UNIVERSITÀ DI PARMA
ARCHIVIO DELLA RICERCA

University of Parma Research Repository

Face Synthesis with a Focus on Facial Attributes Translation using Attention Mechanisms

note finali coverpage

(Article begins on next page)

02 January 2025

Face Synthesis with a Focus on Facial Attributes Translation using Attention Mechanisms

Runze Li, *Student Member, IEEE*, Tomaso Fontanini, Andrea Prati, *Senior Member, IEEE* and Bir Bhanu, *Life Fellow, IEEE*

Abstract—Synthesis of face images by translating facial attributes is an important problem in computer vision and biometrics and has a wide range of applications in forensics, entertainment, etc. Recent advances in deep generative networks have made progress in synthesizing face images with certain target facial attributes. However, visualizing and interpreting generative adversarial networks (GANs) is a relatively unexplored area and generative models are still being employed as black-box tools. This paper takes the first step to visually interpret conditional GANs for facial attribute translation by using a gradient-based attention mechanism. Next, a key innovation is to include new learning objectives for knowledge distillation using attention in generative adversarial training, which result in improved synthesized face results, reduced visual confusions and boosted training for GANs in a positive way. Firstly, visual attentions are calculated to provide interpretations for GANs. Secondly, gradient-based visual attentions are used as knowledge to be distilled in a teacher-student paradigm for face synthesis with focus on facial attributes translation tasks in order to improve the performance of the model. Finally, it is shown how "pseudo"-attentions knowledge distillation can be employed during the training of face synthesis networks when teacher and student networks are trained to generate different facial attributes. The approach is validated on facial attribute translation and human expression synthesis with both qualitative and quantitative results being presented.

Index Terms—Facial attributes translation, Face image synthesis, Visual attention maps, Explainable AI, Generative adversarial network, Deep learning

I. INTRODUCTION

SYNTHESIS of facial attributes and their characterization are important for a wide range of computer vision, forensics, security, biometrics and entertainment applications. For instance, synthesizing face images can be used as a data augmentation technique in order to boost large-scale training for deep neural networks. Face synthesis with facial attributes translation can be applied in attribute-based recognition and identification, especially in surveillance and security. Besides, face synthesis can be useful in digital-art applications, e.g., paintings-related tools, like apps/softwares where users virtually customize faces with various attributes. Further, it is worth exploring how to extend face synthesis for video generation, *e.g.*, one can generate face videos where facial attributes are changing.

R. Li and B. Bhanu are with the University of California Riverside, USA (e-mail:rli047@ucr.edu; bhanu@ee.ucr.edu).

T. Fontanini and A. Prati are with the University of Parma, Italy (e-mail:tomaso.fontanini@unipr.it; andrea.prati@unipr.it).
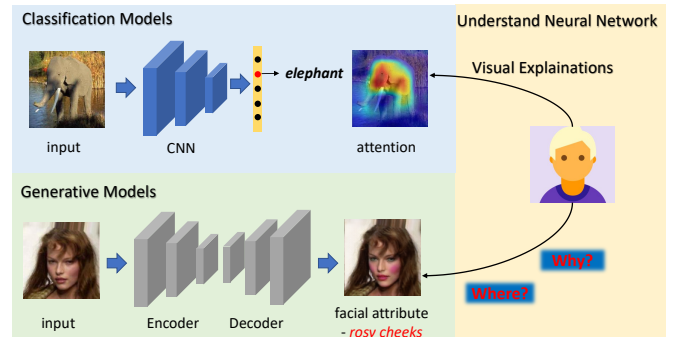
Fig. 1: Exploring interpretability of traditional CNN models (*top*) and generative models (*bottom*).

In applications of forensics, security and biometrics, face synthesis with various facial attributes can be applied for disguised and concealed identity recognition by using either synthesized face images or videos. Moreover, adversarial attacks have become hot topics in recent years and synthesizing face images with various facial attributes can be effectively utilized for dealing with adversarial attacks in deep learning models for improving their robustness of deep models. With these broad applications, we argue that translating facial attributes in face images synthesis is an under-explored problem, specifically from the perspective of visual interpretations, because it generally happens as a "black-box" process, without interpretations. However, security critical applications demand a clear understanding of the reasoning behind algorithmic processes. Consequently, there has been substantial recent interest in understanding and interpreting *how* facial attributes are translated in generative models.

