FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping

Lingzhi Li[1*] Jianmin Bao[2] Hao Yang[2] Dong Chen[2] Fang Wen[2]
[1]Peking University [2]Microsoft Research
lilingzhi@pku.edu.cn {jianbao,haya,doch,fangwen}@microsoft.com

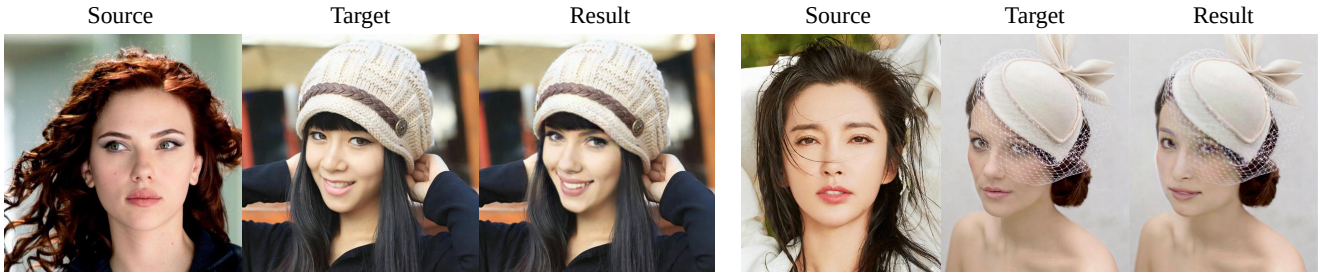

Figure 1: The face in the source image is taken to replace the face in the target image. Results of FaceShifter appear in the right.

Abstract

In this work, we propose a novel two-stage framework, called FaceShifter, for high fidelity and occlusion aware face swapping. Unlike many existing face swapping works that leverage only limited information from the target image when synthesizing the swapped face, our framework, in its first stage, generates the swapped face in high-fidelity by exploiting and integrating the target attributes thoroughly and adaptively. We propose a novel attributes encoder for extracting multi-level target face attributes, and a new generator with carefully designed Adaptive Attentional Denormalization (AAD) layers to adaptively integrate the identity and the attributes for face synthesis. To address the challenging facial occlusions, we append a second stage consisting of a novel Heuristic Error Acknowledging Refinement Network (HEAR-Net). It is trained to recover anomaly regions in a self-supervised way without any manual annotations. Extensive experiments on wild faces demonstrate that our face swapping results are not only considerably more perceptually appealing, but also better identity preserving in comparison to other state-of-the-art methods.

1. Introduction

Face swapping is the replacement of the identity of a person in the target image with that of another person in the source image, while preserving attributes *e.g.* head pose, facial expression, lighting, background *etc*. Face swapping has attracted great interest in vision and graphics community, because of its potential wide applications in movie composition, computer games, and privacy protection [34].

The main difficulties in face swapping are how to extract and adaptively recombine identity and attributes of two images. Early replacement-based works [6, 42] simply replace the pixels of inner face region. Thus, they are sensitive to the variations in posture and perspective. 3D-based works [7, 12, 25, 30] used a 3D model to deal with the posture or perspective difference. However the accuracy and robustness of 3D reconstruction of faces are all unsatisfactory. Recently, GAN-based works [21, 27, 28, 29, 4] have illustrated impressive results. But it remains challenging to synthesize both realistic and high-fidelity results.

In this work, we focus on improving the fidelity of face swapping. In order to make the results more perceptually appealing, it is important that the synthesized swapped face not only shares the pose and expression of the target face, but also can be seamlessly fitted into the target image without inconsistency: the rendering of the swapped face should be faithful to the lighting (*e.g.* direction, intensity, color) of the target scene, the pixel resolution of the swapped face should also be consistent with the target image resolution. Neither of these can be well handled by a simple alpha or Poisson blending. Instead, we need a *thorough and adaptive integration of target image attributes during the synthesis of the swapped face*, so that the attributes from the target image, including scene lighting or image resolution, can help make the swapped face more realistic.

However, previous face swapping methods either neglect the requirement of this integration, or lack the ability to per-

* Work done during an internship at Microsoft Research Asia

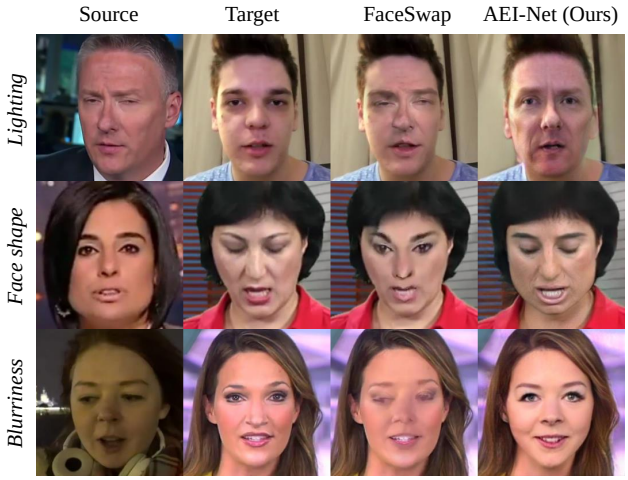

Figure 2: Failure cases of a previous method on FaceForensics++ [35] dataset. From left to right we show the input source images, the input target images, the results of FaceSwap [2], and the results of our method. FaceSwap follows the strategy that, first synthesizes the inner face region, then blends it into the target face. Such strategy causes artifacts, such as the defective lighting effect on the nose (row 1), failing to preserve the face shape of the source identity (row 2) and the mismatched image resolutions (row 3). While our method addresses all these issues.

form it in a thorough and adaptive way. In specific, many previous methods use only pose and expression guidances from the target image to synthesize the swapped face, the face is then blended into the target image using masks of the target faces. This process is easy to cause artifacts, because: 1) Besides pose and expression, it leverages little knowledge about the target image when synthesizing the swapped face, which can hardly respect target attributes like the scene lightings or the image resolutions; 2) Such a blending will discard all the peripheral area of the source face that locates outside the target face mask. Thus these methods cannot preserve the face shape of the source identity. We show some typical failure cases in Figure 2.

