electronics

MDPI

Article

Secure Dual Network for Reversible Facial Image Anonymization Through the Latent Space Manipulation

Yi-Lun Pan [1,2], Jun-Cheng Chen [3] and Ja-Ling Wu [1,*]

[1] Department of Computer Science & Information Engineering, National Taiwan University, Taipei City 10617, Taiwan; d06922016@csie.ntu.edu.tw or serenapan@narlabs.org.tw
[2] National Center for High-Performance Computing, Hsinchu City 30076, Taiwan
[3] Research Center for Information Technology Innovation, Academia Sinica, Taipei City 11529, Taiwan; pullpull@citi.sinica.edu.tw
* Correspondence: wjl@cmlab.csie.ntu.edu.tw

Abstract: We develop a method to automatically and stably anonymize and de-anonymize face images with encoder-decoder networks and provide a robust and secure solution for identity protection. Our fundamental framework is a Neural Network (NN)-based encoder-decoder pair with a dual inferencing mechanism. We denote it as the Secure Dual Network (SDN), which can simultaneously achieve multi-attribute face de-identification and re-identification without any pre-trained/auxiliary model. In more detail, the SDN can take responsibility for successfully anonymizing the face images while generating surrogate faces, satisfying the user-defined specific conditions. Meanwhile, SDN can also execute the de-anonymization procedure and visually indistinguishably reconstruct the original ones if re-identification is required. Designing and implementing the loss functions based on information theory (IT) is one of the essential parts of our work. With the aid of the well-known IT-related quantity, Mutual Information, we successfully explained the physical meaning of our trained models. Extensive experiments justify that with pre-defined multi-attribute identity features, SDN generates user-preferred and diverse appearance anonymized faces for successfully defending against attacks from hackers and, therefore, achieves the goal of privacy protection. Moreover, it can reconstruct the original image nearly perfectly if re-identification is necessary.

Keywords: anonymization; de-anonymization; dual inferencing; encoder-decoder framework; mutual information

Citation: Pan, Y.-L.; Chen, J.-C.; Wu, J.-L. Secure Dual Network for Reversible Facial Image Anonymization Through the Latent Space Manipulation. *Electronics* **2024**, *13*, 4398. https://doi.org/10.3390/electronics13224398

Academic Editor: Aniello Castiglione

Received: 3 October 2024
Revised: 5 November 2024
Accepted: 8 November 2024
Published: 9 November 2024

1. Introduction

With the vigorous development of high-capacity IoT devices and high-resolution cameras, it is becoming more accessible and easier to precisely capture personally identifiable information (PII) [1]. The widespread use of mobile devices, such as smartphones, vastly reduces the barricades for generating facial images. At the same time, the development of social media has promoted the acquisition, spreading, and dissemination of facial image-associated pieces of information. Furthermore, with recent progress in Machine Learning, Artificial Intelligence, and Computer Vision, the quality and performance of Facial Image-related applications have been enhanced significantly.

On one side, the ease of capturing and processing high-quality PII, precisely the facial image, is beneficial to many identity-related applications, such as gate control, dataset access control, and website registration, to name a few. Conversely, the abovementioned factors seriously threaten image holders' privacy and security. Facing this menace to privacy, in 2018, the European Union officially revised and upgraded the General Data Protection Regulation (GDPR) [2] and regulated all data related to PII (especially face images), whether collected or organized, analyzed, and infused, must be carefully protected.

Since the launch of GDPR, topics related to Privacy-Preserving Information Processing (PPIP) have received rocket-high attention in various communities. The top-listed task

to conquer the PPIP challenge is to protect sensitive data from abuse or use by malicious people by anonymizing any PII information and keeping the ability to recover the identity correctly when necessary. This application scenario presents a conflicting trade-off between privacy and utility [3]. To balance the pre-described trade-off, we propose a controllable and reversible Neural Network (NN)-based privacy-preserving framework for facial image anonymization. In the proposed framework, we can complete the tasks of facial image anonymity, also known as de-identification (or de-ID for short), and de-anonymity, also referred to as re-identification (or re-ID for short), simultaneously.

The desired properties of de-ID and re-ID can further be elaborated as follows. In general, generating high-fidelity anonymized images while keeping the original data distribution intact is challenging (where we preserve the invariance of distribution to ease the de-ID task). In other words, with the help of a robust system, we tried to replace the target face with a surrogate one so that the system could synthesize a natural-looking complexion and a realistic face with a background satisfying user-specified conditions. Therefore, the proposed system needs to take into account two essential properties. First, the transition between the original and the anonymous faces must be seamless. Second, the system should still provide strong security protection under other desired features that are forcedly specified.

To our best understanding, this study is the first to incorporate a dual inference mechanism to address the properties mentioned earlier, particularly for facial image de-identification purposes. For example, users can specify that the anonymized faces retain their original hair color by selecting it as one of the pre-defined system parameters associated with specified facial attributes. In this work, we use a multi-attribute feature vector to define the facial qualities involved in the de-ID processes, where the feature vector broadly has the following two classes of attributes: identity-related attributes such as gender, age, and facial expression, and style-related attributes such as hair color and skin color. We refer to the former as identity features and the latter as style features.

On the contrary, during the re-ID process, it is crucial to have a robust system that can recover the original faces from the anonymized ones while also considering the supplemented multi-attribute information such as facial attributes and user-selected passwords. Like its de-ID counterpart, this requirement again presents various challenges to the system design. To tackle these issues, we have developed a reversible privacy protection system for anonymizing and de-anonymizing facial images called the Secure Dual Network (SDN), which utilizes a single NN capable of handling de-ID and re-ID processes equipped with a user-specified password and a user-selected facial attribute feature vector. Experimental results justified that SDN can simultaneously fulfill the requirements of anonymizing and de-anonymizing facial images without compromising the data distribution. Notably and counter-intuitively, SDN will reconstruct near-original images while generating diverse anonymized face images to deceive malicious attackers when the received multi-attribute combinations or passwords are incorrect. This characteristic lifts our system's security level and empowers its usage in privacy-preserving applications.

Figure 1 illustrates two application scenarios of the SDN receiving various inputs for generating anonymized faces. In the first case (lower portion), users provide a single password as input to the SDN to generate the anonymized face. Upon activation of the de-ID process, a well-designed password scheme ensures that the SDN produces a high-quality near-original surrogate image; otherwise, the SDN generates another anonymized image for security consideration. In the second case (upper portion), the SDN receives a multi-attribute combination as input, which could include a password and other facial attributes, such as gender and hair color. The SDN will generate various anonymized faces based on the inputted combinations. When de-anonymization is activated, users must input both the correct password and the multi-attribute combination to obtain an accurate reconstruction result. More precisely, the SDN can reconstruct a near-original image (although it is not the same as the original, we guarantee its visual indistinguishability with the original and pass the inspection through ID verification tools) when the designed password and

multi-attribute combinations are valid. However, suppose users input only the correct password but an incorrect multi-attribute combination. In that case, the system cannot generate the near-original image, and instead produces other anonymized pictures (visually distinguishable from the original and cannot pass the inspection of ID verification).

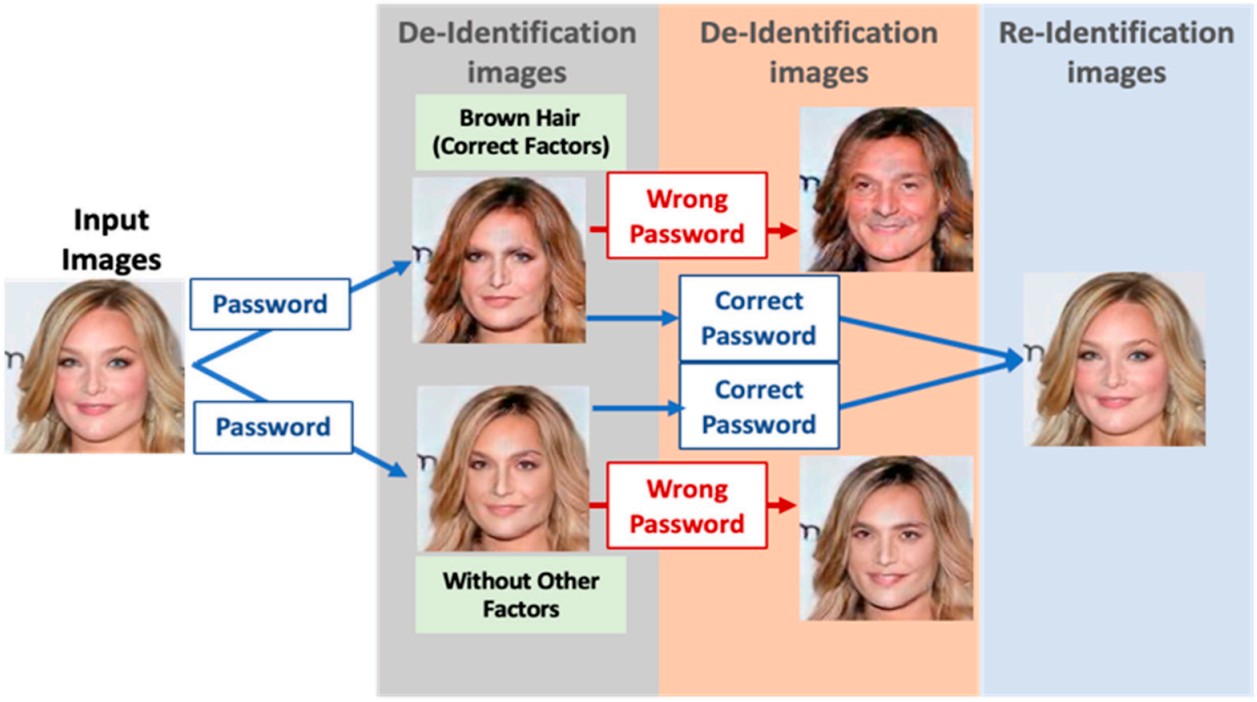


Figure 1. The schematic diagram of the proposed SDN for re-ID.

To conduct our experiments, we utilized three datasets—FaceScrub [4], CASIA-WebFace [5], and CelebA-HQ/CelebA [6]—to train and test our proposed system. Our experimental results demonstrate that our system achieves its multi-task learning objective, generating photo-realistic anonymized images while recovering the original face images without compromising privacy protection. We chose the state-of-the-art anonymization approaches [7] as the comparison benchmark to justify SDN's performance superiority. To further support our system's effectiveness, we present quantitative and qualitative results and compare them to competing de-ID techniques in Section 6.

2. Related Work

Recent breakthroughs in deep generative learning models have significantly improved face de-ID and re-ID techniques. This section briefly overviews recent related work in two fields: (1) Facial de-ID and re-ID and (2) Deep Face Generation.

2.1. Facial De-Identification and Facial Re-Identification

Facial de-ID involves transferring various facial attributes among users and has motivated several exciting research works. Initially, research on facial de-ID focused on using simple image processing operations, such as blacking-out, pixilation, and blurring, to remove privacy-sensitive information from a facial image [8,9]. However, these methods produced poorly anonymized faces and did not substantially enhance privacy protection by altering the data distribution [9]. As a result, they are not widely adopted, particularly when an alteration in data distribution is not allowed because this change of distribution may make re-ID a nearly impossible task.

To overcome these limitations, researchers proposed eigenvector-based solutions that reconstruct faces by combining a fraction of the eigenfaces to hide ID information [10]. Another similar approach proposed by [11] used watermarking, hashing, and PCA data

representations to hide ID information. Other researchers have developed multi-attribute models that unify linear, bilinear, and quadratic data fitting solutions, but require an active appearance model (AAM) to provide the landmark information [12]. Despite these advances, the anonymized quality of the generated images may need to be made more realistic.

Subsequently, the k-Same family algorithms emerged as the most popular methods for face de-ID, which have been widely used in recent works [13]. These algorithms have implemented an effective k-Anonymity algorithm [14] for generating face images. However, the resulting images may contain "ghosting" artifacts due to minor alignment errors [13]. With the popularity of deep neural networks (DNNs), Meden et al. [15] proposed the k-Same-Net scheme to produce photo-realistic de-identified faces by integrating the k-Anonymity algorithm with generative neural networks (GNNs). Although this approach achieved state-of-the-art results at the time, it still had three main limitations. First, selecting cluster centroids using the traditional PCA algorithm demanded substantial computational resources. Additionally, the training process for GNNs was quite time-intensive. Finally, this approach required downsampling the original images during training and synthesis, potentially impacting the resulting images' quality.

The follow-up works on NN-based facial de-ID research have focused on implementing effective and efficient k-Same family algorithms for generating anonymized face images [13–17]. However, these algorithms suffer from weaknesses such as producing "ghosty" artifacts, needing downsampling of original images, and generating unrealistic surrogate images.

In order to overcome the shortages mentioned above, Hukkelås et al. proposed DeepPrivacy [18] to automate the image anonymization process without altering the original data distribution. However, it still suffers from the same issue of producing unrealistic surrogate images. Pan et al. [19] leveraged Generative Adversarial Network (GAN), a labeling scheme, and the k-Same algorithm to generate de-ID image sets without using the data downsampling process. Unfortunately, although the method in Ref. [19] used high-resolution, non-downsampled images, and speeded up the training, the appearance of the generated surrogate is not natural and realistic enough as expected. Jeong et al. [20] proposed a method that uses controllable features to develop more diverse and realistic de-ID images. Inspired by the work done in [21], we derived our current solution that applies dual inference mechanisms for de-ID to increase the diversity of agents' faces and make their appearances more realistic to humans. Additionally, we exploited the properties of dual learning, as concluded in [21], for a theoretical analysis of the proposed SDN.

Regarding re-ID, Yamac et al. [22] and Li et al. [23] proposed reversible privacy-preserving compression approaches that integrated multi-level encryption with compressive sensing techniques. The advantages of these methods include their ability to provide semi- and fully-authorized decryption schemes and a progressive augmentation learning strategy for achieving unsupervised domain adaptive person re-ID. These strengths motivated us to develop a network with enhanced system security.

Gu et al. [24] proposed a generative adversarial learning scheme based on an anchor image and inputted passwords, which inspired us to incorporate password functionality into our system. The idea is to train multiple generative models that can only reconstruct the original input image when users provide the correct password during image recovery. In our development, we found several difficulties associated with their method [24]. For instance, its architecture requires pre-trained networks for additional training, and the to-be-de-identified images must be included in the training set. Therefore, its ability to handle new (i.e., unseen) data is limited. Additionally, it also has limited flexibility of facial features such as facial expression and hair color, which leads the generated images to have the same attributes as the input images. We parameterized the facial features so that extra attributes, such as passwords, are included to enhance system security.

