–2651. JMLR. org, 2017.

[38] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, pages 2234–2242, 2016.

[39] Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Computer vision with a single (robust) classifier. In *ArXiv preprint arXiv:1906.09453*, 2019.

[40] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[41] 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 *WACV*, pages 1–9. IEEE, 2016.

[42] Satya Narayan Shukla, Anit Kumar Sahu, Devin Willmott, and J. Zico Kolter. Black-box adversarial attacks with bayesian optimization. *ArXiv*, abs/1909.13857, 2019.

[43] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *NIPS*. 2018.

[44] Simen Thys, Wiebe Van Ranst, and Toon Goedeme. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In *CVPRW*, June 2019.

[45] 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 *CVPR*, pages 5265–5274, 2018.

[46] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515. Springer, 2016.

[47] F. Xu, K. Luu, and M. Savvides. Spartans: Single-sample periocular-based alignment-robust recognition technique applied to non-frontal scenarios. *TIP*, 24:4780–4795, 2015.

[48] J. Xu, K. Luu, M. Savvides, T. D. Bui, and C. Y. Suen. Investigating age invariant face recognition based on periocular biometrics. In *IJCB*. IEEE, 2011.

[49] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *ACM CCS*, pages 225–240, 2019.

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

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

[52] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *CVPR*, 2019.

[53] Andrey Zhmoginov and Mark Sandler. Inverting face embeddings with convolutional neural networks. *ArXiv*, abs/1606.04189, 2016.