In order to achieve high-fidelity face swapping results, in the first stage of our framework, we design a GAN-based network, named *Adaptive Embedding Integration Network* (AEI-Net), for a thorough and adaptive integration of target attributes. We made two improvements to the network structure: 1) we propose a novel *multi-level attributes encoder* for extracting target attributes in various spatial resolutions, instead of compressing it into a single vector as RSGAN [28] and IPGAN [5]. 2) we present a novel generator with carefully designed *Adaptive Attentional Denormalization* (AAD) layers which adaptively learns where to integrate the attributes or identity embeddings. Such an adaptive integration brings considerable improvements over the single level integration used by RSGAN [28], FSNet [27] and IPGAN [5]. With these two improvements, the proposed AEI-Net can solve the problem of inconsistent illumination and face shape, as shown in Figure 2.

Moreover, handling facial occlusions is always challenging in face swapping. Unlike Nirkin *et al*. [29, 30] that trains face segmentation to obtain occlusion-aware face masks, our method can learn to recover face anomaly regions in a self-supervised way without any manual annotations. We observe that when feeding the same face image as both the target and source into a well trained AEI-Net, the reconstructed face image deviates from the input in multiple areas, these deviations strongly hint the locations of face occlusions. Thus, we propose a novel *Heuristic Error Acknowledging Refinement Network* (HEAR-Net) to further refine the result under the guidance of such reconstruction errors. The proposed method is more general, thus it identifies more anomaly types, such as glasses, shadow and reflection effects, and other uncommon occlusions.

The proposed two-stage face swapping framework, FaceShifter, is subject agnostic. Once trained, the model can be applied to any new face pairs without requiring subject specific training as DeepFakes [1] and Korshunova *et al*. [21]. Experiments demonstrate that our method achieves results considerably more realistic and more faithful to inputs than other state-of-the-art methods.

2. Related Works

Face swapping has a long history in vision and graphics researches. Early efforts [6, 42] only swap faces with similar poses. Such a limitation is addressed by recent algorithms roughly divided in two categories: 3D-based approaches and GAN-based approaches.

3D-Based Approaches. Blanz *et al*. [7] considers 3D transform between two faces with different poses, but requiring user interaction and not handling expressions. Thies *et al*. [38] captures head actions from a RGB-D image using 3DMM, turning a static face into a controllable avatar. It is extended for RGB references in Face2Face [39]. Olszewski *et al*. [31] dynamically inferences 3D face textures for improved manipulation quality. Kim *et al*. [19] separately models different videos using 3DMM to make the portraits controllable, while Nagano *et al*. [26] needs only one image to reenact the portrait within. Recently, Thies *et al*. [37] adopt neural textures, which can better disentangle geometry in face reenactment. However, when applied on face swapping, these methods hardly leverage target attributes like occlusions, lighting or photo styles. To preserve the target facial occlusions, Nirkin *et al*. [30, 29] collected data to train an occlusion-aware face segmentation network in a supervised way, which helps predict a visible target face mask for blending in the swapped face. While our method find the occlusions in a self-supervised way without any manually annotations.

GAN-Based Approaches. In the GAN-based face swap-

ping methods, Korshunova *et al*. [22] swap faces like transfer styles. It separately models different source identities, such as a CageNet for Nicolas Cage, a SwiftNet for Taylor Swift. The recently popular DeepFakes [1] is another example of such subject-aware face swapping: for each new input, a new model has to be trained on two video sequences, one for the source and one for the target.

This limitation has been addressed by subject-agnostic face swapping researches: RSGAN [28] learns to extract vectorized embeddings for face and hair regions separately, and recombines them to synthesize a swapped face. FSNet [27] represents the face region of source image as a vector, which is combined with a non-face target image to generate the swapped face. IPGAN [5] disentangles the identity and attributes of faces as vectors. By introducing supervisions directly from the source identity and the target image, IPGAN supports face swapping with better identity preservation. However, due to the information loss caused by the compressed representation, and the lack of more adaptive information integration, these three methods are incapable of generating high-quality face images. Recently, FSGAN [29] performs face reenactment and face swapping together. It follows a similar reenact and blend strategy with [31, 26]. Although FSGAN utilizes an occlusion-aware face segmentation network for preserving target occlusions, it hardly respects target attributes like the lighting or image resolution, it can neither preserve the face shape of the source identity.

3. Methods

Our method requires two input images, *i.e*., a source image X_s to provide identity and a target image X_t to provide attributes, *e.g*., pose, expression, scene lighting and background. The swapped face image is generated through a two-stage framework, called FaceShifter. In the first stage, we use an *Adaptive Embedding Integration Network* (AEI-Net) to generate a high fidelity face swapping result $\hat{Y}_{s,t}$ based on information integration. In the second stage, we use the *Heuristic Error Acknowledging Network* (HEAR-Net) to handle the facial occlusions and refine the result, the final result is denoted by $Y_{s,t}$.

3.1. Adaptive Embedding Integration Network

In the first stage, we aim to generate a high fidelity face image $\hat{Y}_{s,t}$, which should preserve the identity of the source X_s and the attributes (*e.g*. pose, expression, lighting, background) of the target X_t. To achieve this goal, our method consist of 3 components: i) the *Identity Encoder* $\boldsymbol{z}_{id}(X_s)$, which extracts identity from the source image X_s; ii) the *Multi-level Attributes Encoder* $\boldsymbol{z}_{att}(X_t)$, which extracts attributes of the target image X_t; iii) *Adaptive Attentional Denormalization (AAD) Generator*, which generates swapped face image. Figure 3(a) shows whole network structure.

Identity Encoder: We use a pretrained state-of-the-art face recognition model [13] as identity encoder. The identity embedding $\boldsymbol{z}_{id}(X_s)$ is defined to be the last feature vector generated before the final FC layer. We believe that by training on a large quantity of 2D face data, such a face recognition model can provide more representative identity embeddings than the 3D-based models like 3DMM [7, 8].

Multi-level Attributes Encoder: Face attributes, such as pose, expression, lighting and background, require more spatial informations than identity. In order to preserve such details, we propose to represent the attributes embedding as multi-level feature maps, instead of compressing it into a single vector as previous methods [5, 28]. In specific, we feed the target image X_t into a U-Net-like structure. Then we define the attributes embedding as the feature maps generated from the U-Net decoder. More formally, we define

$$\boldsymbol{z}_{att}(X_t) = \left\{\boldsymbol{z}^1_{att}(X_t), \boldsymbol{z}^2_{att}(X_t), \cdots \boldsymbol{z}^n_{att}(X_t)\right\}, \quad (1)$$

where $\boldsymbol{z}^k_{att}(X_t)$ represents the k-th level feature map from the U-Net decoder, n is the number of feature levels.