As for the passwords, intuitively, longer passwords provide better protection, but they may lead to instability and longer training times. Regarding diversity, Ref. [24] focused only on generating diversified anonymized images if incorrect passwords have

been inputted, without considering how to control the variety of rendered images, which can be crucial in real-world face de-ID and re-ID applications. In a subsequent study, Pan et al. proposed a method based on multi-attribute combinations to increase the diversity of agent faces for serving de-ID/re-ID purposes [25]. However, their approaches depend on pre-trained models and perform poorly on unseen testing data. These limitations led us to explore the latent space to consider the issues of password length, unseen data, and diversity of generated faces. Of course, requiring no pre-training plays the kernel role in our system design.

Recently, one prominent approach involves reversible anonymization using cyclic learning, where models like CycleGAN transform images between the de-identified and original domains [26]. This method effectively obscures identifiable features while enabling later restoration with high fidelity. The process includes training the model to translate features between domains while maintaining privacy, but with strict controls to prevent unauthorized re-identification, such as cryptographic keys for accessing the re-ID model. This prevents attackers from replicating the model's restoration capabilities only using the output images. However, this method requires significant computational resources and time for training, especially for high-resolution images. The mode collapse is also a critical issue. In contrast, the proposed SDN addresses several limitations observed in CycleGAN-like networks.

Another method called "IDeudemon [27]", leverages 3D morphable models (3DMMs) and neural radiance fields (NeRFs) to de-identify faces while preserving utility in non-identifying areas. The method de-identifies images by embedding Gaussian noise into specific identity features while allowing for high-quality restoration if necessary. This two-step approach involves estimating 3D parameters and using NeRFs for precise adjustments to de-identify faces without compromising the image's quality. The method IDeudemon presents several drawbacks. These include computational complexity and resource intensity, as the processes involved require substantial computational power and can lead to longer processing times.

Additionally, there is a dependency on high-quality data, meaning the method's effectiveness can diminish when applied to lower-quality images. Another concern is the potential for imperfect de-identification, where identifying features may still be retrievable, mainly if the embedded noise is insufficient. Lastly, the method may have limited applicability for diverse faces, as 3DMMs might not generalize well across different facial structures or unique features.

In short, the specific characteristics of SDN include the following four parts: First, according to our observations, most related studies did not include theoretic-based analyses in their system design, which could provide further insights and a more straightforward explanation of the approach's physical meaning. We provide information-theoretic-based cost function designs in SDN in response to this issue. Furthermore, we justify the effectiveness of these functions through a series of experiments. Second, we integrated the dual inference mechanism into our encoder-decoder network to simultaneously complete the anonymization and de-anonymization tasks in the latent space. Third, we utilize structural and style facial image features to broaden the diversity of the appearances of the generated facial images. Finally, we group passwords and facial attributes as the system's hyperparameters to regulate the diversity of the reconstructed images and enhance the system's security level.

2.2. The Deep Face Generation

Numerous researchers have recently investigated ways of using GANs to synthesize and edit realistic human faces at the pixel level, as demonstrated in [28–31]. In our work, GANs are used to synthesize face images, and we devised ways to improve the quality of generated images by exploring a series of related studies such as StarGAN [32], StarGAN v2 [33], and Domain-supervised GAN (DosGAN) [34], which are particularly relevant to our work. DosGAN is an unpaired image-to-image translation framework that takes a

step toward direct image domain supervision [35], and we laid the foundation of our SDN on DosGAN to reflect the progress in DNNs. Additionally, we incorporated the StarGAN function into our SDN's encoder module to increase the diversity and conceal the flaws of the synthesized images.

We used the Residual Net structure to construct our SDN and opted for the PatchGAN discriminator (D-PatchGAN) [36]. Using PatchGAN, we could focus the SDN discriminator's attention on the local image patch structure, which helps enhance the image's quality. Furthermore, we introduced multiple loss functions to DosGAN to develop our SDN for de-identification (de-ID) and re-identification (re-ID) tasks. Inspired by Ref. [37], we employed a latent space-based modeling approach to promote the diversity of SDN.

3. The Proposed Approach

Beyond controllable and reversible, the ability to solve de-ID and re-ID tasks simultaneously in a single unified network is another characteristic of SDN. In this section, we provide a more detailed description of SDN according to the following three main aspects:

1. The characteristics of the network architecture.
2. The disentangled efficacy of the designed algorithm and the adopted loss functions.
3. The analysis of cost function design from the information-theoretic point of view.

When using SDN for anonymization, users only need to input an image, a password, and the desired combination of attributes. On the other hand, during de-anonymization, users must provide the correct password and selected attributes to reconstruct a face image close to the original. If users enter incorrect attribute combinations, SDN will generate a face image associated with a different identity from the original. Additionally, SDN can produce diverse anonymized faces without duplication related to varying combinations of attributes and passwords.

3.1. The SDN's Architecture

3.1.1. The Components of SDN

SDN is built on an encoder-decoder structure and uses the information maximization technique [38] to create a privacy-preserving network designed for de-identification (de-ID) and re-identification (re-ID) tasks. The most critical feature of SDN is integrating a dual inference mechanism within the entire network architecture, as shown in Figure 2, which greatly enhances the diversity of de-ID images. SDN consists of four main subnetworks. First, the Feature Extractor/Classifier Module processes randomly selects three-channel color images from the training datasets to generate a latent feature space. Next, the Encoder Module takes a three-channel color original image, desired attributes (for modifying the latent feature space), and passwords as inputs, producing a de-identified image. Then, the Decoder Module reconstructs the re-ID image associated with the correct ID by taking the de-ID image, correcting specific attributes, and selecting the correct password as input. If incorrect passwords or attributes are provided, the Decoder generates a different de-ID image associated with an ID other than the correct one. Finally, the Discriminator Module employs PatchGAN-D [31] to evaluate the similarity between the original and re-ID images, adjusting the background images accordingly.

As depicted in Figure 2, the Feature Extractor/Classifier subnetwork (labeled as "1" and represented by green rectangular blocks) is primarily responsible for processing randomly selected three-channel color images from the dataset as inputs and generating the latent feature space. The Encoder subnetwork comprises two submodules. One is the Identity Network (labeled as "2" and represented by blue rectangular blocks), which takes attributes and passwords as inputs and generates the latent identity space. Next, the module performs vertical element-wise addition on the latent feature and identity spaces. Then, it feeds the results to another submodule of the Encoder subnetwork, the Erasing Network (labeled as "3" and represented by orange rectangular blocks). The Feature Extractor/Classifier and the Encoder subnetworks execute the so-called Dual Inference Process (DIP). At the same time, the Feature Extractor/Classifier subnetwork performs

another task: classification check. We minimize the cross-entropy between the Feature Extractor/Classifier and the Encoder outputs until the reduction process converges to a bound. In later paragraphs, we will discuss the DIP further. SDN's Decoder subnetwork comprises two submodules, the Identity Network and the Reveal Network (labeled as "4" and described by red rectangular blocks). At this moment, the Identity Network checks whether the entered password is correct. If both attribute and password are valid, the Reveal Network will extract, restore, and identify the embedded features of SDN. Conversely, the Reveal Network will still generate another de-ID image when any one of the attributes or the password is incorrect.

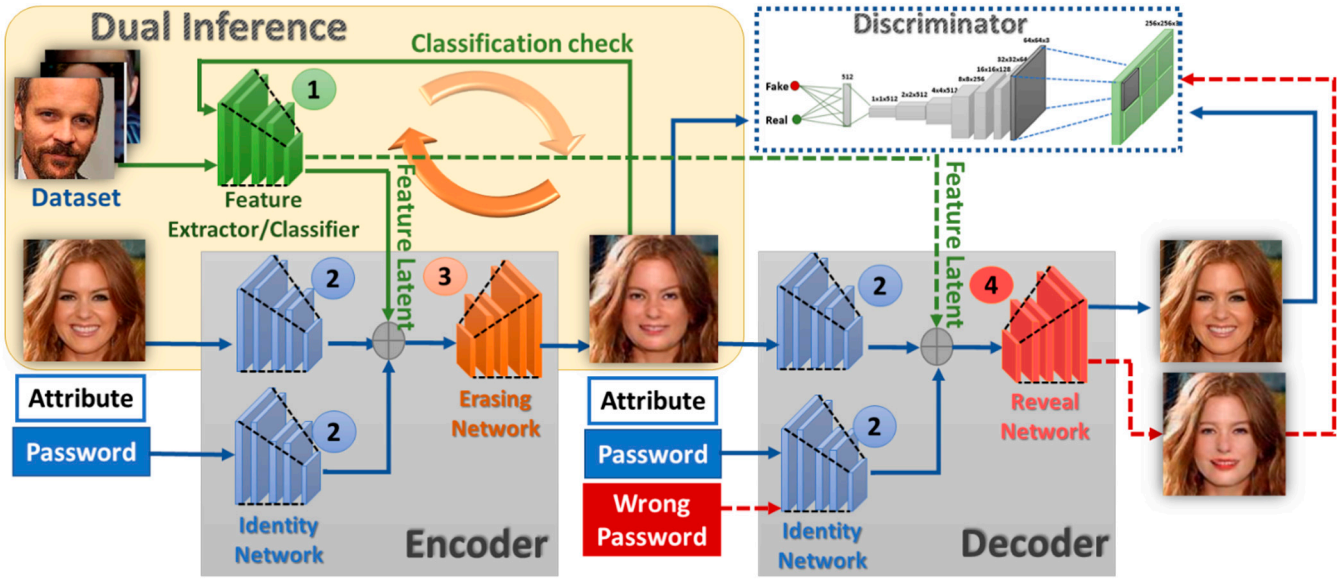


Figure 2. SDN's systematic structure looked at from the viewpoints of subnetworks and functional modules.

3.1.2. The Architecture Perspective of SDN

From a network architecture perspective, the entire SDN involves using a Feature Extractor/Classifier subnetwork to downsample and generate facial feature vectors in the latent space. Next, we use an Encoder subnetwork to compress the identity features to the so-called bottleneck representation, also on the latent space. Then, we conduct element-wise latent space vector additions on both feature spaces. The following Erasing Network performs the upsampling task to generate the de-ID image. The Identity Network takes de-ID-related attributes and authentic or fake passwords as input during the re-ID process and extracts the identity features on the latent space. Finally, we feed the latent identity feature vector into the Reveal Network to reconstruct the three-channel color images. As these functions are similar, the Erasing and Reveal Networks share the same basic structure, but their goals are distinct. The specific function of the Erasing Network aims to mask identity-related features and produce the de-ID image. Conversely, Reveal Network's primary objective is to restore the three-channel color image and accomplish the re-ID task after obtaining the appropriate identity-related features. If incorrect identity-related features are received, our Reveal Network will still generate a three-channel color image associated with a different ID to confuse the attackers.

Based on its architecture, we can also comprehend the information flow of SDN as a set of interactions among the following functional modules: The Dual Inference Module is responsible for the interactions between the Encoder and Feature Extractor/Classifier subnetworks (as indicated by the two brown-colored arcs with an arrow in Figure 2). The Encoder Module is responsible for the interactions between the Identity Network and the Erasing Network (as indicated by the left grey-colored block in Figure 2). The Decoder Module is responsible for the interactions between the Identity Network and the

Reveal Network (as indicated by the right grey-colored block in Figure 2). Finally, the Discriminator Module is responsible for the Discriminator Network only (as noted in the blue-dash line-surrounded block in Figure 2).

Associating with labels of the subnetworks and functional modules addressed above, we redrew the structures and hyperparameters of each layer of the individual subnetwork of the proposed SDN in detail in the following figure. Figure 3 helps interested readers rebuild the SDN and reproduce the experiments conducted in Section 6.

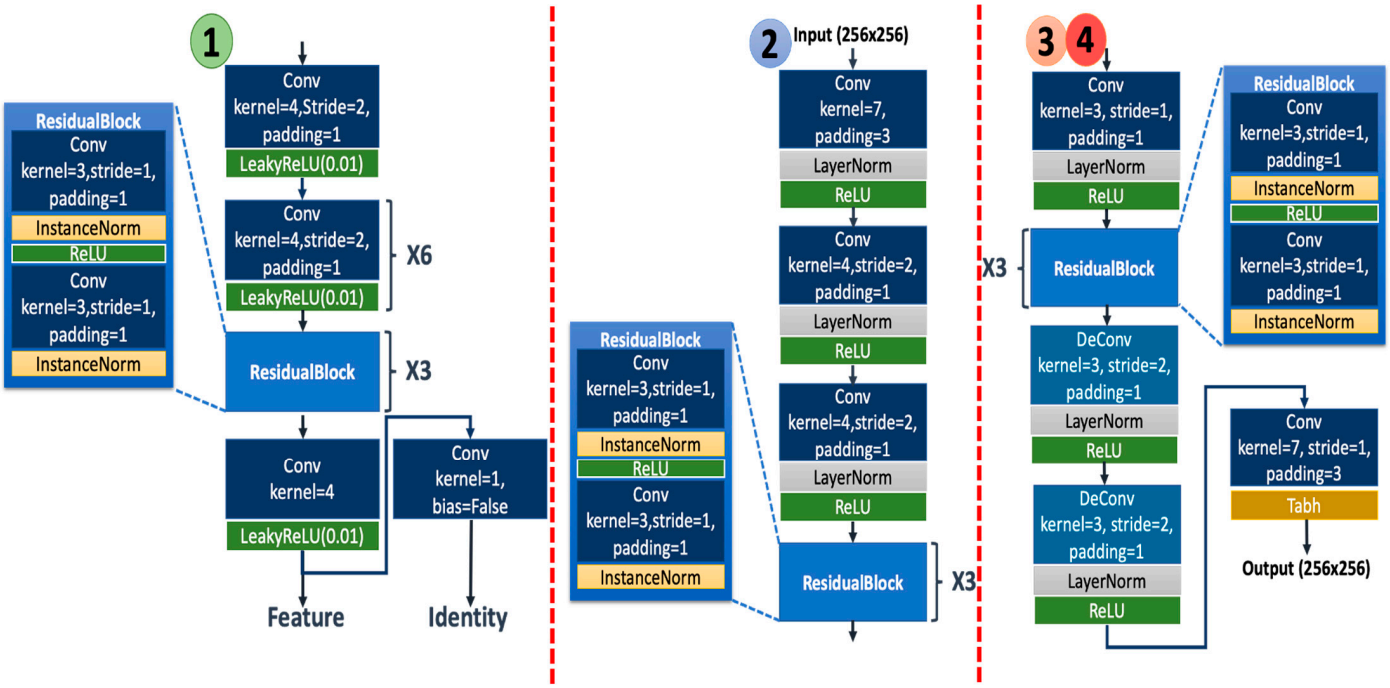


Figure 3. The detailed structures and hyperparameters of each layer within the SDN's subnetworks are depicted in Figure 2.

