Information Bottleneck Disentanglement for Identity Swapping

Gege Gao Huaibo Huang Chaoyou Fu Zhaoyang Li Ran He*
National Laboratory of Pattern Recognition, CASIA
Center for Excellence in Brain Science and Intelligence Technology, CAS
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{gege.gao,huaibo.huang}@cripac.ia.ac.cn {chaoyou.fu,rhe}@nlpr.ia.ac.cn zhaoyang0427@gmail.com

Abstract

Improving the performance of face forgery detectors often requires more identity-swapped images of higher-quality. One core objective of identity swapping is to generate identity-discriminative faces that are distinct from the target while identical to the source. To this end, properly disentangling identity and identity-irrelevant information is critical and remains a challenging endeavor. In this work, we propose a novel information disentangling and swapping network, called InfoSwap, to extract the most expressive information for identity representation from a pre-trained face recognition model. The key insight of our method is to formulate the learning of disentangled representations as optimizing an information bottleneck trade-off, in terms of finding an optimal compression of the pre-trained latent features. Moreover, a novel identity contrastive loss is proposed for further disentanglement by requiring a proper distance between the generated identity and the target. While the most prior works have focused on using various loss functions to implicitly guide the learning of representations, we demonstrate that our model can provide explicit supervision for learning disentangled representations, achieving impressive performance in generating more identity-discriminative swapped faces.

1. Introduction

Face forgery detection aims to identify whether a given facial image has been modified, and is currently dominated by data-driven approaches [35, 31, 28, 27]. This means that it is difficult to improve the performance of forgery detectors in the absence of high-quality Deepfake data. Therefore, better face-swapping methods are in dire need to help develop powerful forgery algorithms.

Recent works have made significant contributions in this regard. FaceSwap [25] enables face swapping in real-time. RSGAN [30] and FSNet [36] introduce GAN-based methods in synthesizing swapped face. FSGAN [32] proposes a subject-agnostic approach for both face swapping and reenactment. More recently, FaceShifter [26] puts its focus on the occlusion problem and achieves high fidelity.

The core objective of identity swap (*i.e*. face swap) is to keep the identity of the swapped face the same as the source face while sharing the identity-irrelevant perceptual information (*e.g*. pose, expression, and illumination) with the target face. Therefore, proper disentanglement is an essential premise for well representing the identity and the perceptual information. Otherwise, entangled target perception will inevitably bring the target identity into the synthesis process, leading to an identity-mixed result. Despite such importance, it is yet to see a breakthrough for disentangled representation on face swapping. Previous works [25, 30, 4, 19, 32, 26] attempt to constrain the identity and perception of the generated faces by adding multiple loss terms to the objective function. However, due to the lack of explicit supervision, it is still challenging for these works to learn well-disentangled representations.

In this paper, we focus on improving disentangled representation learning in subject-agnostic face swap. The main idea is to learn the minimal sufficient statistics, namely the optimal representations, for both the identity and identity-irrelevant perceptual information from the latent features of a pre-trained model [10]. By introducing the information-theoretic principles [44, 40, 2], we model this learning process as a problem of optimizing the Information Bottleneck (IB) trade-off, performed by a novel information disentanglement network *InfoSwap*. Based on the IB principle, we can provide explicit supervision for disentangled representation learning. Moreover, we improve the IB objectives to further facilitate the representation disentangling. Driven by the intuition that proper swapped faces should be not only close to the source in identity but also distinct from the target, we provide a clear definition for discriminative identities and extent the original IB objectives with a novel *Identity Contrastive Loss* (ICL) as an additional regularization on the generated identities.

Extensive experiments show that our method can better disentangle information and generate more identity-discriminative swapped faces with higher fidelity. A com-

*Corresponding Author

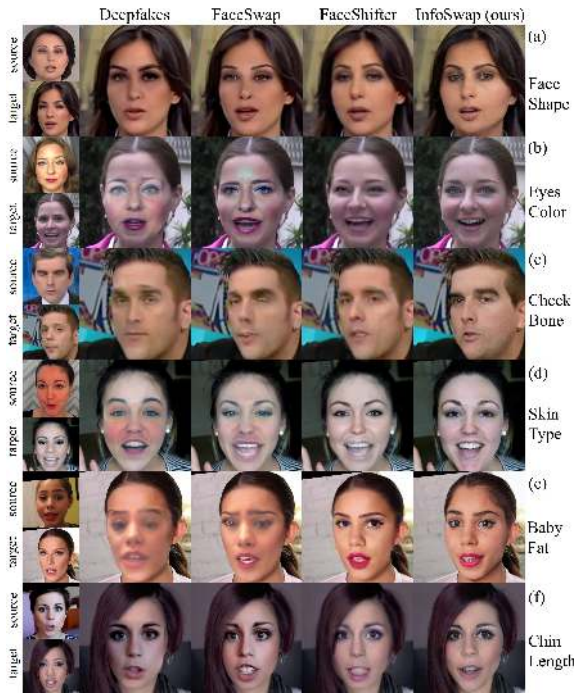

Figure 1. Challenging conditions for generating identity-discriminative faces. More details please refer to Section 4.

parison with state-of-the-art methods is shown in Fig. 1. For example, the face shape of our generated results is closer to the source rather than the target. The empirical outperformance imply that our model can provide data-driven detection algorithms with more realistic Deepfake data to improve their performance. The main contributions of this paper are the following:

- We adopt the IB principle for disentangled representation to extract the minimal sufficient identity and perceptual information. The IB principle provides a guarantee that in the latent space, areas scored identity-irrelevant indeed contribute little information to predict identity, thus enabling explicit supervision for disentanglement.
- We extent the IB objectives with a novel identity contrastive loss to further facilitate the disentanglement by requiring the generated identities to keep proper distances from the targets.
- We provide a novel metric to evaluate if the generated identity is discriminative based on its statistical features.
- Experimental results show that our method is robust and can produce more identity-discriminative swapped faces with high-fidelity.