Our attributes embedding network does not require any attribute annotations, it extracts the attributes using self-supervised training: we require that the generated swapped face $\hat{Y}_{x_t}$ and the target image X_t have the same attributes embedding. The loss function will be introduce in Equation 7. In the experimental part (Section 4.2), we also study what the attributes embedding has learned.

Adaptive Attentional Denormalization Generator: We then integrate such two embeddings $\boldsymbol{z}_{id}(X_s)$ and $\boldsymbol{z}_{att}(X_t)$ for generating a raw swapped face $\hat{Y}_{s,t}$. Previous methods [5, 28] simply integrate them through feature concatenation. It will lead to relatively blurry results. Instead, we propose a novel *Adaptive Attentional Denormalization* (AAD) layer to accomplish this task in a more adaptive fashion. Inspired by the mechanisms of SPADE [32] and AdaIN [14, 16], the proposed AAD layers leverage denormalizations for feature integration in multiple feature levels.

As shown in Figure 3(c), in the k-th feature level, let $\boldsymbol{h}^k_{in}$ denote the activation map that is fed into an AAD layer, which should be a 3D tensor of size $C^k \times H^k \times W^k$, with C^k being the number of channels and $H^k \times W^k$ being the spatial dimensions. Before integration, we perform instance normalization [40] on $\boldsymbol{h}^k_{in}$:

$$\bar{\boldsymbol{h}}^k = \frac{\boldsymbol{h}^k_{in} - \boldsymbol{\mu}^k}{\boldsymbol{\sigma}^k}. \quad (2)$$

Here $\boldsymbol{\mu}^k \in \mathbb{R}^{C^k}$ and $\boldsymbol{\sigma}^k \in \mathbb{R}^{C^k}$ are the means and standard deviations of $\boldsymbol{h}^k_{in}$'s channel-wise activations. Then, we design 3 parallel branches from $\bar{\boldsymbol{h}}^k$ for 1) attributes integration, 2) identity integration, 3) adaptively attention mask.

For attributes embedding integration, let $\boldsymbol{z}^k_{att}$ be the attributes embedding on this feature level, which should be a 3D tensor of size $C^k_{att} \times H^k \times W^k$. In order to integrate $\boldsymbol{z}^k_{att}$ into the activation, we compute an attribute activation

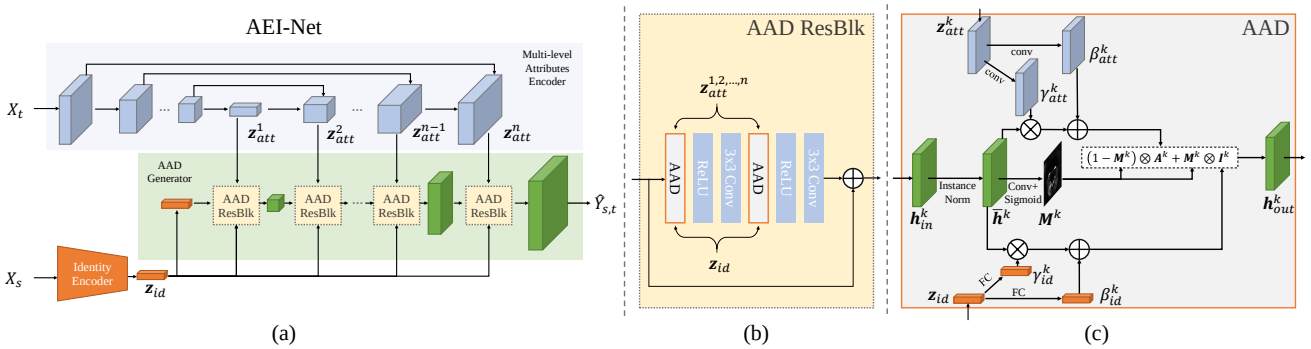

Figure 3: AEI-Net for the first stage. AEI-Net is composed of an Identity Encoder, a Multi-level Attributes Encoder, and an AAD-Generator. The AAD-Generator integrates informations of identity and attributes in multiple feature levels using cascaded AAD ResBlks, which is built on AAD layers.

$\boldsymbol{A}^k$ by denormalizing the normalized $\bar{\boldsymbol{h}}^k$ according to the attributes embedding, formulated as

$$\boldsymbol{A}^k = \gamma^k_{att} \otimes \bar{\boldsymbol{h}}^k + \beta^k_{att}, \tag{3}$$

where γ^k_{att} and β^k_{att} are two modulation parameters both convolved from $\boldsymbol{z}^k_{att}$. They share the same tensor dimensions with $\bar{\boldsymbol{h}}^k$. The computed γ^k_{att} and β^k_{att} are multiplied and added to $\bar{\boldsymbol{h}}^k$ element-wise.

For identity embedding integration, let $\boldsymbol{z}^k_{id}$ be the identity embedding, which should be a 1D vector of size C_{id}. We also integrate $\boldsymbol{z}^k_{id}$ by computing an identity activation $\boldsymbol{I}^k$ in a similar way to integrating attributes. It is formulated as

$$\boldsymbol{I}^k = \gamma^k_{id} \otimes \bar{\boldsymbol{h}}^k + \beta^k_{id}, \tag{4}$$

where $\gamma^k_{id} \in \mathbb{R}^{C^k}$ and $\beta^k_{id} \in \mathbb{R}^{C^k}$ are another two modulation parameters generated from $\boldsymbol{z}_{id}$ through FC layers.

One key design of the AAD layer is to adaptively adjust the effective regions of the identity embedding and the attributes embedding, so that they can participate in synthesizing different parts of the face. For example, the identity embedding should focus relatively more on synthesizing the face parts that are most discriminative for distinguishing identities, *e.g.* eyes, mouth and face contour. Therefore, we adopt an attention mechanism into the AAD layer. Specifically, we generate an attentional mask $\boldsymbol{M}^k$ using $\bar{\boldsymbol{h}}^k$ through convolutions and a sigmoid operation. The values of $\boldsymbol{M}^k$ are between 0 and 1.