Starting from classification tasks, inspired by Zeiler and Fergus [1], much effort has been dedicated in visualizing and understanding feature activations in convolutional neural networks (CNNs) by generating attention maps that highlight regions which are considered to be important for the network goals. As shown in the top stream in Fig. 1, given a trained CNN model, attention maps show how CNN responds to input images by indicating where an object is located in an image, *e.g.*, an elephant is classified correctly and highlighted visually in the attention map. CAM [2], Grad-CAM [3] and Grad-CAM++ [4] were proposed to generate this kind of visualization by means of attention maps to help us interpret why this image is classified to the *elephant* class in a more discriminative way.
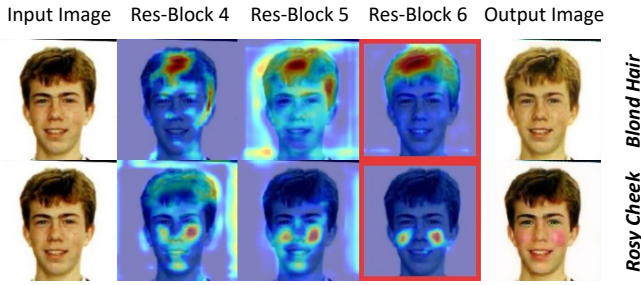
Fig. 2: Visual interpretations obtained from different residual blocks over the conditional GAN for facial attributes translation. The input to the model is a source face image with target facial attributes and the output is the translated face image where target attributes are expected to be applied on. The attention maps highlight the pixels which contribute to the output class.

Following this line of research, various researchers exploited the model explainability and explored the utilization of valuable information contained in attention maps in order to improve the performance of CNNs. Li *et al.* [5] used attention maps as weak supervised visual guidance for training CNNs and observed improvements in model generalizability in image classification and segmentation tasks. Dhar *et al.* [6] proposed an approach with attention distillation loss for incremental learning for classification tasks. Wenqian *et al.* [7] proposed a technique to visually interpret Variational Autoencoders (VAEs) and utlized attention maps for anomaly localization and disengled representation learning. However, in spite of these significant advances, enabling explainability of GANs by using visual attentions is an area that has not been fully explored yet. For example, how GANs respond to input under different conditions still remains unresolved, which results in GANs being used mainly as a black-box tool.

With the introduction of GANs and their many variants, image generation has made a gigantic forward leap in recent years, especially in terms of photo-realism. Still, often it is necessary to condition the generative models in order to have control over their outputs. This is the case of facial attribute translation, where the objective is to have a more useful representation of input face images that can be later used for various down-streaming tasks. For example, as shown in the bottom stream of Fig. 1, generative models that are able to produce highly detailed images with expected facial attributes and their outputs are almost indistinguishable from real face images. Understanding how these generative models perform such kind of tasks for a set of diverse conditions on the input is crucial for us to interpret them.

In this paper, we posit that exploiting visual interpretations in conditional GANs (cGANs) is a fundamental step in order to improve upon them. For this reason, we first study the generation of visual attention in (cGANs) for facial attribute translation by means of a gradient-based method. Facial attribute translation task requires the model to translate a input face image into different face images with specified different facial attributes using only a generator and a discriminator as a GAN system. Then, a fundamental step is taken to produce visual attentions which highlight the spatial features where the network is focusing on for a certain facial attribute and also allows to identify which layers in the generator are relatively more devoted to the facial attribute translations. Some examples of attentions are presented in Fig. 2.

Next, we focus on the utilization of these attention maps to derive a knowledge distillation module in a teacher-student paradigm and propose a novel framework called Attention Knowledge Distillation Generative Adversarial Network (**AKD-GAN**) for facial attribute translation, thus improving performance of translating target facial attributes on input face images. In other words, the teacher network suggests meaningful visual attention for each attribute, that will guide the training of the student network. Further, using attention knowledge distillation helps us in removing biases that are common in facial-attribute translation datasets (*e.g.*, "gender bias" where the selection of a certain attribute also changes the gender of the input face image).

In addition, another flexible application enabled by our proposed model is the use of so-called "pseudo"-attention maps, that are attention maps generated by the teacher from a set of facial attributes and used as weak supervisions for training a different set of facial attributes in order to help the student network to produce better face images with target attributes. This application is developed as a "pseudo"-attention knowledge distillation module in **AKD-GAN**. Extensive experiments are conducted on publicly available datasets and experimental results on two different settings demonstrate the effectiveness of the proposed model.

The main contributions of this work are as follows:

- A novel framework called **AKD-GAN** that consists of a teacher and a student network trained with attention knowledge distillation, where visual attentions are designed as full supervisions or weak supervisions to improve the performance of the student network and to remove bias in the datasets.
- An approach that enables to visually interpret conditional generative adversarial networks (cGANs) by using a gradient-based attention mechanism. In addition, the paper shows how these visual attention maps can be used for multiple purposes.
- A demonstration of the proposed method's advantages in improving facial attribute translations with extensive experiments both in distilling attention knowledge among various facial attributes and alleviating observed bias in generated face images.