2. Related Works

2.1. Identity Swapping

The research on face swap starts with the influential work [6]. Yet it requires human interaction and cannot preserve the target expression. Early efforts [5, 47, 8, 29] are mainly based on the 3D method. Face2Face [43] addresses the limitation of expression transfer by fitting a 3D morphable model (3DMM) [7] to both the source and target face. Nirkin *et al*. [33] proposes a 3D-based face segmentation method for seamless face transfer. Neural Textures [42] enables the synthesis of photo-realistic images with noisy and incomplete 3D geometry. Besides, learning-based methods have enabled great progress in face swap. FaceSwap [25] enables real-time subject-aware face-swapping by building image-to-image translation models case-by-case. RSGAN [30] swaps the face by learning representations of the face and the other area separately. FSGAN [32] achieves both face swap and face reenactment in a subject-agnostic pipeline. FaceShifter [26] proposes to tackle the occlusion problem via a secondary residual learning network. While there are many works on feature disentanglement, [45, 17, 16, 48, 13, 14] are for classification tasks and [50, 37, 12] are not focused on disentangling identity, for example. [11] proposes a reference-based generation for face rotation and manipulation, which can be transferred to face swap. Little attention has been paid to generating better disentangled and highly discriminative identities, which is the focus of this work. Recently, some important advances have been made in face forgery detection. For example, Faceforensics++ [35] provides an automated benchmark as well as a large database of manipulated images for building stronger detection algorithms. However, improving the accuracy of data-driven forgery detectors requires more high-quality face-swapped data.

2.2. Information Bottleneck

The idea of viewing Deep Neural Network (DNNs) in the plane of the Mutual Information is first pointed out in [44], suggesting that the goal of DNNs is to optimize the Information Bottleneck (IB) trade-off between the compression and the predictive power of the internal representations. After that, [3] proposes a variational inference to approximate the bounds on mutual information by using the reparameterization trick [24], so that it becomes easier to optimize the information bottleneck objective when applying to DNNs. More recently, [39] proposes to adopt the information bottleneck trade-off in attribution by quantifying the amount of information that an image region can provide for classification tasks. We will provide a more detailed description of the Information Bottleneck Principle and how it relates to our work in Section 3.2.

3. Method

Given two facial images, *i.e*. a source X_s and a target X_t, our proposed InfoSwap generates a face-swapped image $Y_{s,t}$ that shares the identity with X_s and the perception with X_t. An overview of the swapping process is shown in Fig. 2. InfoSwap consists of two learnable modules, an *Informative Identity Bottleneck* (IIB in **blue**) and an *Adaptive Information Integrator* (AII in **brown**), while the IIB

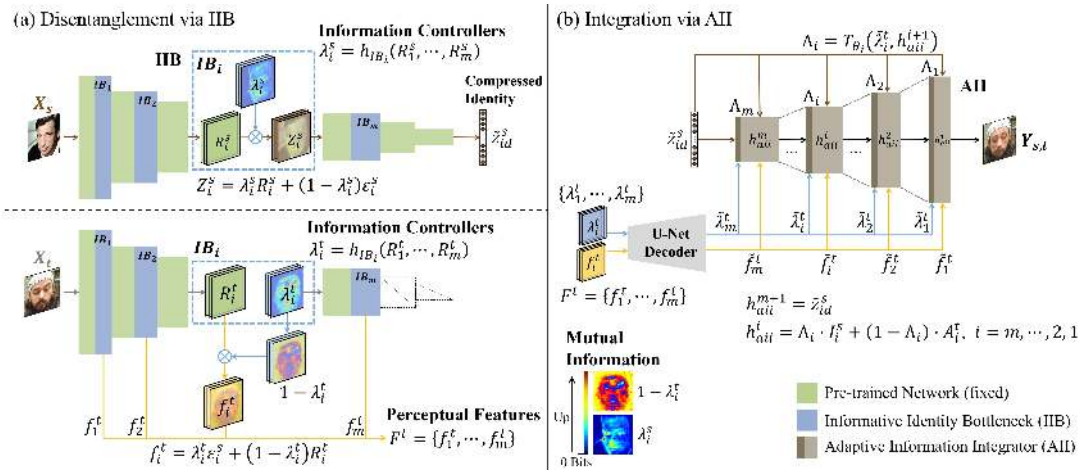

Figure 2. An overview of InfoSwap. It consists of two learnable modules, an Informative Identity Bottleneck (IIB in blue) and an Adaptive Information Integrator (AII in brown), while the pre-trained face recognition network (in green) is fixed. The IIB for both X_s and X_t are the same (share weights). Note that the controllers λ_i^t and λ_i^s differ from the heat maps. They are calculated based on the mutual information, quantifying the contribution of each feature area in bits. Please refer to Section 3 for more details.

for both X_s and X_t are the same (sharing weights).

Instead of training several new encoders [26], we learn the identity and perceptual representations in one go by quantifying and compressing the flow of information through a pre-trained face recognition model [10] in the forward pass. As shown in Fig. 2(a), given a pre-trained network (in green), denoted as a deterministic encoding function $f(\cdot)$, we extract its 512-dimensional feature embedding $z_{id} = f(X)$ and internal features $R = \{R_1, R_2, \cdots, R_m\}$ of the first m intermediate layers to represent the identity and perceptual information respectively. We compress the internal features $\{R_i\}$ by adding noise to them via information controllers $\{\lambda_i^s\}$ and $\{\lambda_i^t\}$, to find an optimal disentanglement between identity and perceptual information. The controllers are predicted by the information bottlenecks in IIB. Then, in Fig. 2(b), the compressed identity and perceptual features are sent into AII and integrated based on the information controllers, outputting the final swapped face $Y_{s,t}$. During training, the pre-trained network is fixed.

In this section, we illustrate intuitions behind InfoSwap and the key designs for learning disentangled representations. We start by defining the discriminative identities for swapped faces.

3.1. Discriminative Identities in Face Swapping

The fake identities of swapped faces are more discriminative if there is no overlap between the interval estimations of two identity similarities: i) the fake and source identity similarities; ii) the fake and target identity similarities. A visual explanation is shown in Fig. 3. The overlap of two intervals means that some fake identities are on the *angle bisector* between the source and target identity. Namely, these generated identities are equally similar to the source and the target, therefore are less discriminative.