Finally, the output of this AAD layer $\boldsymbol{h}^k_{out}$ can be obtained as a element-wise combination of the two activations $\boldsymbol{A}^k$ and $\boldsymbol{I}^k$, weighted by the mask $\boldsymbol{M}^k$, as shown in Figure 3(c). It is formulated as

$$\boldsymbol{h}^k_{out} = (1 - \boldsymbol{M}^k) \otimes \boldsymbol{A}^k + \boldsymbol{M}^k \otimes \boldsymbol{I}^k. \tag{5}$$

The AAD-Generator is then built with multiple AAD layers. As illustrated in Figure 3(a), after extracting the identity embedding $\boldsymbol{z}_{id}$ from source X_s, and the attributes embedding $\boldsymbol{z}_{att}$ from target X_t, we cascade AAD Residual Blocks (AAD ResBlks) to generate the swapped face $\hat{Y}_{s,t}$, the structure of the AAD ResBlks is shown in Figure 3(b).

For the AAD ResBlk on the k-th feature level, it first takes the up-sampled activation from the previous level as input, then integrates this input with $\boldsymbol{z}_{id}$ and $\boldsymbol{z}^k_{att}$. The final output image $\hat{Y}_{s,t}$ is convolved from the last activation.

Training Losses We utilize adversarial training for AEI-Net. Let $\mathcal{L}_{adv}$ be the adversarial loss for making $\hat{Y}_{s,t}$ realistic. It is implemented as a multi-scale discriminator [32] on the downsampled output images. In addition, an identity preservation loss is used to preserve the identity of the source. It is formulated as

$$\mathcal{L}_{id} = 1 - cos(\boldsymbol{z}_{id}(\hat{Y}_{s,t}), \boldsymbol{z}_{id}(X_s)), \tag{6}$$

where $cos(\cdot,\cdot)$ represents the cosine similarity of two vectors. We also define the attributes preservation loss as $\mathcal{L}$-2 distances between the multi-level attributes embeddings from X_t and $\hat{Y}_{s,t}$. It is formulated as

$$\mathcal{L}_{att} = \frac{1}{2}\sum_{k=1}^{n} \left\| \boldsymbol{z}^k_{att}(\hat{Y}_{s,t}) - \boldsymbol{z}^k_{att}(X_t) \right\|_2^2. \tag{7}$$

When the source and target images are the same in a training sample, we define a reconstruction loss as pixel level $\mathcal{L}$-2 distances between the target image X_t and $\hat{Y}_{s,t}$

$$\mathcal{L}_{rec} = \begin{cases} \frac{1}{2}\left\| \hat{Y}_{s,t} - X_t \right\|_2^2 & if\ X_t = X_s \\ 0 & otherwise \end{cases}. \tag{8}$$

The AEI-Net is finally trained with a weighted sum of above losses as

$$\mathcal{L}_{\texttt{AEI-Net}} = \mathcal{L}_{adv} + \lambda_{att}\mathcal{L}_{att} + \lambda_{id}\mathcal{L}_{id} + \lambda_{rec}\mathcal{L}_{rec}, \tag{9}$$

with $\lambda_{att} = \lambda_{rec} = 10, \lambda_{id} = 5$. The trainable modules of AEI-Net include the Multi-level Attributes Encoder and the ADD-Generator.

3.2. Heuristic Error Acknowledging Refinement Network

Although the face swap result $\hat{Y}_{s,t}$ generated with AEI-Net in the first stage can well retain target attributes like

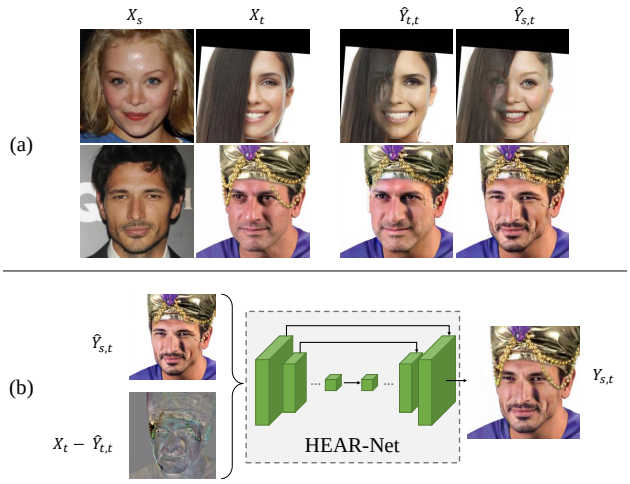

Figure 4: HEAR-Net for the second stage. $\hat{Y}_{t,t}$ is the reconstruction of the target image X_t, *i.e.*, $\hat{Y}_{t,t} = \texttt{AEI-Net}(X_t, X_t)$. $\hat{Y}_{s,t}$ is the swapped face from the first stage.

pose, expression and scene lighting, it often fails to preserve the occlusions appeared on the target face X_t. Previous methods [30, 29] address face occlusions with an additional face segmentation network. It is trained on face data containing occlusion-aware face masks, which require lots of manual annotations. Besides, such a supervised approach may hardly recognize unseen occlusion types.

We proposed a heuristic method to handle facial occlusions. As shown in Figure 4(a), when the target face was occluded, some occlusions might disappear in the swapped face, *e.g.*, the hair covering the face or the chains hang from the turban. Meanwhile, we observe that if we feed the same image as both the source and target images into a well trained AEI-Net, these occlusions would also disappear in the reconstructed image. Thus, the error between the reconstructed image and its input can be leveraged to locate face occlusions. We call it the *heuristic error* of the input image, since it heuristically indicates where anomalies happen.

Inspired by the above observation, we make use of a novel HEAR-Net to generate a refined face image. We first get the heuristic error of the target image as

$$\Delta Y_t = X_t - \texttt{AEI-Net}(X_t, X_t). \tag{10}$$

Then we feed the heuristic error ΔY_t and the result of the first stage $\hat{Y}_{s,t}$ into a U-Net structure, and obtain the refined image $Y_{s,t}$:

$$Y_{s,t} = \texttt{HEAR-Net}(\hat{Y}_{s,t}, \Delta Y_t). \tag{11}$$

The pipeline of HEAR-Net is illustrated in Figure 4(b).