3.1.3. The Dual Inference Process of SDN

In running the DIP, a Feature Extractor/Classifier (denoted as EC(.)) is used to generate the feature vectors on the latent space Y from the three-channel color image space X. We denote the inference in this direction as the Primal task in Figure 4. In other words, we address the Primal task as a map EC: x $\rightarrow$ y through a conditional probability density function of y given x, parameterized by $\varnothing$, $q_{\varnothing}(y|x)$, where $x \in X$, $y \in Y$. We use an Encoder (denoted as En(.)) to convert the bottleneck representation on the latent space Y back to the three-channel color image space X. Similarly, we denote the inference in this direction as the Dual task in Figure 4. Furthermore, we also express it as a map En(.): y $\rightarrow$ x through a conditional probability density function $p_{\varnothing}(x|y)$. Notice that we conduct element-wise latent space vector additions through the Encoder. Functionally, the designed Feature Extractor/Classifier continues to perform the classification check and calculate the cross-entropy between the outputs of the Feature Extractor/Classifier and the Encoder until the calculated cross-entropy reaches a pre-defined threshold. Therefore, the mathematical expressions of the Primal and Dual tasks can be expressed as Equations (1) and (2), respectively. In summary, Algorithm 1 lists the pseudo-codes of the detailed operational procedures of the proposed SDN.

Algorithm 1: The procedures and the pseudo-codes of the proposed SDN.

Input: A set of face images with face identity labels and multiple face-related attributes
Network Architecture: The Encoder ***En***, Decoder ***De***, Discriminator ***D***, and the Feature Extractor/Classifier $\boldsymbol{EC} \triangleq \{\boldsymbol{F_{ext}}, \boldsymbol{F_{cls}}\}$
Operation: Conduct network training for ***J*** iterations
Output: the SDN Model

For $\boldsymbol{i}$ = 1 to ***J*** **do**
  Randomly select a set of face images $\{\boldsymbol{I_x}, \boldsymbol{I_y}, \boldsymbol{I_z}\}$ with face identity label $\{\boldsymbol{FaceID_x}, \boldsymbol{FaceID_y}, \boldsymbol{FaceID_z}\}$ and attributes labels $\{\boldsymbol{A_x}, \boldsymbol{A_y}, \boldsymbol{A_z}\}$
  Let $\boldsymbol{P}$ be defined as $\{p_{1:\mathbf{password}}, p_{2:\mathbf{multi-attribute}}\}$ ($\triangleq \{p_1, p_2\}$ for short). That is, $\boldsymbol{P}$ stands for the correct **password with specific multi-attribute combinations**.
  Let $\hat{\boldsymbol{P}}$ be an incorrect password with some specific multi-attribute combinations
  Let $\boldsymbol{I'_x} = \{\boldsymbol{I_x}, \hat{\boldsymbol{P}}\}$
  Let $\boldsymbol{I'_y} = \{\boldsymbol{I_y}, \hat{\boldsymbol{P}}\}$
  Generate ***de-identified*** *Image* $Q_y \leftarrow En\ (\boldsymbol{I'_y}, F_{ext}(I_x))$
  Let $\boldsymbol{Q'_y} = \{\boldsymbol{Q_y}, \hat{\boldsymbol{P}}\}$
  if $(\boldsymbol{i} + 1)$ mod 10! = 0 **then**
    Train the Feature Extractor/Classifier ***EC*** by Equation (5)
      Constrain $\boldsymbol{I_x}$ with $\boldsymbol{FaceID_x}$ and $\boldsymbol{Q_y}$ with $\boldsymbol{FaceID_x}$ by minimizing their cross-entropies
    Train **D** by Equation (11)
      Constrain $D_{cls:id}(I_x)$ and $\boldsymbol{F_{cls}}(I_x)$ by minimizing their distance
      Constrain I_x with $\boldsymbol{A_x}$ and Q_y with $\boldsymbol{A_y}$ by minimizing their cross-entropies
  else
      Generate image by applying the same face feature $S_y \leftarrow En\ (\boldsymbol{I'_y}, F_{ext}(I_y))$
      Revert the ***de-identified*** *image* $V_y \leftarrow En\ (\boldsymbol{Q'_y}, F_{ext}(I_y))$
      Generate the ***re-identified*** *image* $R_y \leftarrow De\ (\boldsymbol{Q_y}, P, F_{ext}(I_z))$
      Generate the false ***re-identified image*** $R_z \leftarrow De\ (\boldsymbol{Q_y}, \hat{\boldsymbol{P}}, F_{ext}(I_z))$
      Train the encoder En by Equation (8)
    Compute the L_1 distance between I_y and S_y, and compute the L_1 distance between I_y and V_y
      Constrain Q_y with A_y by minimizing the associated entropy
    Train the decoder **De** by Equation (14)
      Compute the L_2 distance between I_y and R_y
    Constrain R_z with $FaceID_z$ by minimizing the corresponding cross-entropy
  end if

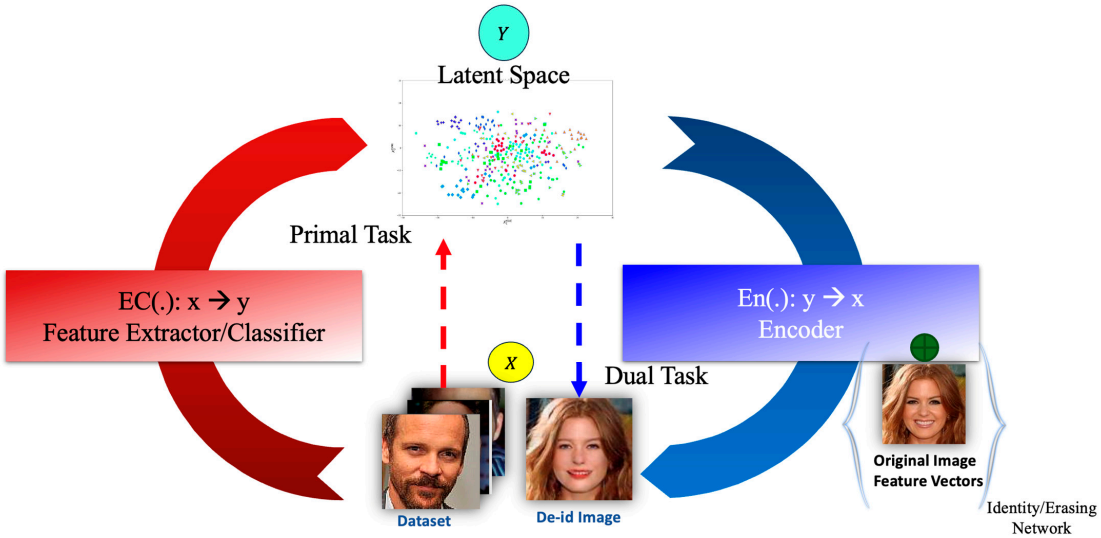


Figure 4. The conceptual schematic diagram of the operational logic and the information flows of the proposed Dual Inference Process.

3.2. The Disentangle Efficacy of the Designed Algorithm

As indicated in Algorithm 1, we first regulate the loss functions to allow the Discriminator and the Feature Extractor/Classifier subnetworks to judge faces faster and more precisely. Then, after every ten runs of computing and updating the Discriminator and Feature Extractor/Classifier subnetworks' losses, the Encoder and Decoder subnetworks perform their regulations on loss and update the learning parameters per iteration. According to the objectives, we explain Algorithm 1 from the following two perspectives.

(a) Anonymization process

We cite the Dual inference mechanism to explain the anonymization process better. In SDN, the Feature Extractor/Classifier mainly executes the Primal task, while the Encoder executes its dual task. Given n data pairs $\{(x_i, y_i)\}_{i=1}^{n}$ which are i.i.d. sampled from the joint space $\mathbf{X} \times \mathbf{Y}$. The training goal is to maximize the likelihood estimation of the parameterized conditional probability distribution, giving rise to the following dual inference optimization problems:

$$\boldsymbol{EC}(x; \theta_{x \to y}) = argmax_{\theta_{x \to y}} q(y|x; \theta_{x \to y}), \tag{1}$$

$$\boldsymbol{En}(y; \theta_{y \to x}) = argmax_{\theta_{y \to x}} p(x|y; \theta_{y \to x}). \tag{2}$$

Notice that we do not re-train or change the models of both the Primal and the Dual tasks. Let us first focus on the calculation of the DIP's loss functions. According to their inference directions, we divide the associated loss function into two parts: $\boldsymbol{\ell}_{\mathbf{EC}}$ and $\boldsymbol{\ell}_{\mathbf{En}}$, which can be understood as the negative log-likelihood of the proposed SDN. Mathematically, we have

$$\begin{aligned} EC(x; \theta_{x \to y}) &= argmin_{y' \in \mathcal{Y}} \boldsymbol{\ell}_{\boldsymbol{EC}}(x, y') \\ &= -logq(y'|x; \boldsymbol{EC}), \end{aligned} \tag{3}$$

$$\begin{aligned} En(y; \theta_{y \to x}) &= argmin_{x' \in \mathcal{X}} \boldsymbol{\ell}_{\boldsymbol{En}}(x', y) \\ &= -\log p(x'|y; \boldsymbol{En}) \end{aligned} \tag{4}$$

Conceptually, we can express the Feature Extractor/Classifier's loss function defined as

$$\begin{aligned} \mathcal{L}(F_{ext},\ F_{cls}) &= \mathcal{L}^{bce}{}_{I_x, FaceID_x}(FaceID_x,\ F_{cls}(I_x)) \\ &+ \mathcal{L}^{bce}{}_{I_y, FaceID_x}(FaceID_x, F_{cls}(En(I_y, P, F_{ext}(I_x)))), \end{aligned} \tag{5}$$

where I_x denotes the x-th randomly selected source image from the dataset, and $\mathcal{L}^{bce}(y,\ \hat{y}) = -\sum ylog(\hat{y})$ stands for the cross-entropy function used to measure the similarity of two distributions y and $\hat{y}$. We use $FaceID_x$ to represent the accurate face domain identity distribution of I_x and let A_x be the set of corresponding attribute labels provided by the dataset. Let I_y represent an unseen target image, which represents the de-ID image required. Let $I'_y = \{I_y,\ \hat{P}\}$ be the set that comprises I_y and its associated password and multi-attribute combinations $\hat{P}$. Moreover, we use F_{cls} and F_{ext} to represent the Feature Classifier and the Feature Extractor, which are parts of $\boldsymbol{EC}(.)$ when dealing with different functions.

The Feature Extractor/Classifier conducts the Primal task according to Equation (5). The first item of Equation (5) is to calculate the cross-entropy between the distribution of the classification results when F_{cls} took the source image I_x as its input and the corresponding actual identity distribution $FaceID_x$. Similarly, the second item of Equation (5) calculates the cross-entropy between $FaceID_x$ and the distribution of the classification results, where the input to F_{cls} is the associated distribution of the de-ID images. Notice that the de-ID images are now rebuilt from the Encoder with the latent space attributes $I'_y = \{I_y,\ \hat{P}\}$ and the output of F_{cls} as inputs. Meanwhile, the Encoder conducts the Dual task according to the associated loss functions, which can be sketched as follows.

$$\mathcal{L}_{rec} = ||I_y - En(I_y,\ P,\ F_{ext}(I_y))||_1 + ||\mathrm{I}_y - En(En(I_y,\ P,\ F_{ext}(I_x)),\ F_{ext}(I_y))||_1 \tag{6}$$

$$\mathcal{L}_{attr} = \mathcal{L}^{bce}{}_{I_y,A_x}\left(A_x,\ En\left(I_y,\ P,\ F_{ext}(I_x)\right)\right) \tag{7}$$

The cost $\mathcal{L}_{rec}$ is to guide F_{ext} to extract the correct facial features and then spread the features evenly over the target image I_y through the Encoder. The smaller the loss $\mathcal{L}_{rec}$, the better. While the cost $\mathcal{L}_{attr}$ measures how close the distributions between A_x and $En\left(I_y,\ P,\ F_{ext}(I_x)\right)$ are. Physically, $\mathcal{L}_{rec}$ uses the selected attributes to control and change the target image I_y and generates the set of attributes (such as hair color and gender) in the de-ID image. Likewise, the smaller the loss $\mathcal{L}_{attr}$, the better. Clearly, the total Encoder's loss would be

$$\mathcal{L}_{En} = \mathcal{L}_{rec} + \mathcal{L}_{attr} \tag{8}$$

Moreover, to measure the distance between the actual image and the synthesis de-ID image, we designed cost functions for supporting the discriminator identity classification $\mathcal{L}_{D_{cls:id}}$ and the discriminator attribute classification $\mathcal{L}_{D_{cls:attr}}$ which can, respectively, be written as:

$$\mathcal{L}_{D_{cls:id}} = \left|\left|D_{cls:id}(I_x) - F_{cls}(I_x)\right|\right|_1 \tag{9}$$

$$\begin{array}{c}\mathcal{L}_{D_{cls:attr}} = \mathcal{L}^{bce}{}_{I_x,A_x}(A_x, D_{cls:attr}(I_x)) + \\ \mathcal{L}^{bce}{}_{I_y,A_y}\left(A_y, D_{cls:attr}\left(En\left(I_y,\ P,\ F_{ext}(I_x)\right)\right)\right)\end{array} \tag{10}$$

The cost $\mathcal{L}_{D_{cls:id}}$ is to guide the Discriminator to classify identity correctly with the help of F_{cls} via calculating the L_1 distance between $D_{cls:id}$ and F_{cls} for the same target image I_x. At the same time, the cost $\mathcal{L}_{D_{cls:attr}}$ guides the Discriminator to classify attributes. The purpose of the first item on the right-hand side of Equation (10), $\mathcal{L}_{D_{cls:attr}}$, is to find the proper classification of all images' attributes in the dataset. Therefore, it calculated the cross-entropy between the labeled attributes in the dataset and the extracted attributes from the original images through $\mathcal{L}_{D_{cls:attr}}$. In comparison, the second item indicated that the attributes generated by the Encoder should be consistent with the attributes of the target image. Hence, it calculates the cross-entropy between the attributes of the de-ID image and the attributes of the target image. In summary, the total Discriminator's loss would be

$$\mathcal{L}_D = \mathcal{L}_{D_{cls-id}} + \mathcal{L}_{D_{cls-attr}} \tag{11}$$

(b) De-anonymization process

This subsection focuses on the primary function of the Decoder. We use it to correctly de-anonymize and output the re-ID image when it receives the correct password and multi-attribute combinations. Conversely, when the Decoder receives an incorrect password and the correct or incorrect multi-attribute combinations, we want the Decoder to reconstruct a different de-ID image to confuse the potential attacker and continue to return to the normal anonymization process. That is, we proceed to calculate the Decoder's loss functions. When the Decoder receives the correct password and correct multi-attribute combinations, it should find the corresponding re-ID image by minimizing the following function

$$\begin{array}{rl}\mathcal{L}_{re-id} = & \left|\left|I_y - De\left(En\left(I_y,\ P,\ F_{ext}(I_x)\right),\ P,\ F_{ext}(I_z)\right)\right|\right|_2 \\ & + \left|\left|I_y - De\left(En\left(I_y,\ P,\ F_{ext}(I_x)\right),\ P,\ F_{ext}(I_z)\right)\right|\right|_1,\end{array} \tag{12}$$

where I_z stands for a different de-ID image.