Instead, a discriminative identity should be close to the source identity z_{id}^s, while appropriately away from the target identity. As shown in Fig. 3(a), the generated identity should be within a small angle centered on z_{id}^s, formally denoted as $z_{id}^s \pm k \cdot \delta_{s,t}$. Here $\mu_{s,t}$ and $\sigma_{s,t}$ are the *mean* and the *standard deviation* (std.) of the cosine similarities between source and target identities; $\delta_{s,t}$ is the *arccosine* value of $\sigma_{s,t}$. Under this circumstance, the fake identities within this small interval will keep a proper distance from target identities, with their similarities to the targets in the interval of $\mu_{s,t} \pm 3\sigma_{s,t}$ (*e.g.* $k = 3$), shown as the lightest green belt in Fig. 3(b).

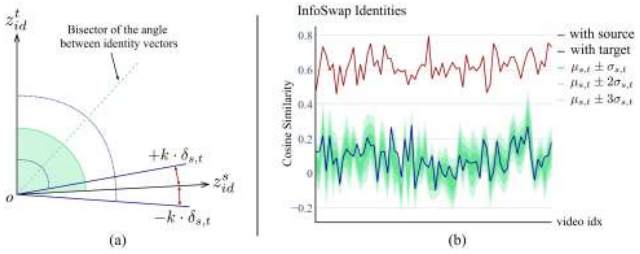

Figure 3. Explanation of the discriminative identities generated by InfoSwap. $\mu_{s,t}$ and $\sigma_{s,t}$: the *mean* and *std.* of cosine similarities between source and target identities; $\delta_{s,t} = arccos(\sigma_{s,t})$.

To make the generated identities highly discriminative, it is essential to avoid the target perception bringing the target identity information into the synthesis process. Therefore, our model aims to improve the information disentanglement and generate identity-discriminative swapped faces.

3.2. Informative Identity Bottleneck

In this subsection, we begin with illustrating the information bottleneck trade-off, and explain how it is used to design a powerful disentangling function.

3.2.1 Intuitions and IB Principle

Revisit. In the view of information theories [3], the goal of deep learning is explained as finding an optimal representation R of the input source X that: i) captures as much as possible the relevant information about the target Y, measured by the Mutual Information $I(R;Y)$, while ii) maximally compressing X by discarding the irrelevant parts

which do not contribute to the prediction of Y. This suggests the following Lagrangian objective:

$$\min_{R} I(X;R) - \beta I(R;Y). \tag{1}$$

where the positive Lagrange multiplier β operates as a trade-off parameter between the compression of the representation complexity, defined by $I(X;R)$, and the predictive power measured by the amount of relevant information in R, defined by $I(R;Y)$.

For face swap. The goal is to learn the most expressive representations R about the identity while maximally compressing the identity-irrelevant perceptual information in it. Generally, we regard such a learning process as a problem of optimizing the information bottleneck trade-off, where X represents the input facial image and Y the ground-truth identity z_{id} given by the pre-trained model. According to Eq. (1), we can optimize this trade-off by minimizing the following objective function:

$$L_{IB} = L_{info} + \beta L_{task}, \tag{2}$$

where L_{info} measures the information available in R, and β is a hyperparameter controlling the trade-off. L_{task} measures the total decline in the performance of face swapping caused by information compression, which includes not only the power of predicting identities but also the power of generating discriminative identities for the final swapped faces. Therefore, we define it as the sum of two objectives: $L_{task} = L_{recog} + L_{icl}$, which are illustrated in detail next.

3.2.2 Disentanglement with IB Objectives

The optimization process is performed in IIB. As shown in Fig. 2(a), we model a total number of m information bottlenecks, denoted as $IB = \{IB_1, IB_2, \cdots, IB_m\}$, and insert them into the first m layers of the pre-trained encoding network. To disentangle the identity and perceptual information in each internal feature R_i, each bottleneck IB_i is designed to predict an *information-controller* λ_i using all m internal features $R_1, R_2, \cdots, R_m$. We define each bottleneck IB_i as a predicting function $h_{IB_i}(*)$, then:

$$\lambda_i = h_{IB_i}(R_1, \cdots, R_m) \in [0, 1], \tag{3}$$

where λ_i is of the same size as R_i. We then compress the information in R_i by adding noise [24, 41, 23] to it. In specific, we apply a linear interpolation between R_i and a Gaussian noise ε_i based on λ_i, and the compressed version of R_i is formulated as:

$$Z_i = \lambda_i R_i + (1 - \lambda_i)\varepsilon_i, \tag{4}$$

where the noise $\varepsilon_i \sim \mathcal{N}(\mu_{R_i}, \sigma^2_{R_i})$ is set to be of the same mean and variance with R_i since the face recognition model is already trained and fixed. Thus $Z_i \sim \mathcal{N}(\mu_{R_i}, \sigma^2_{R_i})$ follows the same distribution as R_i.

Note that λ_i controls the replacement of identity-irrelevant activations with the noise. In areas where $\lambda_i = 1$, then $Z_i = R_i$ and all information is preserved in R_i. Whereas in areas where $\lambda_i = 0$, $Z_i = \varepsilon_i$, all information is damped and replaced by noise. This means that these areas can contribute 0 bits of information for predicting identities, and thus are indeed irrelevant to identity. Therefore, λ_i is different from the heat map (attention map). It is calculated based on the mutual information between feature areas and the task target, quantifying the contribution of areas in bits.