We train HEAR-Net in a fully self-supervised way, without using any manual annotations. Given any target face image X_t, with or without occlusion regions, we utilize the following losses for training HEAR-Net. The first is an identity preservation loss to preserve the identity of the source. Similar as stage one, it is formulated as

$$\mathcal{L}'_{id} = 1 - cos(\boldsymbol{z}_{id}(Y_{s,t}), \boldsymbol{z}_{id}(X_s)). \tag{12}$$

The change loss $\mathcal{L}'_{chg}$ guarantees the consistency between the results of the first stage and the second stage:

$$\mathcal{L}'_{chg} = \left| \hat{Y}_{s,t} - Y_{s,t} \right|. \tag{13}$$

The reconstruction loss $\mathcal{L}'_{rec}$ restricts that the second stage is able to reconstruct the input when the source and target images are the same:

$$\mathcal{L}'_{rec} = \begin{cases} \frac{1}{2} \left\| Y_{s,t} - X_t \right\|_2^2 & \text{if } X_t = X_s \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

Since the number of occluded faces is very limited in most face datasets, we propose to augment data with synthetic occlusions. The occlusions are randomly sampled from a variety of datasets, including the EgoHands [3], GTEA Hand2K [15, 24, 23] and ShapeNet [9]. They are blended onto existing face images after random rotations, rescaling and color matching. Note that *we do not utilize any occlusion mask supervision during training, even from these synthetic occlusions*.

Finally, HEAR-Net is trained with a sum of above losses:

$$\mathcal{L}_{\texttt{HEAR-Net}} = \mathcal{L}'_{rec} + \mathcal{L}'_{id} + \mathcal{L}'_{chg}. \tag{15}$$

4. Experiments

Implementation Detail: For each face image, we first align and crop the face using five point landmarks extracted with [11], the cropped image is of size 256×256 covering the whole face, as well as some background regions. The number of attribute embeddings in AEI-Net is set to $n = 8$ (Equation 1). The number of downsamples/upsamples in HEAR-Net is set to 5. Please refer to the supplemental material for more details concerning the network structure and training strategies.

The AEI-Net is trained using CelebA-HQ [17], FFHQ [18] and VGGFace [33]. While the HEAR-Net is trained using only a portion of faces that have Top-10% heuristic errors in these datasets, and with additional augmentations of synthetic occlusions. Occlusion images are randomly sampled from the EgoHands [3], GTEA Hand2K [15, 24, 23] and object renderings from ShapeNet [9].

4.1. Comparison with Previous Methods

Qualitative Comparison: We compare our method with FaceSwap [2], Nirkin *et al.* [30], DeepFakes [1] and IP-GAN [5] on the FaceForensics++ [35] test images in Figure 5. Comparison with the latest work FSGAN [29] is shown in Figure 6. We can see that, since FaceSwap, Nirkin *et al.*, DeepFakes, and FSGAN all follow the strategy that first synthesizing the inner face region then blending it into the

Figure 5: Comparison with FaceSwap [2], Nirkin *et al*. [30], DeepFakes [1], IPGAN [5] on FaceForensics++ [35] face images. Our results better preserve the face shapes of the source identities, and are also more faithful to the target attributes (*e.g*. lightings, image resolutions).

target face, as expected, they suffer from the blending inconsistency. All faces generated by these methods share exactly the same face contours with their target faces, and ignore the source face shapes (Figure 5 rows 1-4, Figure 6 rows 1-2). Besides, their results can not well respect critical informations from the target image, such as the lighting (Figure 5 row 3, Figure 6 rows 3-5), the image resolutions (Figure 5 rows 2 and 4). IPGAN [5] suffers from decreased resolutions in all samples, due to its single-level attributes representation. IPGAN cannot well preserve expression of the target face, such as the closed eyes (Figure 5 row 2).

Our method addresses all these issues well. We achieve higher fidelity by well preserving the face shapes of the source (instead of the target), and faithfully respecting the lighting and image resolution of the target (instead of the source). Our method also has the ability to go beyond FSGAN [29] to handle occlusions.

Quantitative Comparison: The experiment is constructed on FaceForensics++ [35] dataset. For FaceSwap [2] and DeepFakes [1], the test set consists of 10K face images for each method by evenly sampled 10 frames from each video clip. For IPGAN [5], Nirkin *et al*. [30] and our method, $10K$ face images are generated with the same source and target image pairs as the other methods. Then we conduct quantitative comparison with respect to three metrics: *ID retrieval*, *pose error* and *expression error*.

We extract identity vector using a different face recognition model [41] and adopt the cosine similarity to measure the identity distance. For each swapped face from the test set, we search the nearest face in all FaceForensics++ original video frames and check whether it belongs to the correct source video. The averaged accuracy of all such retrievals

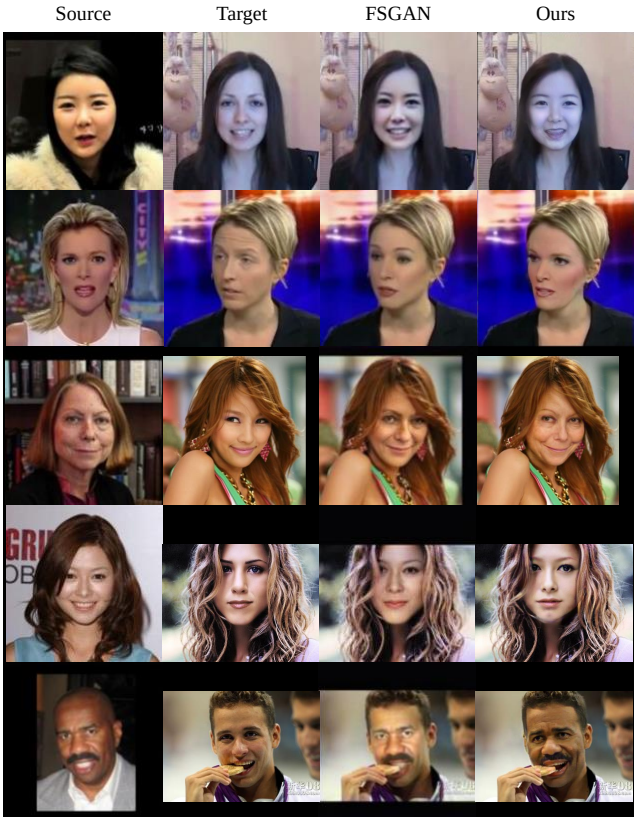

Figure 6: Comparison with FSGAN [29]. Besides the advantages in face quality and fidelity to inputs, our results preserve common occlusions as good as FSGAN. Please also refer to Figures 1, 10 and 11 for more challenging cases.

method	ID retrieval ↑	pose ↓	expression ↓
DeepFakes [1]	81.96	4.14	2.57
FaceSwap [2]	54.19	**2.51**	2.14
Nirkin *et al*. [30]	76.57	3.29	2.33
IPGAN [5]	82.41	4.04	2.50
Ours	**97.38**	2.96	**2.06**

Table 1: Comparison on FaceForensics++ videos.

is reported as the *ID retrieval* in Table 1, serving to measure identity preservation ability. Our method achieves higher *ID retrieval* score with a large margin.