The first item of Equation (12) calculates the L_2 distance between the original target image I_y and the reconstructed image through the Decoder with the correct password and multi-attribute combinations. Since the L_2 distance calculates the point-by-point Euclidean distance between the actual value and the predicted value, the primary effect of the first term of Equation (12) is to constrain the pixel-by-pixel distance between the target image and the reconstructed image. Therefore, minimizing it will keep all the detailed facial textures intact. In contrast, the second item of Equation (12) calculates the L_1 distance (i.e., the sum of all the absolute differences) between the target image and the reconstructed image. Since the target and reconstructed images are calculated using L_1 distance, it can

help to remove the outliers. Thus, we found that constraining this distance renders the non-facial backgrounds of the two images to coincide as precisely as possible.

On the other hand, whether the inputted multi-attribute combinations are correct or not, if the password received by the Decoder is incorrect, the cost function used to guide the de-anonymization process becomes:

$$\mathcal{L}_{rec} = \mathcal{L}^{bce}{}_{I_z, FaceID_z}(FaceID_z,\ F_{cls}(De(En(I_y,\ P,\ F_{ext}(I_x)),\ \hat{P},\ F_{ext}(I_z))) \tag{13}$$

where $\hat{P}$ stands for the set of inputted incorrect passwords and the involved multi-attribute combinations. As pre-described, $\mathcal{L}_{rec}$'s primary effect is forcing the Decoder to generate false de-ID images to confuse the possible attacker when inputting incorrect passwords. In summary, the total Decoder's loss would be

$$\mathcal{L}_{De} = \mathcal{L}_{re-id} + \mathcal{L}_{rec}. \tag{14}$$

Finally, the total objective loss can be written as

$$\mathcal{L}_{total} = \lambda_{En}\mathcal{L}_{En} + \lambda_{De}\mathcal{L}_{De}, \tag{15}$$

where λ_{En} and λ_{De} are gain control parameters for the Encoder and the Decoder, respectively. We use the following parameter settings, $\lambda_{En} = \lambda_{De} = 10$, for all experiments conducted in this work.

4. Analyzing the Cost Functions Involved in SDN Using Information Theory

In the above subsection, we have formulated two specific cost functions to guide the learning process of the proposed SDN. To be addressed in this subsection, the cost functions related to 'dual inference' consider the Decoder's visual realism and recovery radiality. In visual realism, we use the following minimax game as one of the regularization functions. That is

$$min_{MP_{dual}} max_D V_I(D, MP_{dual}) = V(D, MP_{dual}) - \lambda_1 I(c_{\mathrm{id}}; MP_{dual}(Z_{\mathrm{id}},\ P)), \tag{16}$$

where V is the value function similar as that of GAN formulation, $I(x;y)$ represents the Mutual Information between two random variables x and y, D is the discriminator, MP_{dual} denotes the SDN model, λ_1 is a hyperparameter, and c_{id} stands for the latent codes of the original face image Z_{id}. Like the above, we let $P \triangleq \{p_1,\ A_x\}$ be the set of embedded information that consists of the given password p_1 and the given multi-attribute combinations A_x. Then, we can form the affinity relationship between the original target face image Z_{id}(which plays the same role I_y mentioned previously) and the de-ID image Z_{nid} as $I(c_{\mathrm{id}};\ Z_{\mathrm{nid}}) = I(c_{\mathrm{id}};\ MP_{dual}(Z_{\mathrm{id}}, P))$ after working through the proposed SDN model.

To impose constraints on the allowable visual dissimilarity between the original face image and the de-identified image in the minimax game, we use mutual information (MI) as a regularization term. Specifically, we aim to maximize the mutual information between our model's representation of the original face image and the de-identified image. The following discussions treat the probability distribution P as a random variable. By maximizing the mutual information between the original face image and the de-identified image, we can limit the amount of visual difference tolerated between them while preserving as much information as possible about the original image.

Moreover, because the encoding function, MP_{dual} (.), which relates P and Z_{id}, is deterministic and invertible, and now $I(c_{\mathrm{id}}; Z_{\mathrm{nid}}) = I(c_{\mathrm{id}}; MP_{dual}(Z_{\mathrm{id}}, P))$ behaves as a concave function, we can find its maximal value. This interpretation allows us to easily define a cost function that enforces a specific range of visual differences between c_{id} and Z_{nid}, which is a crucial aspect of privacy protection. By maximizing the mutual information between the original image and the de-identified image, we ensure that the de-identified image retains as much information as possible about the original image while minimizing

the visual differences between them. We can then formulate a cost function that penalizes any visual differences between c_{id} and Z_{nid} that exceed a certain threshold. This threshold can be chosen to reflect the desired level of privacy protection, ensuring that any differences between the original and de-identified images are not perceptually significant. In this way, we can balance the competing goals of preserving privacy while retaining as much helpful information as possible.

From a machine learning perspective, the above expression implies that the information encoded in the latent code of c_{id} will be preserved to a significant extent during the generation process of the SDN model. $I(c_{\mathrm{id}}; Z_{\mathrm{nid}})$ can be expressed as:

$$I(c_{\mathrm{id}}; Z_{\mathrm{nid}}) = I(c_{\mathrm{id}}; MP_{dual}(Z_{\mathrm{id}}, P)) = H(c_{\mathrm{id}}) - H(c_{\mathrm{id}}|MP_{dual}(Z_{\mathrm{id}}, P)). \tag{17}$$

While $MP_{dual}(.)$ is a deterministic and invertible function, it can be challenging to directly compute the maximum value of Equation (17) because we need to gain knowledge of the posterior distribution $p(c_{\mathrm{id}}|MP_{dual}(Z_{\mathrm{id}}, P))$. To overcome this challenge, we use a variational approximation to estimate the mutual information of the encoder, as addressed in the next paragraph.

Let $p(x)$ denote the distribution of the data x, and we need to bound $H(c_{\mathrm{id}}|MP_{dual}(Z_{\mathrm{id}}, P))$ suitably. We can use the non-negative property of conditional entropy, and therefore, we have:

$$H(c_{\mathrm{id}}|MP_{dual}(Z_{\mathrm{id}}, P)) \geq 0. \tag{18}$$

Equation (18) can be employed to obtain a lower bound on the prediction error of c_{id}, given the measurement of $MP_{dual}(Z_{\mathrm{id}}, P)$. This fact leads us to the visual realism cost function expressed in Equation (16) as a minimax game. Similarly, we can formulate the recovery fidelity cost function of the Decoder as another minimax game:

$$min_{MP_{dual}} max_D V_I(D,\ MP_{dual}) = V(D, MP_{dual}) - \lambda_2 I\left(c_{\mathrm{id}}; MP_{dual}\left(Z_{nid},\ \hat{P}\right)\right), \tag{19}$$

where λ_2 is a hyperparameter, and $\hat{P}$ stands for the set of input passwords and the multi-attribute combinations inputted from the user or the hacker.

When $\hat{P} = P$, Cheng et al. [39] have shown that by utilizing a variational marginal approximation $r(MP_{dual}(Z_{nid}, P))$, which is a standard normal distribution, we can construct a variational upper bound expressed as follows:

$$\begin{aligned} I(c_{\mathrm{id}}; Z_{id}) &= I(c_{\mathrm{id}}; MP_{dual}(Z_{nid},\ P)) \\ &= \mathbb{E}_{p(c_{\mathrm{id}},\ MP_{dual}(Z_{nid},\ P))}[log\tfrac{p(MP_{dual}(Z_{nid},P)|c_{\mathrm{id}})}{p(MP_{dual}(Z_{nid},P))}] \\ = \mathbb{E}_{p(c_{\mathrm{id}},\ MP_{dual}(Z_{nid},\ P))}\left[log\tfrac{p(MP_{dual}(Z_{nid},P)|c_{\mathrm{id}})}{r(MP_{dual}(Z_{nid},P))}\right] &- KL(p(MP_{dual}(Z_{nid}, P))||r(MP_{dual}(Z_{nid}, P))) \\ \leq \mathbb{E}_{p(c_{\mathrm{id}},\ MP_{dual}(Z_{nid},\ P))}\left[log\tfrac{p(MP_{dual}(Z_{nid},P)|cc_{\mathrm{id}})}{r(MP_{dual}(Z_{nid},P))}\right] &= KL(p(MP_{dual}(Z_{nid}, P)|c_{\mathrm{id}})||r(MP_{dual}(Z_{nid}, P))). \end{aligned} \tag{20}$$

Let us assume that the designed regularization function restricts the value of $KL(p(MP_{dual}(Z_{nid}, P))||r(MP_{dual}(Z_{nid}, P)))$ to be very small. This assumption implies that $r(MP_{dual}(Z_{nid}, P))$ will be compelled to become a well-dense approximation of $p(MP_{dual}(Z_{nid}, P))$ during the learning process. In other words, such a regularization function can facilitate the re-ID process significantly. Similar arguments hold for the case of $\hat{P} \neq P$, where we aim to put bound to $H\left(c_{\mathrm{id}}\middle|MP_{dual}\left(Z_{nid}, \hat{P}\right)\right)$. By using the definition of mutual information and the non-negativity property of the KL divergence, we obtain the following inequality:

$$\begin{aligned} I\left(c_{\mathrm{id}}; MP_{dual}\left(Z_{nid}, \hat{P}\right)\right) &= H(c_{id}) - H\left(c_{\mathrm{id}}\middle|MP_{dual}\left(Z_{nid},\ \hat{P}\right)\right) \\ &\geq H(c_{\mathrm{id}}) + \left\langle \log q\left(c_{\mathrm{id}}\middle|MP_{dual}\left(Z_{nid},\ \hat{P}\right)\right)\right\rangle_{p(c_{\mathrm{id}}, MP_{dual}(Z_{nid}, \hat{P}))} \\ &\triangleq \widetilde{I}\left(c_{\mathrm{id}}; MP_{dual}\left(Z_{nid},\ \hat{P}\right)\right), \end{aligned} \tag{21}$$

where $q\left(c_{\mathrm{id}}\middle|MP_{dual}\left(Z_{nid}, \hat{P}\right)\right)$ uses a different variational distribution, our carefully designed regularization function can ensure that the learned distribution $q\left(c_{\mathrm{id}}\middle|MP_{dual}\left(Z_{nid}, \hat{P}\right)\right)$ will closely approximates the true distribution $p\left(c_{\mathrm{id}}\middle|MP_{dual}\left(Z_{nid}, \hat{P}\right)\right)$. This approximation is equivalent

to a moment-matching approximation of $p(c_{\text{id}}|MP_{dual}(Z_{nid}, \hat{P}))$ by $q(c_{\text{id}}|MP_{dual}(Z_{nid}, \hat{P}))$. We expect to provide readers with a better understanding of the cost functions designed for SDN through the discussions presented above.

5. Experimental Materials and the Chosen Benchmarking Methods

To validate our claims and demonstrate the effectiveness of the proposed SDN, we conducted a series of experiments and compared our results with several selected benchmark works. In this section, we provide details about the experimental settings, including the datasets used, the evaluation metrics employed, and the characteristics of the selected benchmarks. This information will help to contextualize our experimental results and provide a basis for comparing our approach to existing methods.

5.1. The Training Datasets and Evaluation Metrics

We trained our proposed SDN scheme using three different datasets: FaceScrub, CASIA-WebFace, and CelebA-HQ/CelebA. The FaceScrub dataset contains 106,863 face images of 530 male and female celebrities, with 200 images per person, making it one of the largest public face databases. Due to the many images per person, our SDN model can learn face attributes more effectively and be applied to other datasets. Therefore, we used the FaceScrub dataset to train our model and the CASIA-WebFace and CelebA-HQ/CelebA datasets for validation. The CASIA-WebFace dataset contains over 453,453 face images of 10,575 individuals, while the CelebA-HQ/CelebA dataset contains over 30,000 face images of 10,177 individuals. These datasets provide a diverse range of face images, allowing us to test our approach's performance under different conditions.