Supervision on Information Compression: L_{info}

For the first term L_{info} in Eq. (2), we use the mutual information $I(Z_i, R_i)$ to quantify the information shared between Z_i and R_i, *i.e.* the information uncompressed:

$$L_{info} = \frac{1}{m}\sum_{i=1}^{m} I(Z_i, R_i). \tag{5}$$

Given an input X, R_i is constant as the pre-trained network is fixed, thus according to Eq. (4), $Z_i|R_i \sim \mathcal{N}(\lambda_i R_i + (1-\lambda_i)\mu_{R_i}, (1-\lambda_i)^2\sigma^2_{R_i})$. Since the KL-divergence stays the same when both distributions are scaled, we normalize $Z_i|R_i$ and Z_i by using μ_{R_i} and $\sigma^2_{R_i}$, then for Gaussian distribution:

$$\begin{aligned} I(Z_i, R_i) &\triangleq KL[p(Z_i|R_i)||p(Z_i)] \\ &= -\log(1-\lambda_i)^2 \\ &+ \frac{1}{2}[(1-\lambda_i)^2 + (\lambda_i \frac{R_i - \mu_{R_i}}{\sigma_{R_i}})^2 - 1]. \end{aligned} \tag{6}$$

Therefore, when $\lambda_i = 0$ and $Z_i = \varepsilon_i$, both Z_i and $Z_i|R_i$ have the same distribution $\mathcal{N}(\mu_{R_i}, \sigma^2_{R_i})$, thus $I(Z_i, R_i) = 0$, *i.e.* contributing 0 bits information.

As for the second term L_{task} in Eq. (2), it performs supervision on both the latent space and the pixel space.

Supervision on the latent space: L_{recog}

L_{recog} is defined as the average decline of the accuracy in predicting the identities caused by the compression of the latent space. As shown in Fig. 2(a), we insert one bottleneck IB_i a time to the i-th layer and define the current network as $f_{IB_i}(\cdot)$. Then, we replace each R_i with its compressed (by IB_i alone) version Z_i, and all later layers are calculated on Z_i instead of R_i, outputting a compressed identity embedding $\tilde{z}^{(i)}_{id}(X) = f_{IB_i}(X)$. The final compressed identity is the average of all m embeddings (whole IIB), defined as:

$$\tilde{z}_{id}(X) = \frac{1}{m}\sum_{i=1}^{m} \tilde{z}^{(i)}_{id}(X) = \frac{1}{m}\sum_{i=1}^{m} f_{IB_i}(X), \tag{7}$$

Therefore, the supervision for the latent representations is formulated as:

$$L_{recog} = 1 - \cos\langle \tilde{z}_{id}, z_{id} \rangle. \tag{8}$$

where $\cos\langle \tilde{z}_{id}, z_{id} \rangle$ is the cosine distance between the compressed and the original identity.

Supervision on the pixel space via ICL

In addition to L_{recog} which focuses mainly on the latent representations, we introduce another supervision L_{icl} that explicitly requires the swapped face to be identity-discriminative in a contrastive manner.

The intuition behind contrastive learning is to teach a model to distinguish between similar and dissimilar things.

Previously, contrastive losses are usually used in face recognition as a max-margin approach for better separating positive from negative examples [9, 38, 46]. In our case, to make the swapped faces more identity-discriminative as defined in Section 3.1, we require the generated identities to be properly distanced from the target. Based on this intuition, we propose a novel Identity Contrastive Loss (ICL) which consists of a positive part L_{pos} to learn from the source identity, and a negative part L_{neg} to explore information from the distance between the source and target identities:

$$L_{icl} = L_{pos} + L_{neg}. \tag{9}$$

The positive part L_{pos} requires the identity of $Y_{s,t}$ to be close to the source in the cosine distance:

$$L_{pos} = -\cos\langle \tilde{z}_{id}(Y_{s,t}), \tilde{z}_{id}(X_s)\rangle, \tag{10}$$

where $\tilde{z}_{id}(Y_{s,t}) = \sum f_{IB_i}(Y_{s,t})/m$ is the compressed identity of $Y_{s,t}$, as defined in Eq. (7).

As for the negative part L_{neg}, instead of forcing the identity distance between $Y_{s,t}$ and the target X_t to approach a constant value 0, we use the angular margin between the source and target identities to induce a more *proper* distance between $\tilde{z}_{id}(Y_{s,t})$ and $\tilde{z}_{id}(X_t)$. As illustrated in Section 3.1, a discriminative identity should be in a small interval centered around the source identity. Namely, its proper distance from the target identity should be close to that of the source identity. Thus, we use the cosine distance between source and target identities as a better constraint for the generated identity:

$$\begin{aligned} L_{neg} = [&\cos\langle \tilde{z}_{id}(Y_{s,t}), \tilde{z}_{id}(X_t)\rangle \\ &- \cos\langle \tilde{z}_{id}(X_s), \tilde{z}_{id}(X_t)\rangle]^2. \end{aligned} \tag{11}$$

Based on this contrastive loss L_{icl}, we can provide effective supervision on the generated identity to make the swapped face more discriminative.

Therefore, the total IB objective function is:

$$L_{IB} = L_{info} + \beta(L_{recog} + L_{icl}). \tag{12}$$

By minimizing L_{IB}, the values of λ_i in areas efficiently informative about the identity will be close to 1, while in areas less relevant to identity will be compressed near 0. Thus representations can get properly disentangled based on these controllers $\lambda_1, \lambda_2, \cdots, \lambda_m$.

3.3. Adaptive Information Integration

After the disentanglement process, IIB provides two outputs: the compressed identity $\tilde{z}_{id}(X_s)$ as defined in Eq. (7), and the perceptual features which are identity-irrelevant. We define the perceptual features of X_t as:

$$f_i^t = \lambda_i^t \varepsilon_i^s + (1 - \lambda_i^t) R_i^t, \tag{13}$$

multiple as $F^t = \{f_1^t, f_2^t, \cdots, f_m^t\}$, and λ_i^t is the information controller defined in Eq. (3). Note that the controller λ_i^t is used in the way opposite to Eq. (4), since for the target, the information we need is in the areas not relevant to identity. Here the mean and std. of the noise are the same with R_i^s instead of R_i^t, *i.e.* $\varepsilon_i^s \sim \mathcal{N}(\mu_{R_i^s}, \sigma^2_{R_i^s})$, for latterly better integrating with the source identity.

Since the information in R_i is disentangled via the information controllers, it is natural to use the controllers again for re-integration. Therefore, we propose a novel Adaptive Information Integrator (AII) with a new set of *information integrators* $\Lambda^t = \{\Lambda_1^t, \Lambda_2^t, \cdots, \Lambda_m^t\}$, to guide the integration based on the controllers $\tilde{\lambda}_i^t$ which already learnt the identity-relevance of each area in $\tilde{f}_i^t$, measured by the amount of mutual information in bits.