We use a pose estimator [36] to estimate head pose and a 3D face model [10] to retrieve expression vectors. We report the $\mathcal{L}$-2 distances of pose and expression vectors between the swapped face and its target face in Table 1 as the *pose* and the *expression* errors. Our method is advantageous in expression preservation while comparable with others in pose preservation. We do not use the face landmark comparison as [29], since face landmarks involve identity information which should be inconsistent between the swapped face and the target face.

Human Evaluation: Three user studies are conducted to

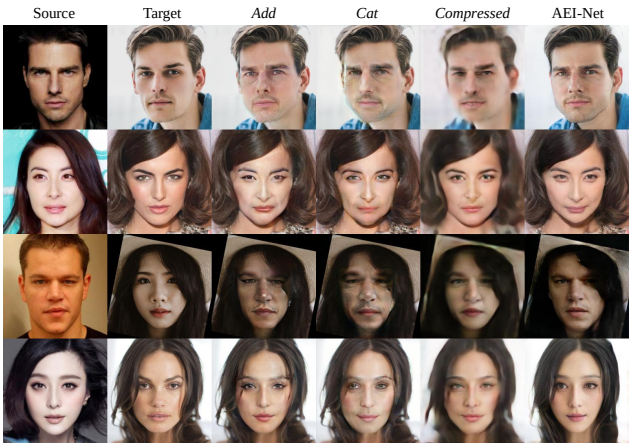

Figure 7: Comparing AEI-Net with three baseline models. The two models *Add* and *Cat* are for ablation studies of the adaptive embedding integration. The model *Compressed* is for ablating multi-level attributes representation.

method	id.	attr.	realism
DeepFakes [1]	13.7	6.8	6.1
FaceSwap [2]	12.1	23.7	6.8
Nirkin *et al*. [30]	21.3	7.4	4.2
Ours	**52.9**	**62.1**	**82.9**

Table 2: User study results. We show the averaged selection percentages of each method.

evaluate the performance of the proposed model. We let the users select: i) *the one having the most similar identity with the source face*; ii) *the one sharing the most similar head pose, face expression and scene lighting with the target image*; iii) *the most realistic one*. In each study unit, two real face images, the source and the target, and four reshuffled face swapping results generated by FaceSwap [2], Nirkin *et al*. [30], DeepFakes [1] and ours, are presented. We ask users to select one face that best matches our description.

For each user, 20 face pairs are randomly drawn from the 1K FaceForensics++ test set without duplication. Finally, we collect answers from 100 human evaluators. The averaged selection percentage for each method on each study is presented in Table 2. It shows that our model surpasses the other three methods all in large margins.

4.2. Analysis of the Framework

Adaptive Embedding Integration: To verify the necessity of adaptive integration using attentional masks, we compare AEI-Net with two baseline models: i) *Add*: element-wise plus operations is adopted in AAD layers instead of using masks $\boldsymbol{M}^k$ as in Equation 5. The output activation $\boldsymbol{h}^k_{out}$ of this model is directly calculated with $\boldsymbol{h}^k_{out} = \boldsymbol{A}^k + \boldsymbol{I}^k$; ii) *Cat*: element-wise concatenation is adopted without using masks $\boldsymbol{M}^k$. The output activation becomes $\boldsymbol{h}^k_{out} =$ $\texttt{Concat}[\boldsymbol{A}^k, \boldsymbol{I}^k]$. Results of the two baseline models, as well as the AEI-Net, are compared in Figure 7. Without a soft mask for fusing embeddings adaptively, the faces generated by baseline models are relatively blurry and contain lots of ghosting artifacts.

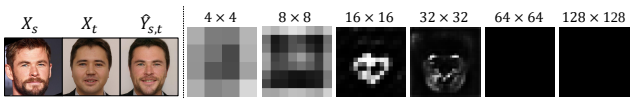

Figure 8: Visualizing attentional masks $\boldsymbol{M}^k$ of AAD layers on different feature levels. These visualizations reflect that identity embeddings are mostly effective in low and middle feature levels.

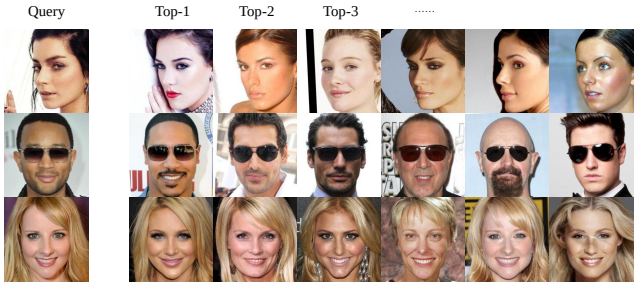

Figure 9: Query results using attributes embedding.

We also visualize the masks $\boldsymbol{M}^k$ of AAD layers on different levels in Figure 8, where a brighter pixel indicates a higher weight for identity embedding in Equation 5. It shows that the identity embedding takes more effect in low level layers. Its effective region becomes sparser in middle levels, where it activates only in some key regions that strongly relates to the face identity, such as the locations of eyes, mouth and face contours.

Multi-level Attributes: To verify whether it is necessary to extract multi-level attributes, we compare with another baseline model called *Compressed*, which shares the same network structure with AEI-Net, but only utilizes the first three level embeddings $\boldsymbol{z}^k_{att}, k = 1, 2, 3$. Its last embedding $\boldsymbol{z}^3_{att}$ is fed into all higher level AAD integrations. Its results are also compared in Figure 7. Similar to IPGAN [5], its results suffer from artifacts like blurriness, since a lot of attributes information from the target images are lost.