In addition to the datasets used, we also employed a variety of evaluation metrics to assess the performance of our proposed method. We also use two objective networks, FaceNet [40] (VGGFace2) and FaceNet (CASIA), and one publicly available face recognition tool [41] to measure the identity distance (ID-distance) and successful protection rate (SPR) after completing the anonymization process. Mathematically, we can calculate the ID-distance and the SPR by Equations (22) and (23), respectively. That is,

$$ID\text{-}distance = E_{distance}\Big(F_{reg}(I_y),\ F_{reg}\Big(En\Big(I'_y,\ \ F_{ext}(I_x)\Big)\Big)\Big), \tag{22}$$

where $E_{distance}$ stands for the conventional Euclidean distance. In Equation (22), we let F_{reg} be a face recognition model and use it to recognize faces and generate the identity vectors. We calculate the Euclidean distance between the two sets of identity vectors and get the corresponding ID distance. The two vectors are judged as different subjects when the ID distance exceeds a preset threshold. In our experiments, we set the threshold for FaceNet as t = 1.1 according to Ref. [40] and t = 0.6 for the face recognition tool. As for computing the SPR, the function SP(t) must be defined first as follows.

$$SP(\mathrm{t}) = \Big\{\Big(I_y,\ En\Big(I'_y,\ F_{ext}(I_x)\Big)\Big) \in \mathrm{P}_{\text{de-id}},\ \text{with } ID\text{-}distance >= \mathrm{t}\Big\}, \tag{23}$$

where $\mathrm{P}_{\text{de-id}}$ denotes the set of all de-ID test pairs, and SP(t) is the total number of the pairs with ID-distance greater than the threshold. And then

$$SPR = \frac{SP(\mathrm{t})}{P_{de-id}}. \tag{24}$$

In addition to the evaluation metrics mentioned earlier, we also conducted experiments to assess the quality of the images produced by SDN. Specifically, we used the following perceptual-based image quality metrics: Learned Perceptual Image Patch Similarity (LPIPS), Frechét inception distance (FID), structural-similarity-index-measure (SSIM), and peak signal-to-noise ratio (PSNR). LPIPS measures the distance between image patches, with higher values indicating more significant differences. SSIM measures the similarity between two images, with higher values indicating a more remarkable similarity. The FID metric

assesses the resemblance between two sets of image data and has been demonstrated to correlate strongly with human assessment of visual fidelity. Therefore, it is commonly utilized to appraise the quality of GAN-generated samples. Finally, PSNR measures the visual quality of images calculated by comparing the error between the two images. Higher PSNR values indicate a smaller amount of distortion between the compared images.

To further demonstrate that our SDN scheme behaves similarly to human perception, we used PieAPP [42], a metric designed to simulate human perception for quality assessment. A lower PieAPP error value indicates better image quality, as the images produced by the SDN are more similar to what a human observer would perceive.

5.2. The Benchmarking Methods

In preparation for analyzing our experimental results, we compared the performances of SDN with several existing methods in both the anonymization and de-anonymization scenarios. For anonymization, we compared our results with those of DeepPrivacy [18], Gu et al. [24], MfM [25], CIAGAN [43], and Cao et al. [44]. For de-anonymization, we compared our results with Gu et al. [24], MfM [25], and Cao et al. [44]. Additionally, we implemented the works of Gu et al. [24], Maximov et al. [43], and Cao et al. [44] to expand our comparison ranges.

The following section will provide detailed analyses of our experimental results, including performance comparisons with benchmarked methods and our latent space manipulation outcomes. By delivering comprehensible analyses, we aim to demonstrate the effectiveness and superiority of our SDN scheme in both the anonymization and de-anonymization scenarios.

6. Experimental Results and Latent Space Manipulation Analysis

In this section, we conduct quantitative and qualitative experiments to demonstrate the proposed approach's effectiveness and compare them with the existing relative anonymization and de-anonymization works.

6.1. The Anonymization and De-Anonymization Performances of the SDN

The SDN's quantitative evolution results in dealing with the anonymization task as compared with those of DeepPrivacy [18], Gu et al. [24], MfM [25], CIAGAN [43], and Cao et al. [44] are presented in Figure 5 (in terms of ID-distance) and Figure 6 (in terms of SPR). From the two figures, we observed that SDN is effective for identity protection and superior in providing a more considerable ID distance and a higher SPR. In addition to these superiorities, the SDN can achieve compatible anonymization ability without any pre-train/auxiliary model. This property is very different from the other benchmarked works.

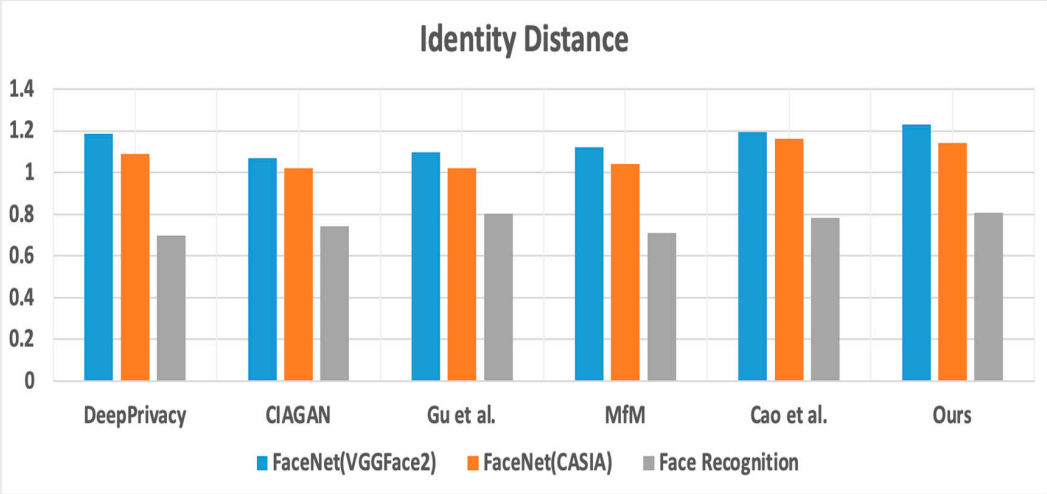


Figure 5. The comparison of the anonymization process's quantitative evolution results between the SDN and benchmarked works regarding ID-distance concerning various testing datasets [15,21,22,38,39].

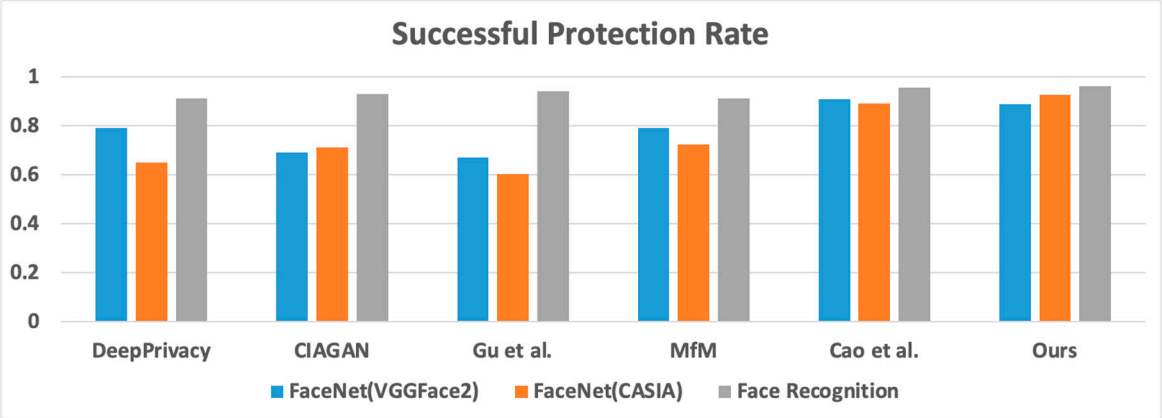


Figure 6. The comparison of the anonymization process's quantitative evolution results between the SDN and benchmarked works regarding SPR concerning various testing datasets [15,21,22,38,39].

The qualitative evaluation results of the SDN's anonymization and de-anonymization performances in utility-related experiments are presented in Tables 1 and 2. Table 1 compares SDN with several state-of-the-art works for de-ID. According to Table 1, Yang et al. achieve the best performance in terms of LPIPS, but SDN performs comparable and better than most current works. Regarding FID, SDN shows some gap compared to MfM and Gu et al., but its performance is still better than most other methods. As for SSIM, Table 1 confirms that SDN delivers the best performance. This information demonstrates that SDN's de-identified synthesized agent faces match well with the original images, including background, facial contours, and edges, to name a few examples.

Table 1. The perceptual quality comparison of the anonymization task between the SDN and benchmarked works concerning four image quality measures.

Method	LPIPS	FID	SSIM
Gu et al. [24]	0.17	28	0.95
MfM [25]	0.35	27	0.83
FaceBERT [45]	0.15	123	0.83
Yang et al. [46]	0.12	144	0.81
A^3GAN [47]	0.29	93	0.87
Khorzooghi et al. [48]	0.28	101	0.86
Xue et al. [49]	0.16	127	0.83
CIAGAN [43]	0.28	108	0.85
Cao et al. [44]	0.29	43	0.93
Ours	0.15	28	0.96
Real Images	-	-	1

Table 2. The perceptual quality comparison of the de-anonymization task between the SDN and benchmarked works concerning four image quality measures.

Method	LPIPS	PSNR	SSIM	PieAPP
Gu et al. [24]	0.038	28.11	0.809	0.532
MfM [25]	0.069	27.52	0.823	0.581
Cao et al. [44]	0.072	27.10	0.85	0.63
Ours	0.034	28.91	0.872	0.451
Real Images	-	-	1	0

Table 2 compares different approaches for de-anonymization tasks based on subjective and objective quality measures. As shown in the boldface numbers in Table 2, SDN performs the best in all tested quality metrics. In other words, besides its controllable and reversible characteristics, SDN performs better regarding LPIPS, and its lower value in LPIPS indicates a lower degree of diversity in the generated images. At the same time, SDN's higher PSNR values suggest that the distortion between the compared images is relatively minor. Moreover, its higher SSIM values indicate relatively good subjective quality in the de-anonymization images produced by SDN. Finally, the most crucial observation from Table 2 is SDN's superior behavior in PieAPP. PieAPP measures the distance between two distributions of authentic and generated images, with lower values indicating better performance. SDN's excellence in PieAPP measures suggests that it produces high-quality de-anonymization images. It is worth noting that although SDN achieves a relatively good SSIM score compared to other works in the re-ID process, it only reaches the value of 0.872. The reason for this could be the loss of high-frequency information, leading to a more significant drop in SSIM than its de-ID counterpart.

The following experiments focus on the effects of generated faces using identity- and style-related attributes and mixing both attributes.

Figure 7 presents the results of using password and the style-related attribute hair color as the tested multi-attribute combinations in SDN's anonymization and de-anonymization processes. The first column shows the input images, and the second to fourth columns offer the resulting surrogates corresponding to different passwords with varying bit patterns. Finally, the fifth to the seventh columns show the de-ID images obtained using the multi-attribute combinations that jointly consider the password and the hair color. The results indicate that the proposed system generates high-quality surrogates that preserve identity-related attributes while concealing identity-related information. Moreover, the system can effectively use multi-attribute combinations to generate reasonable de-ID images,

similar to the original input images, but without revealing sensitive information. The excellent de-anonymized picture quality, shown in the eighth column of Figure 7, provides strong evidence of the effectiveness of SDN in generating high-quality surrogates for anonymization and de-anonymization tasks.

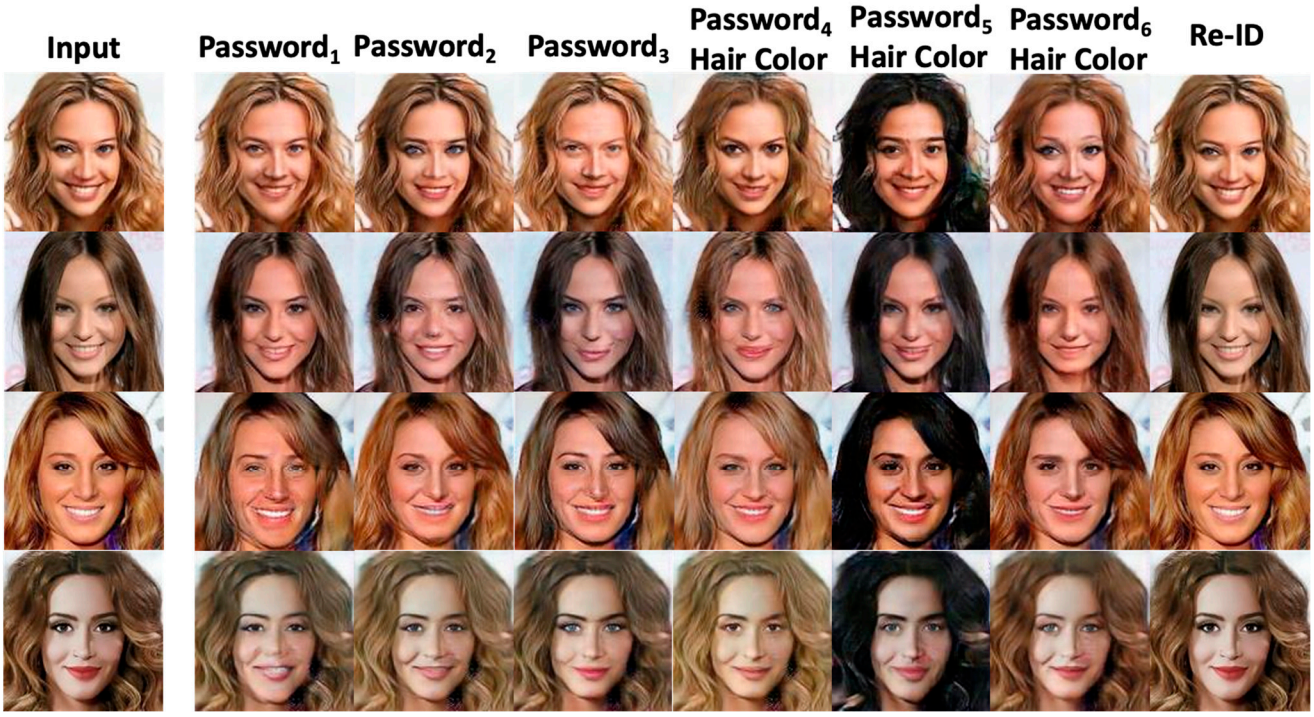


Figure 7. The SDN generated surrogate faces' snapshots using the style-related attribute Hair color and the password as the evaluated multi-attribute combinations.

The above processes of generating agent faces using different passwords and attribute-based features verified the effectiveness of the basic functionality of the designed SDN. Next, we visually compared target face pollution situations in the synthesized images between competing systems, including CIAGAN [43], Cao et al.'s work [44], MfM [25], and SDN. As shown in Figure 8, CIAGAN exhibits gaze deviation in the agent's face and significant synthetic seams when trying to fit the agent's face to the target face. In the comparison between SDN and Cao et al.'s work, we found that Cao et al.'s method produces style-related pollution, with the most obvious problem being the extreme unnaturalness of the Bangs synthesis, as well as a severe background color deviation. Finally, for MfM, we also observed similar issues to those of Cao et al., including style-related pollutions and background color deviation.