As shown in Fig. 2(b), to get rid of the size of features in the pre-trained network and flexibly generate higher resolution images such as 512×512 and 1024×1024, we first use a U-Net decoder $dec(\cdot)$ to extend the spacial size of the perceptual feature f_i^t and controller λ_i^t, defined as:

$$\tilde{f}_i^t = dec(f_i^t) \text{ and } \tilde{\lambda}_i^t = dec(\lambda_i^t), \tag{14}$$

multiple as $\tilde{F}^t = \{\tilde{f}_1^t, \cdots, \tilde{f}_m^t\}$ and $\tilde{\lambda}^t = \{\tilde{\lambda}_1^t, \cdots, \tilde{\lambda}_m^t\}$. Then following AdaIN [18, 34], we use the affine parameters from $\tilde{z}_{id}^s$ and $\tilde{f}_i^t$ to normalize the internal activation in AII , defined as I_i^s and P_i^t respectively referring to the identity and perceptual activation. After that, the integration process based on the integrator Λ_i^t is formulated as:

$$\begin{aligned} h_{aii}^i &= \Lambda_i^t \cdot I_i^s + (1 - \Lambda_i^t) \cdot A_i^t, \\ \Lambda_i^t &= T_{\theta_i}(\tilde{\lambda}_i^t, h_{aii}^{i+1}) \in [0, 1], \end{aligned} \tag{15}$$

where T_{θ_i} is a parametric function (one convolution layer with $sigmoid$ activation) for the i-th feature level. h_{aii}^{i+1} is the output activation of the previous level, $h_{aii}^{m+1} = \tilde{z}_{id}^s$. The final result $Y_{s,t}$ is generated from the last activation h_{aii}^1.

Since the integrator Λ_i^t is learned based on the amount of mutual information each feature area contains about the identity, this new integration formula makes more sense than integrating based on the former activation h_{aii}^{i+1} alone. We demonstrate this by an ablation test in Section 4.3.

3.4. Training Losses

Adversarial Loss: To make the swapped face $Y_{s,t}$ more realistic, we employ a multi-scale discriminator D from [34] to train our model in an adversarial way, and adopt the relativistic adversarial loss from [20]:

$$\begin{aligned} L_{adv}^G &= -\mathbb{E}_{(X_s, Y_{s,t})}[\log(\sigma(D(Y_{s,t}) - D(X_s)))], \\ L_{adv}^D &= -\mathbb{E}_{(X_s, Y_{s,t})}[\log(\sigma(D(X_s) - D(Y_{s,t})))], \end{aligned} \tag{16}$$

where $\sigma(\cdot)$ denotes the sigmoid activation. The total adversarial loss is formulated as $L_{adv} = L_{adv}^G + L_{adv}^D$.

Perceptual Loss: Given the swapped result $Y_{s,t}$, we further use the pre-trained network to extract its perceptual features in the same way as Eq. (13) and (14):

$$\tilde{f}_i(Y_{s,t}) = dec(\lambda_i^{Y_{s,t}} \varepsilon_i^s + (1 - \lambda_i^{Y_{s,t}}) R_i^{Y_{s,t}}), \tag{17}$$

where the superscript $Y_{s,t}$ means the corresponding variables of $Y_{s,t}$. We define the multi-level perceptual loss, *i.e.* L_2 loss between the target perceptual features and $Y_{s,t}$ as:

$$L_{per} = \frac{1}{m} \sum_{i=1}^{m} [\tilde{f}_i(Y_{s,t}) - \tilde{f}_i^t]^2. \tag{18}$$

Cycle-consistency Loss: Additionally, we adopt a cycle-consistency loss L_{cyc} to further improve the preservation of the source identity. In specific, we use the compressed identity of $Y_{s,t}$ and the disentangled perceptual features of X_s to reconstruct the source image. We define the information integration process in AII as a generating function $g(*)$, then the reconstructed source can be formulated as:

$$\hat{X}_s = g(\tilde{z}_{id}(Y_{s,t}), \tilde{F}^s, \tilde{\lambda}^s), \tag{19}$$

where $\tilde{F}^s$ and $\tilde{\lambda}^s$ are the decoded perceptual features and controllers of X_s. Based on this, we define L_{cyc} as the $\mathcal{L}$-1 distance between X_s and its reconstruction $\hat{X}_s$:

$$L_{cyc} = \|X_s - \hat{X}_s\|_1. \tag{20}$$

Total Objective for InfoSwap: In summary, the final objective function for training InfoSwap is given by:

$$L_{obj.} = L_{IB} + \beta_1 L_{adv} + \beta_2 L_{per} + \beta_3 L_{cyc}, \tag{21}$$

where β_1, β_2, β_3 are hyper parameters.

4. Experiments

In this section, we first generally compare our method with several state-of-the-art methods using quantitative and qualitative metrics. Then, we further evaluate the performance of these methods on generating discriminative identities. After that, we analyze the impact of the proposed IB optimization on our method. We also report an ablation study to quantify the improvement brought by each component of InfoSwap. We start with implementation details.

During training, the internal features are extracted from the first 10 layers of the pre-trained network [10], *i.e.* $m = 10$, as these features are spatially larger. Accordingly, IIB consists of 10 information bottlenecks $IB = \{IB_1, IB_2, \cdots, IB_{10}\}$, and AII contains 10 integrating layers (Eq. (15)). The pre-trained network is not involved in updates, and other parts are trained end-to-end according to the total objective $L_{obj.}$ (Eq. (21)). The images used for the training are from the FFHQ [22] and the CelebA-HQ [21] datasets, with an initial resolution of 1024 pixels. We align these images by facial landmarks [49] and crop them to 512×512. The final training set consists of 96000 images after pre-processing, with the other 4000 images used for the test set. For more details on architecture and training strategies, please refer to the supplementary material.

4.1. Quantitative and Qualitative Results

In this subsection, we present quantitative and qualitative comparisons between InfoSwap and state-of-the-art methods on the preservation of *source identities*, *target poses* and *expressions*. The experiments are conducted on the FaceForensics++ (FF++) datasets [35].