The paper is organized as follows: Section II summarizes related works. Section III describes our approach to generate facial attributes translation attentions and our proposed framework for improved face synthesis with facial attributes translation; Section IV presents extensive experiments with our framework; SectionV provides conclusions of our work.

II. RELATED WORK

CNNs visual attention explanation. Deep Convolutional Networks have achieved astounding results in most computer

vision tasks and interpreting their behaviours by visualizing "where they look" when making a decision has attracted lots of interest in the past years. Following the initial work of Zeiler *et al.* [1] and Mahendran *et al.* [8], Zhou *et al.* [2] provided a method for generating *class activation mappings* (CAM) by using the global average pooling. Grad-CAM [3] and its variant Grad-CAM++ [4] were proposed by using gradients of the output score and intermediate feature maps to obtain the gradient-based class-discriminative attention maps. Compared with the response-based approaches [2], [9], [10] which introduce additional trainable units, they are applicable to a wide range of architectures without requiring any structural change in the network and without retraining the models. Recently, the concept of visual attention was also extended to GANs [11], [12], [13]; However, these papers mainly studied self-attention modules which required a large number of additional training units consisting of a series of convolutional layers with 1×1 kernel size. In particular, MU-GAN [12] introduced an additive attention mechanism to build attention-based U-Net connections and a self-attention mechanism in the convolutional layers. In addition, Kim *et al.* [14] calculated attention in the discriminator of a GAN in order to use it as a mask to preserve the attribute-irrelevant regions. As compared to this work, we take the first step into visualizing and employing attention that is calculated from the GAN generator to directly improve the performance of face images synthesis.

Knowledge distillation in neural networks. Knowledge distillation involves transferring knowledge from a more complex network (teacher) to a simpler and lighter network (student) sharing the same task [15]. The goal is to have the student to reach almost the same results as the teacher. Many techniques have been developed in this area [16], [17], [18], with [19] being the first to use attention transfer to improve the performance of a student classification network. Previous methods were almost entirely used for recognition or classification models, while [20] introduced a method working on unconditional GANs. Recently, Li *et al.* [21] proposed a compression algorithm for cGANs using feature distillation and neural architecture search.

Conditional GANs for facial attribute translation. Generative Adversarial Networks (GANs) [22], [23], [24], in their many variations represent the state-of-the-art for photo-realistic image synthesis nowadays. In particular, when a much finer control over the output is required, conditional GANs (cGANs) [25] allow the generation of images from text [26], [27], [28], class labels [29], [30] for natural images, sketches or rich textures [31], [32], [33], [34], [35], [30]. Besides, Di *et al.* [36] studied to synthesize attribute-preserved visible face images from thermal imagery for cross-modal matching. Furthermore, while initially a paired dataset was required [37], CycleGAN [38] showed that a conditional GAN can be successfully trained in an unpaired way. Another relevant feature that most cGANs lack is the ability of producing images belonging to different classes or domains using a single architecture. Some models [39], [40] achieve this by using adaptive instance normalization layers [41] combined with a class-specific encoder and a content-specific encoder whereas
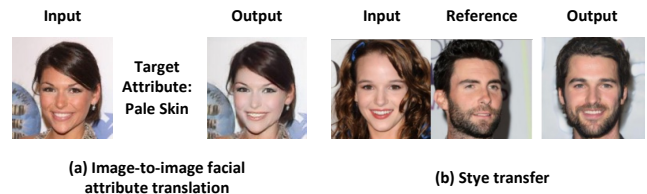

Fig. 3: Examples of image-to-image facial attribute translation (a) and style transfer (b).

StarGAN [42] and its variants [43], [44], [45], [46] take as input both an image and the target label learning to flexibly execute the translation using only one underlying representation. Unlike the above frameworks, models like StarGANv2 [47] focus on different tasks and share different principles in network design. StarGANv2 designs the architecture by heavily relying on using a mapping network to learn multiple styles and utilizing Adaptive Instance Normalization (AdaIN) layers [41] that can only perform global editing over the input images. Thus, StarGANv2 performs more of a style transfer (like most of AdaIN-based methods) than attributes translation as we focus in this paper. We argue that the purpose of StarGANv2 is significantly different from ours. To perform style transfers, the model is trained to take an image to be transferred and a reference image as the input, and outputs a new image by transferring features of reference image over the image to be transferred, and the output image could be completely different in identities. Fig. 3 presents a comparison of results of image-to-image facial attributes translation and style transfer. Fig. 3(a) shows an example of facial attributes translation on a face image, where the identity of the face and the appearance of the face are maintained, but only facial attributes are changed. Fig. 3(b) shows an example of style transfer, where the identity and almost the entire image have been changed in the output. Indeed, when transferring styles, no single facial attribute is changed, but the style and appearance of the entire image have changed and only the pose of the source image is maintained. In this work, we focus on image-to-image translation for facial attributes translation. Finally, cGANs can also be combined with meta-learning for greater flexibility and robustness [48].