To understand what is encoded in the attributes embedding, we concatenate the embeddings $\boldsymbol{z}^k_{att}$ (bilinearly upsampled to 256×256 and vectorized) from all levels as a unified attribute representation. We conduct PCA to reduce vector dimensions as 512. We then perform tests querying faces from the training set with the nearest $\mathcal{L}$-2 distances of such vectors. The three results illustrated in Figure 9 verify our intention, that the attributes embeddings can well reflect face attributes, such as the head pose, hair color, expression and even the existence of sunglasses on the face. Thus it also explains why our AEI-Net sometimes can preserve occlusions like sunglasses on the target face even without a

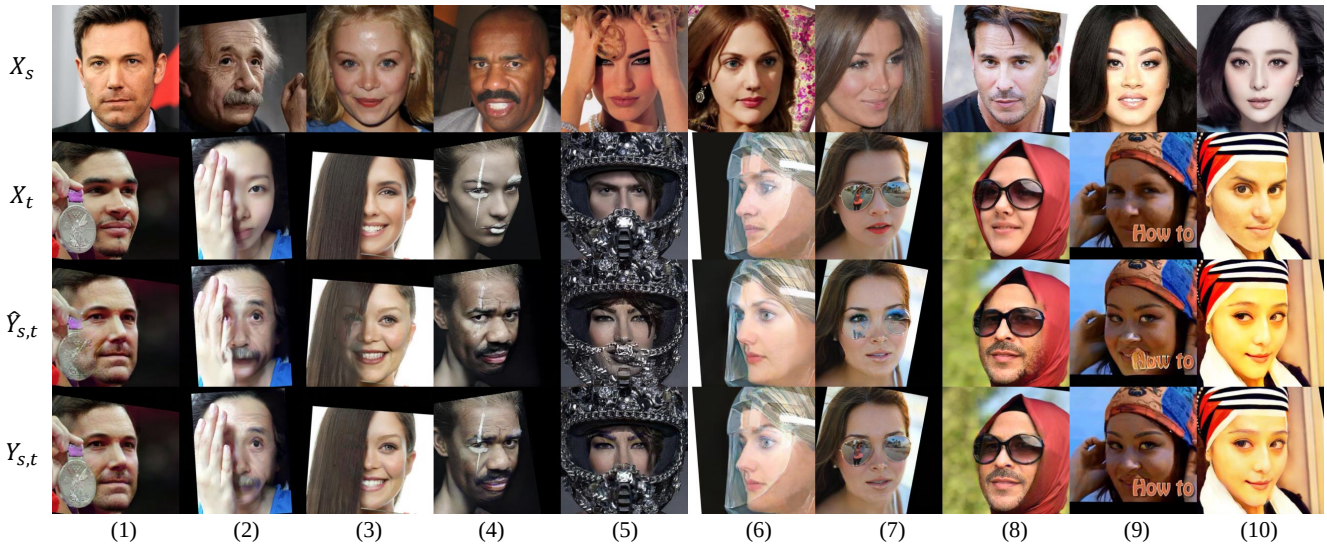

Figure 10: Second-stage refining results presenting the strong adaptability of HEAR-Net on various kinds of errors, including occlusions, reflections, slightly shifted pose and color etc.

Figure 11: Our face swapping results on wild face images under various challenging conditions. All results are generated using a single well-trained two-stage model.

second stage (Figure 10(8)).

Second Stage Refinement: Multiple samples are displayed with both one-stage results $\hat{Y}_{s,t}$ and two-stage results $Y_{s,t}$ in Figure 10. It shows that the AEI-Net is able to generate high-fidelity face swapping results, but sometimes its output $\hat{Y}_{s,t}$ does not preserve occlusions in the target. Fortunately, the HEAR-Net in the second stage is able to recover them.

The HEAR-Net can handle occlusions of various kinds, such as the medal (1), hand (2), hair (3), face painting (4), mask (5), translucent object (6), eyeglasses (7), headscarf (8) and floating text (9). Besides, it is also able to correct the color-shift that might occasionally happen in $\hat{Y}_{s,t}$ (10). Moreover, the HEAR-Net can help rectify the face shape when the target face has a very large pose (6).

4.3. More Results on Wild Faces

Finally, we demonstrate the strong capability of FaceShifter by testing on wild face images downloaded from Internet. As shown in Figure 11, our method can handle face images under various conditions, including large poses, uncommon lightings and occlusions of very challenging kinds.

5. Conclusions

In this paper, we proposed a novel framework named FaceShifter for high fidelity and occlusion aware face swapping. The AEI-Net in the first stage adaptively integrates the identity and the attributes for synthesizing high fidelity re-

sults. The HEAR-Net in the second stage recovers anomaly region in a self-supervised way without any manual annotations. The proposed framework shows superior performance in generating realistic face images given any face pairs without subject specific training. Extensive experiments demonstrate that the proposed framework significantly outperforms previous face swapping methods.

References

[1] DeepFakes. https://github.com/ondyari/FaceForensics/tree/master/dataset/DeepFakes. Accessed: 2019-09-30. 2, 3, 5, 6, 7

[2] FaceSwap. https://github.com/ondyari/FaceForensics/tree/master/dataset/FaceSwapKowalski. Accessed: 2019-09-30. 2, 5, 6, 7

[3] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1949–1957, 2015. 5, 11

[4] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2745–2754, 2017. 1

[5] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 5, 6, 7

[6] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K Nayar. Face swapping: automatically replacing faces in photographs. In *ACM Transactions on Graphics (TOG)*, volume 27, page 39. ACM, 2008. 1, 2

[7] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. In *Computer Graphics Forum*, volume 23, pages 669–676. Wiley Online Library, 2004. 1, 2, 3

[8] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999. 3

[9] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5, 11

[10] Bindita Chaudhuri, Noranart Vesdapunt, and Baoyuan Wang. Joint face detection and facial motion retargeting for multiple faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9719–9728, 2019. 6

[11] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pages 109–122. Springer, 2014. 5

[12] 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'09: Posters*, page 29. ACM, 2009. 1