Quantitative Results: We first use three quantitative metrics to evaluate the swapping performance of each method. For Deepfakes [1], FaceSwap [25] and FaceShifter [26] which provide their manipulated videos in FF++, we evenly extract 10 frames from each video and build a test set of size 10k for each method following the same protocol with [26]. As for FSGAN [32], DiscoFaceGAN [11] and our method, the test sets of the equal size are generated using the same source-to-target pairs as the others. As shown in Tab. 1, the *ID retrieval* (%) is the mean accuracy of a classification among swapped faces and all FF++ original faces used to measure the identity preservation. Following the same testing protocol with [26], we use the face recognition model [46] to extract the identity embeddings. For each swapped face, we find out the nearest identity (on cosine distance) in all original faces and check whether it belongs to the correct source. The results show that our method has better performance on identity preservation.

method	ID retrieval↑	pose↓	expression↓
FSGAN [32]	60.41	0.626	**0.028**
Deepfakes [1]	<u>81.96</u>	1.092	0.114
FaceSwap [25]	<u>54.19</u>	0.488	0.029
DiscoFaceGAN[11]	93.12	1.197	0.159
FaceShifter [26]	<u>97.38</u>	0.511	0.032
InfoSwap	**99.67**	**0.443**	0.030

Table 1. General comparisons with SOTA methods. ↑: the higher the better; ↓: the lower the better. Values underlined are from [26], others are computed following the same protocol.

The *pose* and *expression* reported in Tab. 1 are the mean square error between the swapped faces and the corresponding targets, measuring the preservation of target perceptual information. Since the estimation models used in [26] are not available by now, we use another 3D face alignment model [15] to estimate the pose and expression parameters. The results demonstrate that our method is comparable to other methods in preserving perceptual information. The expression errors of FSGAN [32] and FaceSwap [25] are slightly lower than ours, probably due to their strategy of only generating the inner face regions and swapping them to the target face by blending. However, such a strategy could cause the problem of blending inconsistency.

Qualitative Results: (I) Results on FF++. As shown in Fig. 1, we compare our results with Deepfakes [1], FaceSwap [25], and the latest work FaceShifter [26] on preserving various identity traits of the source, including face shape, eyes color, cheekbone shape, skin type, baby fat, and chin length. The comparisons are based on the test data provided by FF++. We can see that results produced by our method share these identity traits with the source much better than other methods. This demonstrates the strong ability of our method to better preserve source identity and make the generated identities more discriminative. (II) Results on the test set of FFHQ and CelebA-HQ. We present more test results in Fig. 4, which demonstrate the strong performance of InfoSwap on producing high-quality swapped faces across large gaps between different genders, ages, skin colors, and lighting conditions. More swapped results are provided in the supplementary material.

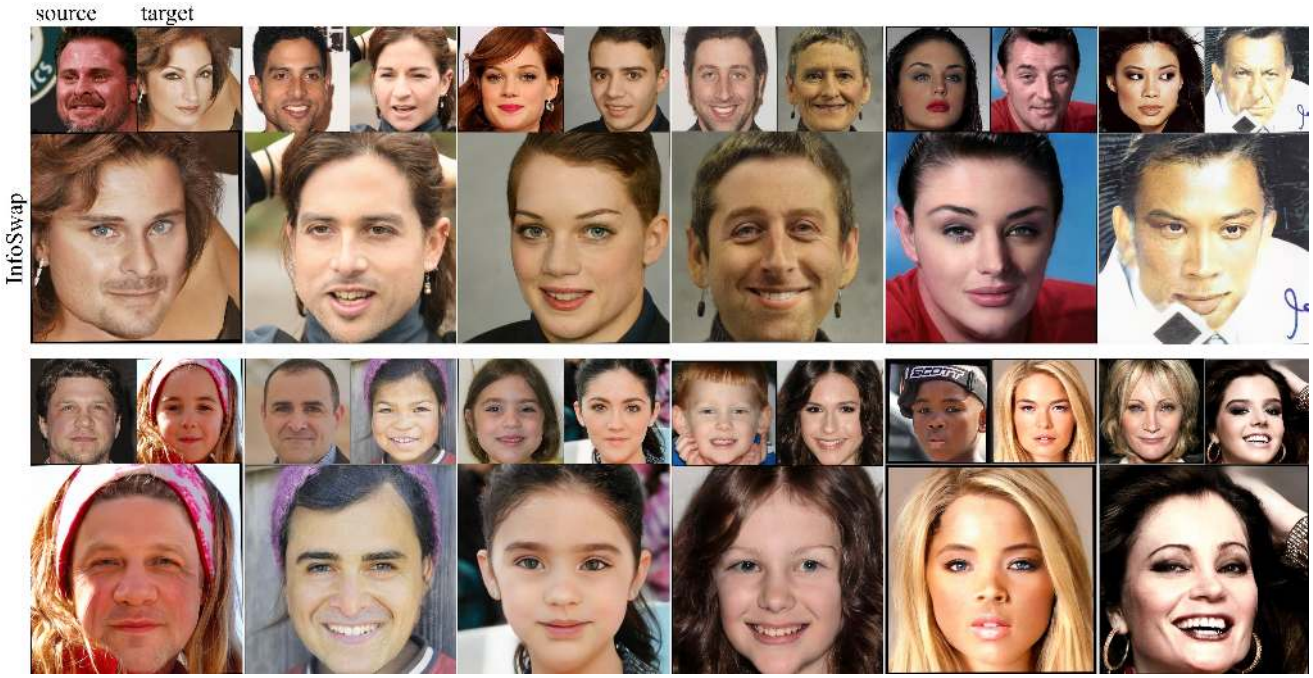

Figure 4. More test results swapped by InfoSwap across large gaps in genders, ages, skin colors and lighting conditions.

User Evaluation: We conduct a user survey to evaluate the performance of each method concerning preserving source identity and target perception. For each user, we randomly sample 30 source-to-target pairs of faces (frames) from all 1000 FF++ videos without duplication. Users are asked to select from the 4 results (produced by 4 methods using each pair), the one: (i) *most identical to the source face*; (ii) *with the most similar expression and posture to target*; (iii) *looks most realistic*. The results reported in Tab. 2 are based on the answers from 50 users, showing that our method significantly outperforms others in all three aspects.

method	Identity	Perception	Fidelity
Deepfakes [1]	0.131	0.052	0.026
FaceSwap [25]	0.120	0.244	0.050
FaceShifter [26]	0.238	0.267	0.246
InfoSwap	**0.511**	**0.437**	**0.678**

Table 2. User Study. Percentage of each method being selected.