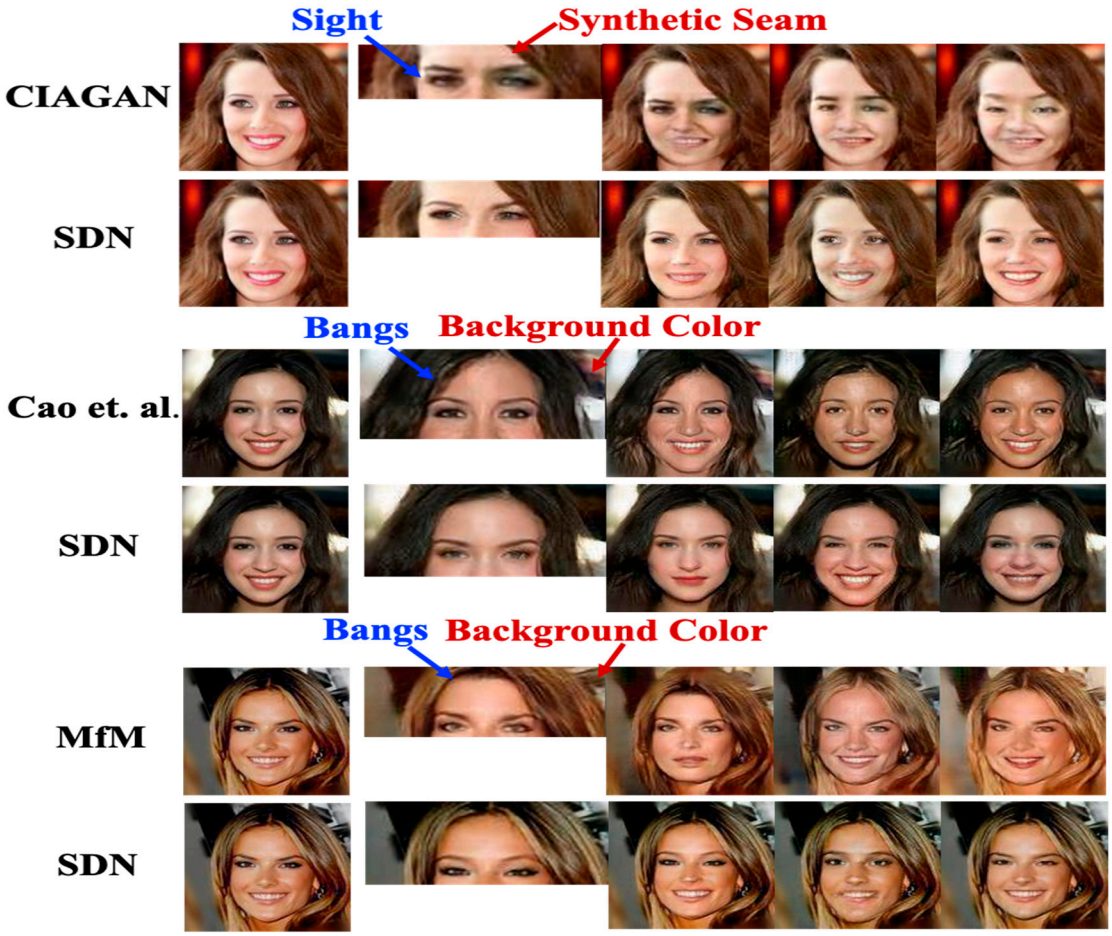


Figure 8. The visual comparison of polluted situations in the synthesized images between competing systems, including CIAGAN [43], MfM [25], Cao et al. [44], and our SDN.

In contrast, the proposed SDN completely avoids the problems above. In Table 3, we have summarized the comparison results for the four benchmarked methods. Moreover, we assessed the de-anonymization performance of the SDN model using face recognition techniques, including InsightFace, InsightFace_IR50_MS1M, and FaceNet. The original images were compared with both the anonymized and de-anonymized versions. A score of 0 was given if the de-anonymized image matched the original and 1 if it did not. Figure 9 shows that the distances between the anonymized and input images are far enough by inputting them into a publicly available facial recognition tool and successfully passing the validation test. These facts indicate that the quality of the anonymized images is good enough for privacy protection, and SDN can effectively preserve the privacy of the individuals while still maintaining the utility of the data.

Table 3. This table tabulates the Visualizable shortages (indicated by solid circles) of polluted situations among benchmarked works associated with Figure 8.

Issues / Method	Bangs	Sight	Synthetic Seam	Background Color
CIAGAN [43]	Non-pollution Issue	Pollution Issue	Pollution Issue	Non-pollution Issue
Cao et al. [44]	Pollution Issue	Non-pollution Issue	Non-pollution Issue	Pollution Issue
MfM [25]	Pollution Issue	Non-pollution Issue	Non-pollution Issue	Pollution Issue
Ours	Non-pollution Issue	Non-pollution Issue	Non-pollution Issue	Non-pollution Issue

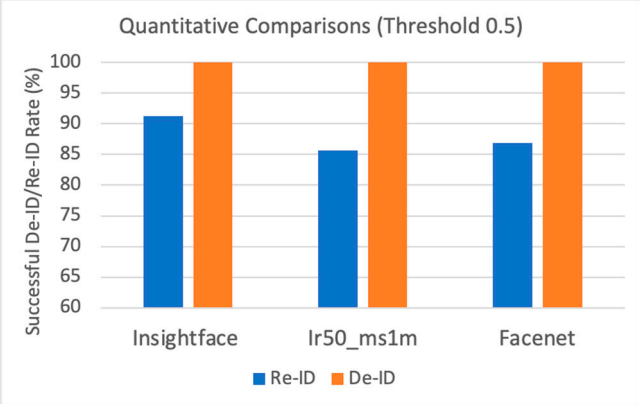


Figure 9. The successful de-ID/re-ID Rates.

Figure 10 presents the outcomes of an experiment examining similar passwords' impacts on a fixed anonymized image. In this testing, we randomly select some input images from the benchmark datasets and generate anonymized images using the smile-attribute-based features and 1-bit difference passwords as testing targets. The first column shows the input images, and the second to the seventh columns are the generated anonymized images concerning different 1-bit difference passwords associated with varying degrees of smiling. The experimental results demonstrate that SDN can effectively use different attribute combinations to generate high-quality surrogates with diverse appearances that preserve identity-related attributes while concealing sensitive information. Of course, users can choose the preferred attributes as they wish, and we will justify this claim with experiments later.

6.2. Latent Space Manipulation Analyses

We have used PCA as the compression method in the latent space for computational efficiency. This decision has freed our original approach from its constraints on working in image space, allowing SDN to work directly on NN-preferred latent space. Figure 11 depicts the block diagram of our Latent Space Manipulation Module (LSMM), and the function of LSMM is to find the shifting guidance for a target manipulation. Initially, we gather bottleneck features for a set of images produced by model G and calculate the eigenvectors of the associated correlation matrix of features. In our LSMM design, f_c represents the eigenvalues obtained from the PCA of the features. Subsequently, we employ the support vector machine (SVM) mechanism to derive the shifting guidance for specific attributes. That is,

$$f_{c_n} = PCA(En(x_n)) \; and \; v_i = SVM(f_{c_n}, label_n).$$

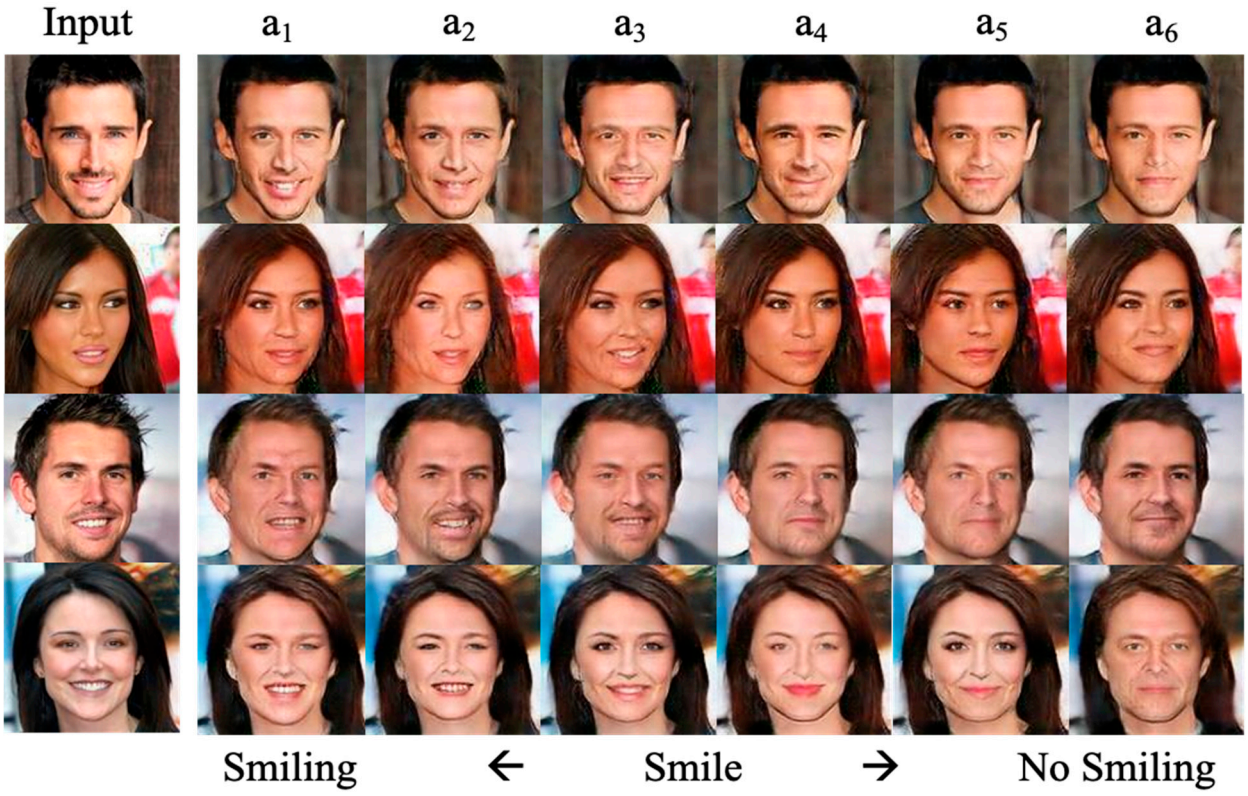


Figure 10. The above snapshots examine the impacts of 1-bit difference passwords on a given anonymized image with varying degrees of smiling attributes, which "a_1–a_6" are the abbreviations for "anonymized image 1–anonymized image 6".

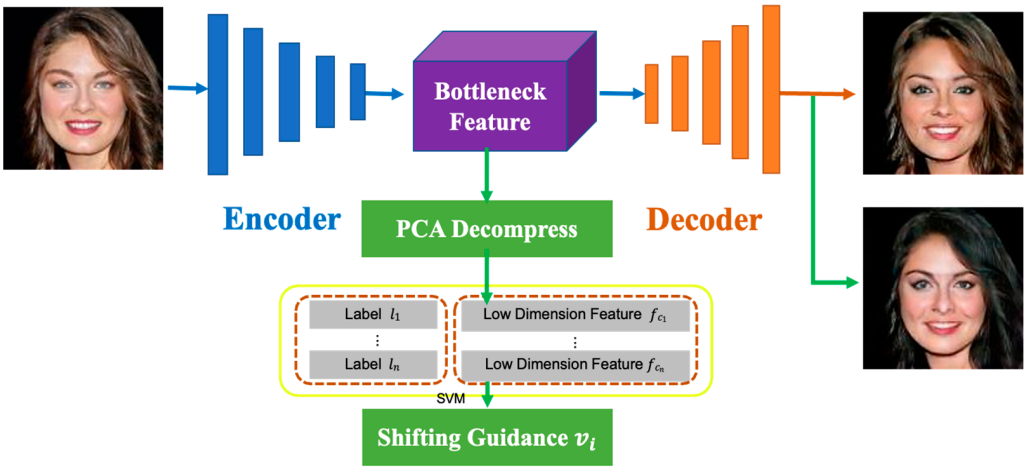


Figure 11. Block diagram of the proposed Latent Space Manipulation Module and how it is used to find the shifting guidance for a target latent space manipulation.

During the implementation phase, we begin by compressing the bottleneck feature, f. We then perform an element-wise addition of the obtained shifting guidance, v, to generate the manipulated and compressed feature, f'_n. Next, we employ inverse-PCA to decompress the manipulated and compressed feature back to the original feature space, resulting in the

feature, f'. This feature is subsequently passed to the decoder to obtain the manipulated image, i'. Mathematically, we can express the above derivations as:

$$f_c = PCA(En(x)),$$
$$f'_n = f_n + v,$$
$$f' = PCA_{inv}(f'_c), \text{ and } i' = De(f').$$

Our approach combines Yujun Shen's technique [50] for accurate single-feature control and Erik Härkönen's system design [51] to extend applicability to any latent space. However, since SVM is used for analysis, the method remains supervised, requiring labeled data. Additionally, utilizing PCA to reduce feature space dimensionality may lead to the loss of high-frequency information and blurred output images. Therefore, we applied an additional model (as shown in Figure 12) to recover the missing details from the manipulated bottleneck features to enhance the manipulation effects before sending it back to the decoder.