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 3

[14] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 3

[15] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011. 5, 11

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 3

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 5, 11

[19] Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):163, 2018. 2

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 11

[21] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3677–3685, 2017. 1, 2

[22] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3677–3685, 2017. 3

[23] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3216–3223, 2013. 5, 11

[24] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015. 5, 11

[25] Yuan Lin, Shengjin Wang, Qian Lin, and Feng Tang. Face swapping under large pose variations: A 3d model based approach. In *2012 IEEE International Conference on Multimedia and Expo*, pages 333–338. IEEE, 2012. 1

[26] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. pagan: real-time avatars using dynamic textures. In *SIGGRAPH Asia 2018 Technical Papers*, page 258. ACM, 2018. 2, 3

[27] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Fsnet: An identity-aware generative model for image-based face swapping. In *Asian Conference on Computer Vision*, pages 117–132. Springer, 2018. 1, 2, 3

[28] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint arXiv:1804.03447*, 2018. 1, 2, 3

[29] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7184–7193, 2019. 1, 2, 3, 5, 6

[30] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 98–105. IEEE, 2018. 1, 2, 5, 6, 7

[31] Kyle Olszewski, Zimo Li, Chao Yang, Yi Zhou, Ronald Yu, Zeng Huang, Sitao Xiang, Shunsuke Saito, Pushmeet Kohli, and Hao Li. Realistic dynamic facial textures from a single image using gans. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5429–5438, 2017. 2, 3

[32] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 3, 4

[33] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, page 6, 2015. 5

[34] Arun Ross and Asem Othman. Visual cryptography for biometric privacy. *IEEE transactions on information forensics and security*, 6(1):70–81, 2010. 1

[35] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *arXiv preprint arXiv:1901.08971*, 2019. 2, 5, 6

[36] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2074–2083, 2018. 6

[37] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *arXiv preprint arXiv:1904.12356*, 2019. 2

[38] Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6):183–1, 2015. 2

[39] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 2

[40] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 3

[41] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 6

[42] Hong-Xia Wang, Chunhong Pan, Haifeng Gong, and Huai-Yu Wu. Facial image composition based on active appearance model. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 893–896. IEEE, 2008. 1, 2

A. Network Structures

Detailed structures of the AEI-Net and the HEAR-Net are given in Figure 12.

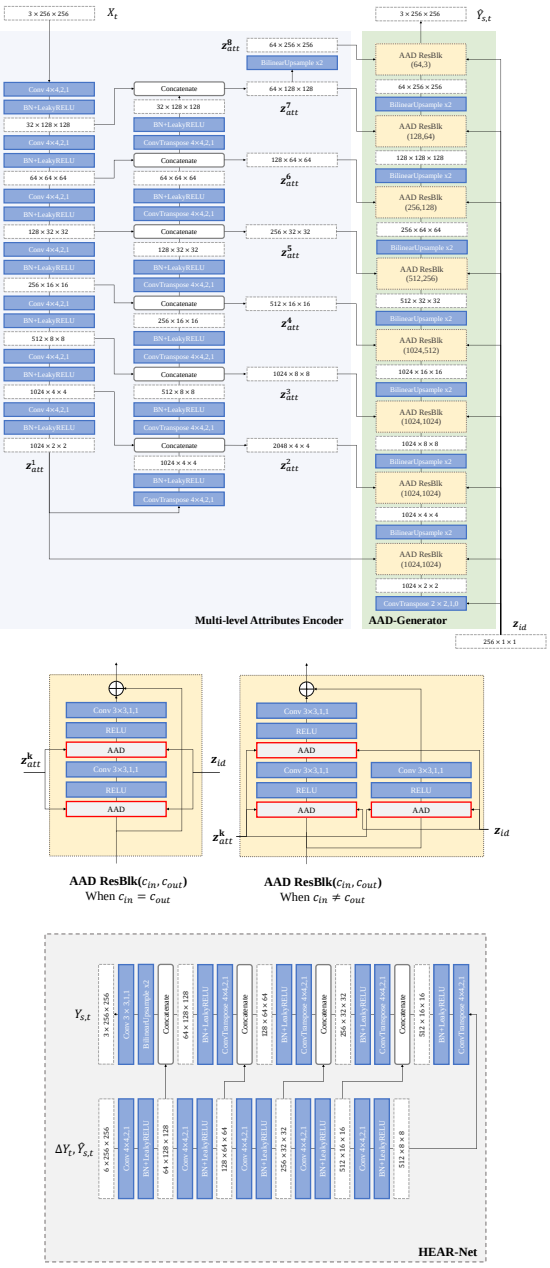

Figure 12: Network structures. *Conv k,s,p* represents a Convolutional Layer with kernel size k, stride s and padding p. *ConvTranspose k,s,p* represents a Transposed Convolutional Layer with kernel size k, stride s and padding p. All *LeakyReLU*s have $\alpha = 0.1$. *AAD ResBlk(c_{in}, c_{out})* represents an AAD ResBlk with input and output channels of c_{in} and c_{out}.

B. Training Strategies

In AEI-Net, we use the same multi-scale discriminator as [18] with the adversarial loss implemented as a hinge loss. The ratio of training samples that have $X_t = X_s$ is 80% when training the AEI-Net, it is 50% when training the HEAR-Net. ADAM [20] with $\beta_1 = 0, \beta_2 = 0.999, lr = 0.0004$ is used for training all networks. The AEI-Net is trained with 500K steps while the HEAR-Net is trained with 50K steps, both using 4 P40 GPUs with 8 images per GPU. We use synchronized mean and variance computation, i.e., these statistics are collected from all the GPUs.

Besides sampling hand images from the EgoHands [3] and GTEA Hand2K [15, 24, 23], we use a public code[1] for rendering ShapeNet [9] objects in occlusion data augmentation. Some synthetic occlusions are shown in Figure 13.

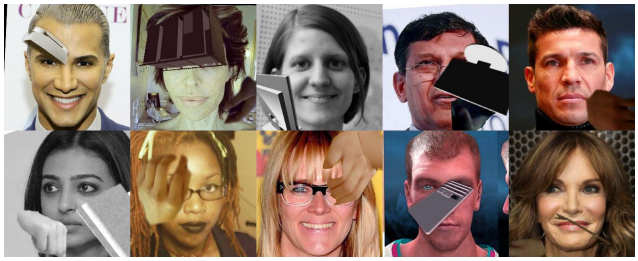

Figure 13: Augmentation with synthetic occlusions.

[1] https://github.com/panmari/stanford-shapenet-renderer