4.2. Identity Preserving Evaluation

In this subsection, we provide a detailed comparison of identity preserving. We demonstrate that identities generated by InfoSwap are more discriminative by providing statistical analyses on their similarities to the source and target.

As shown in Tab. 3, we calculate the cosine similarities of identities generated by InfoSwap and four SOTA methods. The second column shows the mean and std. of the similarities between fake identities and the source identities, while the third column shows the values with the target, and the last row shows the values between the source and the target. Fig. 5 is the visualization of this table.

It is obvious that comparing with other methods, most of the similarities between InfoSwap identities and source identities (the fifth red box in Fig. 5) are higher than others.

method	with source		with target	
	mean	std.	mean	std.
FSGAN [32]	0.3874	0.1722	0.3478	0.1444
Deepfakes [1]	0.4784	0.1398	0.2666	0.1287
FaceSwap [25]	0.4328	0.1409	0.3236	0.1274
FaceShifter [26]	0.5295	0.1418	0.3108	0.1418
InfoSwap	**0.6332**	**0.0983**	**0.0770**	**0.1035**
source with target			0.0669	0.1025

Table 3. Cosine similarities of identities. The test sets are expanded to 100k frames (100 frames per video) to better display the distribution since the variability of each sampling distribution decreases as the sample sizes increase.

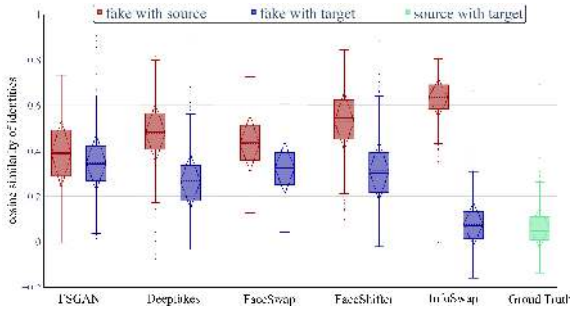

Figure 5. Comparisons with SOTA methods on cosine similarities of identities. The mean similarity of InfoSwap to the source is higher than others, while its mean similarity to the target is much closer to the level between source and target. The narrower boxes of InfoSwap also indicate its superiority in robustness.

While the similarities to the target identities (the fifth blue box) are at a much lower level that very close to the similarities between source and target (the green box). Also, the narrower intervals (box length) of InfoSwap indicate that our method is more robust than other methods.

More importantly, there are zero overlaps between the

two estimated intervals of InfoSwap (the fifth pair of red and blue boxes), which is not observed in other methods. The zero overlap directly indicates that our generated identities are more discriminative, as mentioned in Section 3.1. Experimental results on FF++ show that, similarities between the target and the identities generated by InfoSwap are of 97.61% falling into the range of $\mu_{s,t} \pm 3\sigma_{s,t}$, shown as the lightest green belt in Fig. 3(b), 88.57% into $\mu_{s,t} \pm 2\sigma_{s,t}$, and 59.44% into $\mu_{s,t} \pm \sigma_{s,t}$, indicating that InfoSwap is powerful in generating highly discriminative identities.

4.3. Analysis of Components

In this subsection, we start by visualizing the information varied in the internal features of the pre-trained model with and without the informative optimization, showing the impact of inserting IIB directly.

Visualization of IIB Optimization: Fig. 6 shows the information contained in the first 8 intermediate feature maps of the pre-trained model with and without the optimization by IIB (*i.e.* Z_i and R_i). We can see that in the original uncompressed features R_i (Fig. 6 (a)), the information is scattered. Some areas outside the face are regarded as informatively important (*e.g.* the red areas in the hair), which are indeed less relevant to identity. While after the IIB optimization (Fig. 6 (b)), the dispersed information is properly compressed so that in all Z_i, areas considered informatively efficient are concentrated on the face, which well demonstrates the power of IIB in facilitating disentanglement.

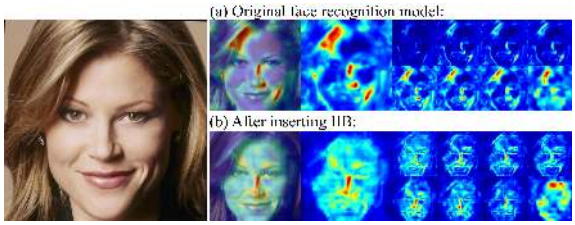

Figure 6. Visualization of information variation in feature maps.

Explanation to Generated Face Shape: As the former mentioned, the face shape of the results from InfoSwap is observed to be closer to the source rather than the target. This may benefit from better disentanglement. During the optimization in IIB, the shape information of the source face is learned to be relevant and allowed to pass through the bottleneck, thus retained in the compressed source identity to join in the generation. On the other hand, the shape information of the target face is excluded from the identity-irrelevant perceptual representations and hence is not involved in the generation. This may also suggest that the face recognition model [10] in use is sensitive to the shape information.