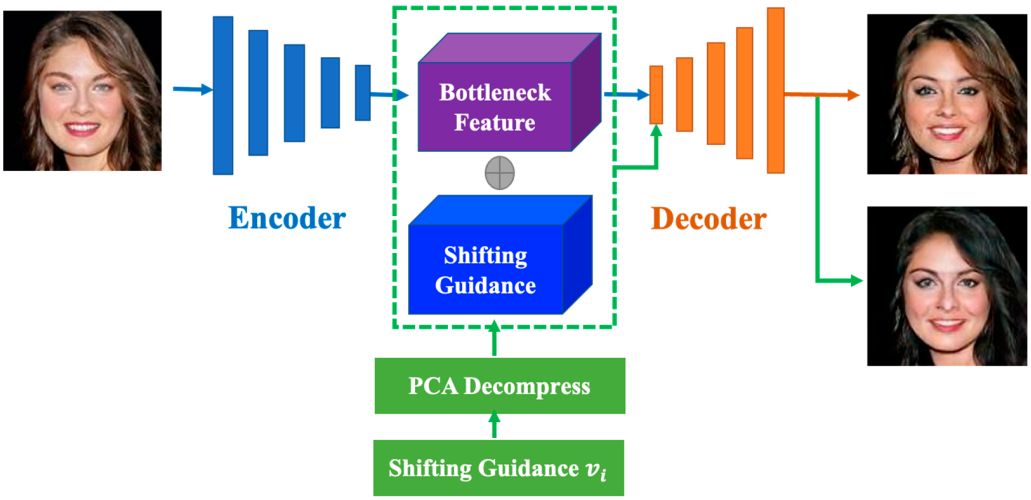


Figure 12. Block diagram of the proposed Latent Space Manipulation Module and how it is used to complete the desired manipulation by adding the shifting guidance in the latent space.

As shown in Figure 11, the space dimension of the bottleneck feature (i.e., the original latent space) is reduced to a particular range through PCA. Next, we perform SVM analysis on the dimension-reduced data to determine the shifting guidance vector associated with each attribute. Since PCA is linear, mathematically, we can perform the feature space modulations in the compressed domain isomorphically (as shown in Figure 12). In other words, we can decompress (and inverse PCA of) the offset control item first and then apply the result to directly manipulate the feature vector for achieving equivalent manipulation results. Thus, we can revise the mathematical expressions associated with Figure 11 to derive the mathematical representations associated with Figure 12 as follows:

$$f = En(x), v' = PCA_{inv}(v), \text{ and}$$
$$i' = De(f + v').$$

Finally, we use SDN as a pre-trained model for finding bottleneck features suitable for image manipulations in the latent space. The model also converts (or encodes) the input images into the so-called intermediate attributes. Then, we element-wisely embed the style-related attributes into the de-ID images following the guidance of vi generated based on the given manipulation conditions, such as from no smile to smile, from no mustache to mustache, and from mouth close to mouth open, as shown in Figure 13. Similarly, in Figure 14, we incorporate the identity-related attribute (female to male) into de-ID images based on the given manipulation condition and the calculated guidance in the latent space.

The snapshots depicted in Figures 13 and 14 show that the SDN as an Encoder will be more likely to preserve the high-frequency components of a facial image.

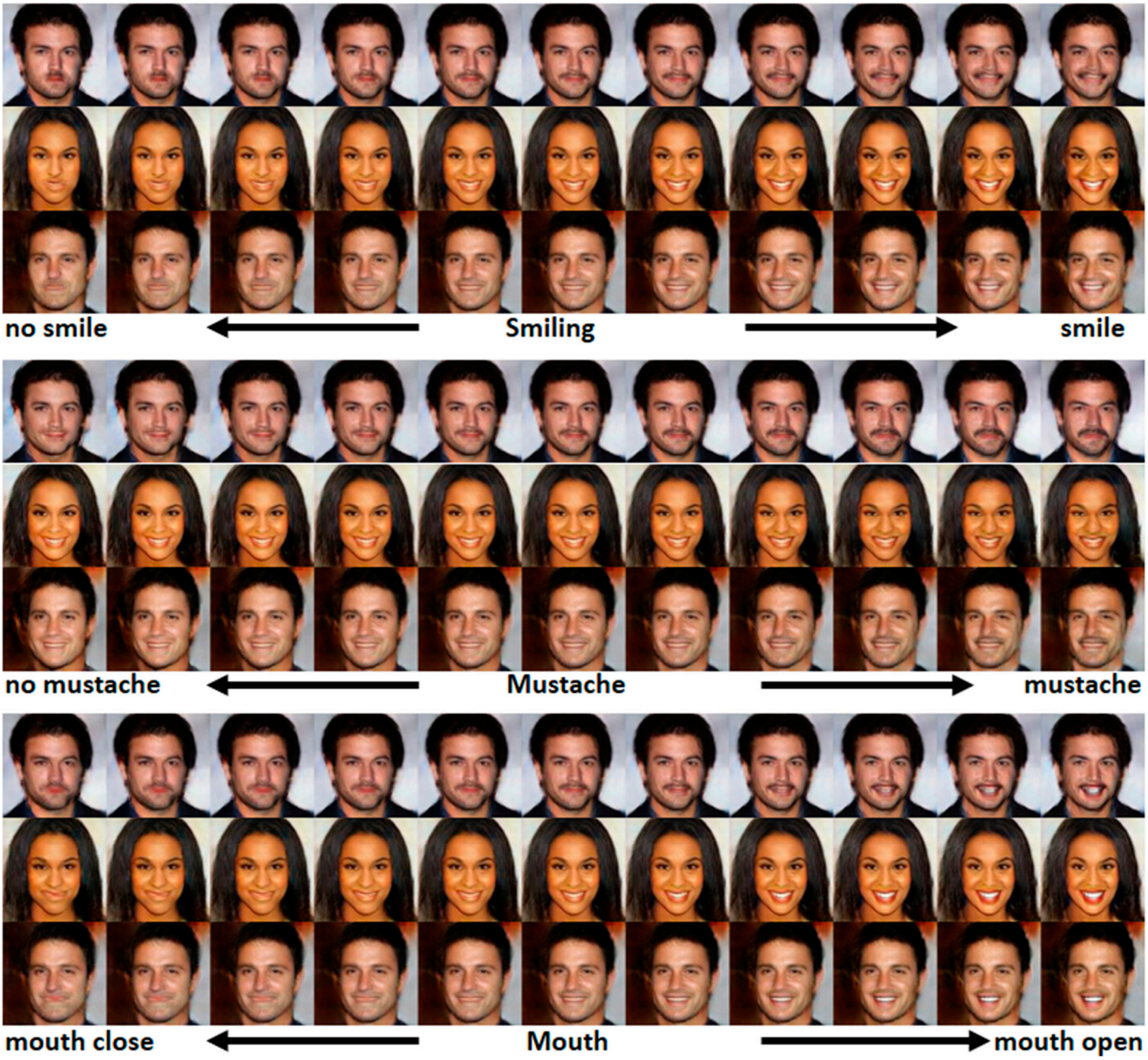


Figure 13. The snapshots of the SDN manipulated results concerning varying degrees of style-related attributes (upper: Smiling, middle: Mustache, and bottom: Mouth Open).

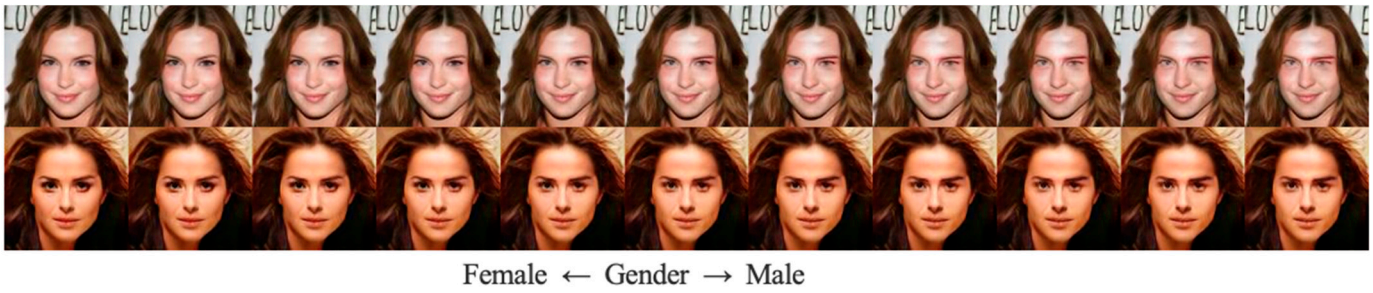


Figure 14. The snapshots of the SDN manipulated results concerning varying degrees of the identity-face-related attributes, and we take gender as a testing target.

Table 4 shows FID, SSIM, and LPIPS scores of the snapshots of the SDN manipulated results concerning varying degrees of style-related attributes. The degree stands for the strength we used to manipulate the origin bottleneck feature in Table 4, while the chosen attribute is the attribute we try to control. Take the row associated with the Mustache attribute as an example. When the degree is −2, it implies that the output results should be closer to no mustache. Conversely, when the degree is +2, the output results should be more immediate to more mustache. As we can see, the distance between the original and the output images increases when the manipulative degree is enlarged.

Table 4. FID, SSIM, and LPIPS scores of the snapshots of the SDN manipulated results concerning varying degrees of style-related attributes.

		Weaken Feature by 2 Degrees	Weaken Feature by 1 Degree	No Enhance/Weaken Features	Enhance Feature by 1 Degree	Enhance Feature by 2 Degrees
FID	Smiling	0.25	8.94	0	11.68	14.08
	Mouth Open	13.73	11.23	0	8.98	9.54
	Mustache	9.75	9.47	0	10.66	12.68
SSIM	Smiling	0.93	0.97	1	0.97	0.93
	Mouth Open	0.94	0.97	1	0.97	0.93
	Mustache	0.93	0.97	1	0.97	0.93
LPIPS	Smiling	0.0155	0.0063	0	0.00646	0.0158
	Mouth Open	0.0157	0.0069	0	0.0082	0.0200
	Mustache	0.0171	0.0067	0	0.0068	0.0176

7. The Ablation Study

Target Face Pollution: We use the same input for target face pollution inspection in all four methods when activating the de-ID process. The agent's face in CIAGAN displays gaze deviation and noticeable synthetic seams if matching the target face. When comparing SDN with Cao et al.'s approach, we discovered that Cao et al.'s technique introduces style-related distortions, particularly in the synthesis of Bangs, resulting in a highly unnatural appearance and a significant deviation in the background colors. Similarly, as in Cao et al.'s approach, we observed similar problems in MfM, such as style-related distortions and background color deviation. In contrast, as expected and shown in Figure 15, SDN avoids all the above-mentioned issues.

Time Comparison of the Evaluation of Dual Inference: This experiment involves a comparison of MfM, Cao et al., Gu et al., and SDN across three datasets—FaceScrub [4], CASIA-WebFace [5], and CelebA-HQ/CelebA [6]. Gu et al. and MfM require starting the testing process from pre-trained models, resulting in a longer total execution time than Cao et al.'s works and SDN. However, statistical analyses of each dataset indicate that the proposed SDN can train the entire mechanism in a shorter time, as shown in Figure 16a–c.

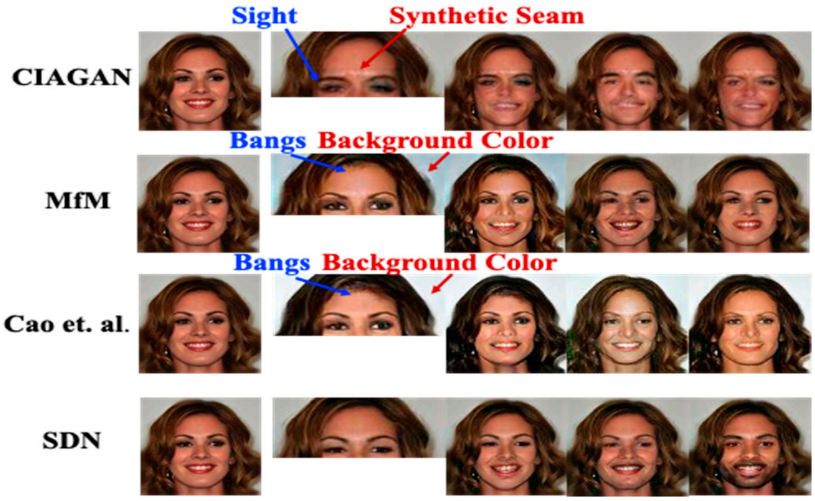
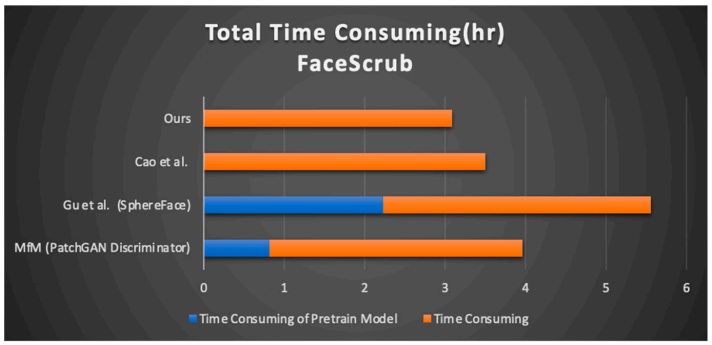
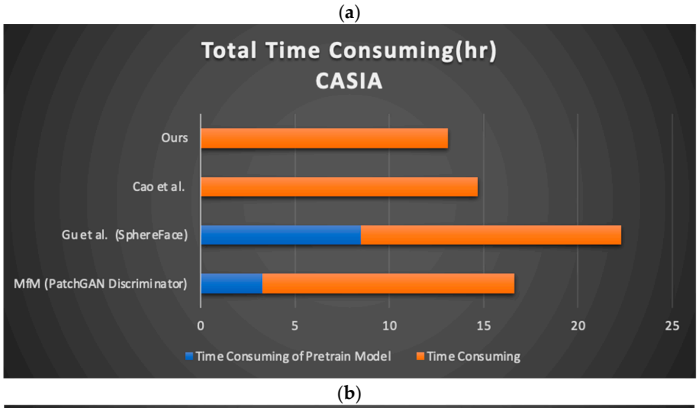


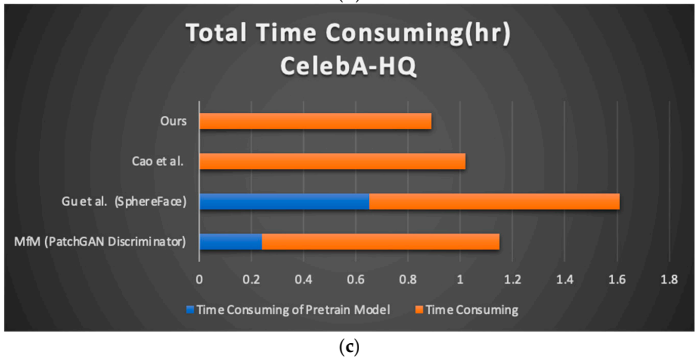
Figure 15. Using the same input image, the visual comparison of polluted situations in the synthesized images among competing systems, including CIAGAN, MfM, and SDN.



(**a**)



(**b**)



(**c**)

Figure 16. Total time consumption across three different datasets: (**a**) FaceScrub [4], (**b**) CASIA [5], and (**c**) CelebA-HQ [6].

Model-related Costs Comparison of Dual Inference (in terms of Multiply–accumulate Operations: MACs and parameter size): This experiment involves comparing MfM, Cao et al., Gu et al., and SDN with the THOP toolkit [52]. As shown in Table 5, SDN needs a parameter size cost that is only slightly less than that of Cao et al., but it achieves the best performance in terms of FLOPs, with only 13.09 G. This result shows that the SDN is more efficient in training.

Table 5. Model-related costs comparison of dual inference (MACs) among benchmarked works.

	Parameters	**MACs**
Ours	269.02 M	13.09 G
Cao et al. [44]	291.02 M	16.66 G
Gu et al. [24]	50.92 M	19.58 G
MfM [25]	189.02 M	13.89 G

8. Conclusions

This paper presents a novel privacy protection system, the SDN, that can automatically and stably anonymize and de-anonymize facial images using one neural network based on DosGAN. Unlike most existing works, SDN provides strong security protection under various attribute conditions. In more detail, SDN allows users to input their preferred face attributes and passwords to form a specific multi-attribute combination. The proposed 'dual inference' mechanism ensures precise facial de-ID and re-ID tasks are performed, as demonstrated and justified by a series of experiments. Additionally, SDN can create photo-realistic surrogate faces that satisfy specified additional conditions and de-anonymize face images using a user-defined multi-attribute combination without altering the data distribution in many cases.

The experimental performance of SDN in the anonymization task, as compared to DeepPrivacy [18], Gu et al. [24], MfM [25], CIAGAN [43], and Cao et al. [44], is illustrated in Figure 5 (ID-distance) and Figure 6 (SPR). These figures demonstrate SDN's effectiveness in identity protection, offering a larger ID distance and higher SPR than other methods. Additionally, unlike other benchmarked approaches, SDN achieves this anonymization capability without requiring any pre-trained or auxiliary models.

Tables 1 and 2 provide qualitative evaluations of SDN's anonymization and de-anonymization performance in utility-related experiments. In the anonymization task, SDN outperforms or matches other studies in LPIPS, FID, and SSIM. For de-anonymization, SDN achieves the best results in LPIPS, SSIM, and PieAPP, with acceptable performance in PSNR as well.

Additionally, the SDN's superior performances are attributed to the newly designed MI-based cost functions. In addition to conducting many experiments to justify our claims, this paper also provides detailed mathematical derivations based on information theory to explain the design insights. While MI is a helpful information-theoretic quantity, incorporating other physically meaningful measures, such as rate-distortion measures and channel capacity, could further enhance the effectiveness and efficiency of a de-ID and re-ID system. These related explorations are, of course, one of our future research directions.

Finally, we summarize several possible contributions of the proposed SDN in the following:

1. Introducing an NN-based privacy protection solution that is both reversible and controllable, allowing for facial images to be anonymized and de-anonymized as needed;
2. Using dual inference theory to ensure better realism of the de-ID image without any pre-trained/auxiliary model to enhance its applicability in practice;
3. Providing techniques to handle unseen images (which need to be de-ID'd and have never appeared in the training dataset) during the inference process;
4. Enforcing the protective function of the agent face generator to output different anonymized facial identities associated with different passwords;

5. Achieving maximum feature distance between an anonymized face and its de-anonymized version, even when multi-attribute combinations are incorrect;
6. Based on information theory, we analyze the physical meanings of the cost functions used in our development.

Author Contributions: Formal analysis, Y.-L.P.; funding acquisition, J.-L.W.; investigation, Y.-L.P., J.-C.C. and J.-L.W.; methodology, Y.-L.P.; project administration, J.-L.W.; resources, J.-L.W.; algorithm development, Y.-L.P.; supervision, J.-L.W.; writing—original draft, Y.-L.P.; writing—review and editing, J.-L.W. and J.-C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Minister of Science and Technology, Taiwan MOST 111-2221-E-002-134-MY3.

Data Availability Statement: The datasets used in this study are publicly available and freely reusable. FaceScrub dataset: https://vintage.winklerbros.net/facescrub.html; CASIA-WebFace dataset: https://paperswithcode.com/dataset/casia-webface; CelebA-HQ/CelebA dataset: https://paperswithcode.com/dataset/celeba-hq. All accessed on 2 October 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fabrègue, B.F.G.; Bogoni, A. Privacy and Security Concerns in the Smart City. *Smart Cities* **2023**, *6*, 586–613. [CrossRef]
2. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance) (OJ L 119 04.05.2016). p. 1. Available online: http://data.europa.eu/eli/reg/2016/679/oj (accessed on 18 May 2018).
3. Morić, Z.; Dakic, V.; Djekic, D.; Regvart, D. Protection of Personal Data in the Context of E-Commerce. *J. Cybersecur. Priv.* **2024**, *4*, 731–761. [CrossRef]
4. Ng, H.W.; Winkler, S. A Data-driven Approach to Cleaning Large Face Datasets. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 343–347.
5. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Learning Face Representation from Scratch. *arXiv* **2014**, arXiv:1411.7923.
6. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
7. Ren, Z.; Lee, Y.J.; Ryoo, M.S. Learning to anonymize faces for privacy-preserving action detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 620–636.
8. Boyle, M.; Edwards, C.; Greenberg, S. The Effects of Filtered Video on Awareness and Privacy. In Proceedings of the ACM Conference on Computer Supported Cooperative Work, Philadelphia, PA, USA, 2–6 December 2000; pp. 1–10.
9. Neustaedter, C.; Greenberg, S.; Boyle, M. Blur Filtration Fails to Preserve Privacy for Home-Based Video Conferencing. *ACM Trans. Comput. Hum. Interact.* **2006**, *13*, 1–36. [CrossRef]
10. Phillips, P. Privacy Operating Characteristic for Privacy Protection in Surveillance Applications. In *Audio- and Video-Based Biometric Person Authentication*; Kanade, T., Jain, A., Ratha, N., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2005; pp. 869–878.
11. Seo, J.; Hwang, S.; Suh, Y.-H. A Reversible Face De-Identification Method based on Robust Hashing. In Proceedings of the 2008 International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 9–13 January 2008. [CrossRef]
12. Gross, R.; Sweeney, L.; Cohn, J.; de la Torre, F.; Baker, S. Face De-identification. In *Protecting Privacy in Video Surveillance*; Senior, A., Ed.; Springer: London, UK, 2009. [CrossRef]
13. Padilla-López, J.R.; Chaaraoui, A.A.; Flórez-Revuelta, F. Visual Privacy Protection Methods: A Survey. *Expert Syst. Appl.* **2015**, *42*, 4177–4195. [CrossRef]
14. Sweeney, L. K-anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2002**, *10*, 557–570. [CrossRef]
15. Meden, B.; Emeršič, Ž.; Štruc, V.; Peer, P. K-same-net: K-anonymity with Generative Deep Neural Networks for Face Deidentification. *Entropy* **2018**, *20*, 60. [CrossRef] [PubMed]
16. Newton, E.M.; Sweeney, L.; Malin, B. Preserving privacy by de-identifying face images. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 232–243. [CrossRef]
17. Jourabloo, A.; Yin, X.; Liu, X. Attribute Preserved Face De-identification. In Proceedings of the 2015 International Conference on Biometrics, ICB 2015, Phuket, Thailand, 19–22 May 2015; pp. 278–285. [CrossRef]
18. Hukkelås, H.; Mester, R.; Lindseth, F. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. In *Advances in Visual Computing. ISVC 2019*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; Volume 11844. [CrossRef]
19. Pan, Y.-L.; Haung, M.-J.; Ding, K.-T.; Wu, J.-L.; Jang, J.-S.R. K-Same-Siamese-GAN: K-Same Algorithm with Generative Adversarial Network for Facial Image De-identification with Hyperparameter Tuning and Mixed Precision Training. In Proceedings

of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, 18–21 September 2019; pp. 1–8.
20. Jeong, Y.; Choi, J.; Kim, S.; Ro, Y.; Oh, T.-H.; Kim, D.; Ha, H.; Yoon, S. FICGAN: Facial Identity Controllable GAN for De-identification. *arXiv* **2021**.
21. Zhao, Z.; Xia, Y.; Qin, T.; Xia, L.; Liu, T.-Y. Dual Learning: Theoretical Study and an Algorithmic Extension. *SN Comput. Sci.* **2021**, *2*, 413. [CrossRef]
22. Yamac, M.; Ahishali, M.; Passalis, N.; Raitoharju, J.; Sankur, B.; Gabbouj, M. Reversible Privacy Preservation using Multi-level Encryption and Compressive Sensing. In Proceedings of the 27th European Signal Processing Conference, A Coruna, Spain, 2–6 September 2019. [CrossRef]
23. Li, Y.; Chen, S.; Qi, G.; Zhu, Z.; Haner, M.; Cai, R. A GAN-Based Self-Training Framework for Unsupervised Domain Adaptive Person Re-Identification. *J. Imaging* **2021**, *7*, 62. [CrossRef]
24. Gu, X.; Luo, W.; Ryooand, M.; Lee, Y. Password-conditioned Anonymization and Deanonymization with Face Identity Transformers. *arXiv* **2020**, arXiv:1911.11759.
25. Pan, Y.-L.; Chen, J.-C.; Wu, J.-L. A Multi-Factor Combinations Enhanced Reversible Privacy Protection System for Facial Images. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6. [CrossRef]
26. Xu, S.; Chang, C.; Nguyen, H.H.; Echizen, I. Reversible anonymization for privacy of facial biometrics via cyclic learning. *EURASIP J. Inf. Secur.* **2024**, *2024*, 24. [CrossRef]
27. Wen, Y.; Liu, B.; Cao, J.; Xie, R.; Song, L. Divide and Conquer: A Two-Step Method for High-Quality Face De-identification with Model Explainability. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 5125–5134. [CrossRef]
28. Bao, J.; Chen, D.; Wen, F.; Li, H.; Hua, G. Towards Open-set Identity Preserving Face Synthesis. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6713–6722.
29. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
30. Boesen, A.; Larsen, L.; Ren, S.; Snderby, K.; Larochelle, H.; Winther, O. Autoencoding beyond pixels using a learned similarity metric. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA; 2016; Volume 48, pp. 1558–1566. Available online: https://proceedings.mlr.press/v48/larsen16.html (accessed on 28 September 2024).
31. Shen, W.; Liu, R. Learning Residual Images for Face Attribute Manipulation. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4030–4038.
32. Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; Choo, J. Star-GAN: Unified generative adversarial networks for multi-domain Image-to-Image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.
33. Choi, Y.; Uh, J.; Ha, J. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8188–8197.
34. Lin, J.; Chen, Z.; Xia, Y.; Liu, S.; Qin, T.; Luo, J. Exploring Explicit Domain Supervision for Latent Space Disentanglement in Unpaired Image-to-Image Translation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1254–1266. [CrossRef] [PubMed]
35. Wang, L.; Chen, W.; Yang, W.; Bi, F.; Yu, F. A State-of-the-Art Review on Image Synthesis with Generative Adversarial Networks. *IEEE Access* **2020**, *8*, 63514–63537. [CrossRef]
36. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-image Translation with Conditional Adversarial Networks. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
37. Shen, Y.; Zhou, B. Closed-Form Factorization of Latent Semantics in GANs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 21–25 June 2021; pp. 1532–1540.
38. Barber, D.; Agakov, F.V. The IM Algorithm: A Variational Approach to Information Maximization. *Adv. Neural Inf. Process. Syst.* **2003**, *16*, 201–208.
39. Cheng, P.; Hao, W.; Dai, S.; Liu, J.; Gan, Z.; Carin, L. CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 1779–1788.
40. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
41. Omkar, M. Parkhi, Andrea Vedaldi and Andrew Zisserman. Deep Face Recognition. In Proceedings of the British Machine Vision Conference (BMVC), Swansea, UK, 7–10 September 2015; Xie, X., Jones, M.W., Tam, G.K.L., Eds.; BMVA Press: Durham, UK, 2015; pp. 41.1–41.12.
42. Prashnani, E.; Cai, H.; Mostofi, Y.; Sen, P. PieAPP: Perceptual Image-Error Assessment Through Pairwise Preference. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
43. Maximov, M.; Elezi, I.; Leal-Taixe, L. Ciagan: Conditional identity anonymization generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5447–5456.

44. Cao, J.; Liu, B.; Wen, Y.; Xie, R.; Song, L. Personalized and Invertible Face De-identification by Disentangled Identity Information Manipulation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3314–3322. [CrossRef]
45. Im, D.H.; Seo, Y.S. FaceBERT: Face De-Identification Using VQGAN and BERT. In Proceedings of the 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 19–21 October 2022; pp. 2013–2015. [CrossRef]
46. Yang, J.; Zhang, W.; Liu, J.; Wu, J.; Yang, J. Generating De-identification facial images based on the attention models and adversarial examples. *Alex. Eng. J.* **2022**, *61*, 8417–8429. [CrossRef]
47. Zhai, L.; Guo, Q.; Xie, X.; Ma, L.; Wang, Y.E.; Liu, Y. A^3GAN: Attribute-Aware Anonymization Networks for Face De-identification. In Proceedings of the 30th ACM International Conference on Multimedia (MM'22), New York, NY, USA, 10–14 October 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 5303–5313. [CrossRef]
48. Seyyed, K.; Shirin, N. StyleGAN as a Utility-Preserving Face De-identification Method. *arXiv* **2022**, arXiv:2212.02611.
49. Xue, H.; Liu, B.; Yuan, X.; Ding, M.; Zhu, T. Face image de-identification by feature space adversarial perturbation. *Concurr. Comput. Pract. Exp.* **2022**, *35*, e7554. [CrossRef]
50. Shen, Y.; Gu, J.; Tang, X.; Zhou, B. Interpreting the latent space of gans for semantic face editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020.
51. Harkonen, E.; Hertzmann, A.; Lehtinen, J.; Paris, S. Ganspace: Discovering interpretable gan controls. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9841–9850.
52. Available online: https://pypi.org/project/thop/ (accessed on 12 May 2024).