Ablation Study: We further perform an ablation study on FF++ with three configurations of InfoSwap: (i) removing the IIB module (w/o IIB); (ii) replacing the ICL by conventional identity loss (measures the cosine distance between the fake and the source identity) (w/o ICL); (iii) discarding the controller $\tilde{\lambda}_i^t$ in Eq. (15) to integrate the activation I_i^s and A_i^t only based on the former layer output h_{aii}^{i+1} (w/o $\tilde{\lambda}_i^t$). We calculate the main metrics introduced above to measure the effect of each component. As shown in Tab. 4, the mean similarity to the source drops significantly (greater than one std.) in all three configurations, with "w/o IIB" declining the most and ending up with the lowest value on the ID retrieval. The similarity to the target increases obviously when replacing ICL with conventional identity loss, indicating that the swapped faces become less distinct from the target. Discarding information controllers $\tilde{\lambda}_i^t$ in integration also degrades the performance. More qualitative results of the ablation experiments please refer to the supplementary.

method	Cosine Similarity		Acc.
	with source ↑	with target ↓	
InfoSwap	**0.633 ± 0.098**	**0.077 ± 0.104**	**99.7**
InfoSwap w/o IIB	0.529 ± 0.118	0.119 ± 0.116	96.3
InfoSwap w/o ICL	0.544 ± 0.111	0.244 ± 0.118	97.9
InfoSwap w/o $\tilde{\lambda}_i^t$	0.550 ± 0.110	0.096 ± 0.111	98.5

Table 4. Ablation results. Acc.: ID retrieval (%) ↑

5. Conclusions

In this paper, we have presented InfoSwap for learning well-disentangled representations. By modeling the learning process as finding an optimal compression of the pre-trained latent features based on the information bottleneck principle, the extracted representations for identity and perceptual information are efficiently disentangled. We extend the IB objectives with the intuition of contrastive learning and enable us to generate identity-discriminative swapped faces. Extensive experiments demonstrate the superiority of InfoSwap in subject-agnostic face swapping, which is an encouraging development for building new benchmarks to improve the performance of data-driven forgery detectors.

6. Broader Impact

Deepfakes, synthetic media that replace a person in an existing image or video with the appearance of someone else, have been in the spotlight since they first appeared. Empowered by it, film-making, computer games, and other mixed realities are about to see a breakthrough. Yet it also sparks serious problems in privacy protection if misused. Identity swapping, the concern of this paper, is one of the main methods for manipulating the appearances of face images. Given such potentially negative impacts, several face forgery detection technologies have recently been proposed to prevent the misuse of Deepfakes. We will further discuss such impacts as well as the corresponding solution regarding our work in the supplementary.

Acknowledgments:

This work is partially funded by National Key Research and Development Program of China (Grant No. 2020AAA0140001), National Natural Science Foundation of China (Grant No. 62006228), and Youth Innovation Promotion Association CAS (Grant No. Y201929).

References

[1] *Deepfakes*, Accessed: 2020-09-21. https://github.com/deepfakes/faceswap.

[2] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *JMLR*, 19(1):1947–1980, 2018.

[3] Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. Deep Variational Information Bottleneck. In *ICLR*, 2017.

[4] Jian-Min Bao, Dong Chen, Fang Wen, Hou-Qiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *CVPR*, 2018.

[5] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K. Nayar. Face swapping: Automatically replacing faces in photographs. In *SIGGRAPH*, 2008.

[6] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. *Computer Graphics Forum*, 23(3):669–676, 2004.

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *PACMCGIT*, 1999.

[8] Yi-Ting Cheng, Virginia Tzeng, Yu Liang, Chuan-Chang Wang, Bing-Yu Chen, Yung-Yu Chuang, and Ming Ouhyoung. 3D-model-based face replacement in video. In *SIGGRAPH*, 2009.

[9] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively with application to face verification. In *CVPR*, 2015.

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

[11] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *CVPR*, 2020.

[12] Chaoyou Fu, Yibo Hu, Xiang Wu, Guoli Wang, Qian Zhang, and Ran He. High fidelity face manipulation with extreme poses and expressions. *TIFS*, 2020.

[13] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dual variational generation for low shot heterogeneous face recognition. In *NeurIPS*, 2019.

[14] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dvg-face: Dual variational generation for heterogeneous face recognition. *TPAMI*, 2021.

[15] Jian-Zhu Guo, Xiang-Yu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *ECCV*, 2020.

[16] Weikuo Guo, Huaibo Huang, Xiangwei Kong, and Ran He. Learning disentangled representation for cross-modal retrieval with deep mutual information estimation. In *ACM ICM*, 2019.

[17] Naama Hadad, Lior Wolf, and Moni Shahar. A two-step disentanglement method. In *CVPR*, 2018.

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

[19] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

[20] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. In *ICLR*, 2019.

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *ICLR*, 2018.

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[23] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020.

[24] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[25] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *ICCV*, 2017.

[26] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *CVPR*, 2020.

[27] Ling-Zhi Li, Jian-Min Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Bai-Ning Guo. Face x-ray for more general face forgery detection. In *CVPR*, 2020.

[28] Yue-Zun Li, Xin Yang, Pu Sun, Hong-Gang Qi, and Si-Wei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *CVPR*, 2020.

[29] Yuan Lin, Sheng-Jin Wang, Qian Lin, and Feng Tang. Face swapping under large pose variations: A 3D model based approach. In *ICME*, 2012.

[30] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. RSGAN: Face swapping and editing using face and hair representation in latent spaces. In *SIGGRAPH*, 2018.

[31] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *BTAS*, 2019.

[32] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, 2019.

[33] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *FG*, 2018.

[34] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.

[35] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019.

[36] Natsume Ryota, Tatsuya Yatagawa, and Shigeo Morishima. Fsnet: An identity-aware generative model for image-based face swapping. In *ACCV*, 2018.

[37] Eduardo Hugo Sanchez, Mathieu Serrurier, and Mathias Ortner. Learning disentangled representations via mutual information estimation. In *ECCV*, 2020.

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

[39] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. In *ICLR*, 2020.

[40] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[41] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[42] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *TOG*, 38(4):1–12, 2019.

[43] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016.

[44] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *ITW*, 2015.

[45] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017.

[46] Hao Wang, Yi-Tong Wang, Zheng Zhou, Xing Ji, Di-Hong Gong, Jing-Chao Zhou, Zhi-Feng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, 2018.

[47] Hong-Xia Wang, Chun-hong Pan, Hai-feng Gong, and Huai-Yu Wu. Facial image composition based on active appearance model. In *ICASSP*, 2008.

[48] Wu Xiang, Huaibo Huang, Vishal M Patel, Ran He, and Zhenan Sun. Disentangled variational representation for heterogeneous face recognition. In *AAAI*, 2019.

[49] Kai-Peng Zhang, Zhan-Peng Zhang, Zhi-Feng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[50] Xinqi Zhu, Chang Xu, and Dacheng Tao. Learning disentangled representations with latent variation predictability. In *ECCV*, 2020.