To bridge the gap of model interpretations of cGANs for synthesizing face images with facial attributes, this work firstly generate visual attention as interpretations for cGANs, and then propose a new framework to utilize visual attention to distill attention knowledge in a teacher-student paradigm to improve the face synthesis performance of the student network. The motivation comes from our observations that attention maps can highlight spatial regions where the target attributes would be applied (examples in Fig. 2), and these attention maps can in turn serve for knowledge-based guidance to further improve model performance. Firstly, we study generating visual attention for facial attributes translation generative adversarial models, *i.e.*, conditional GANs in our framework which has not been explored yet. Secondly, we propose our framework with attention knowledge distillation in a teacher-
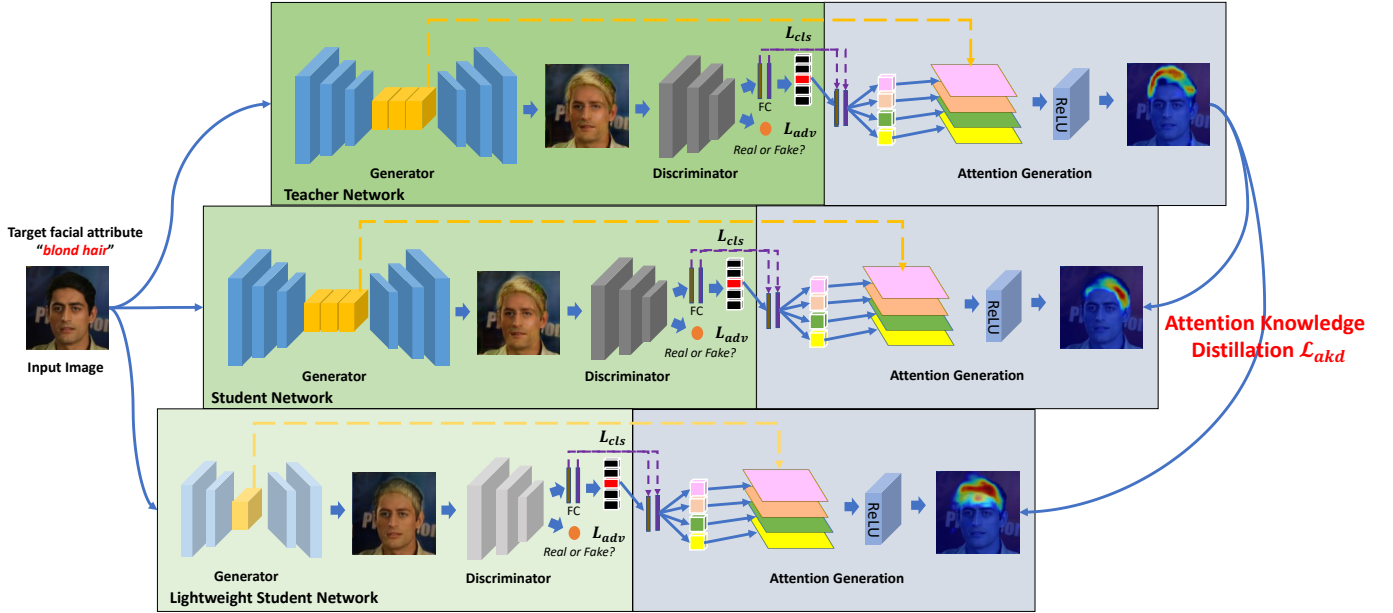
Fig. 4: Summary of the **AKD-GAN** workflow: the *Teacher **T*** network and the *Student **S*** network represent the conditional GANs for facial attribute translation. The *Lightweight Student* (***Lite-S***) network is a lighter student network. During training, our proposed attention distillation loss $\mathcal{L}_{akd}$ is calculated using the attention maps obtained from the teacher and the student or the lightweight student network using the method described in III-A.

student paradigm for facial attributes translation and conduct thorough experiments on CelebA [49] and RaFD [50] datasets, establishing improved facial attributes translation performance under extensive experimental settings. Finally, we study how these attention visualizations can help distilling knowledge among *different* facial attributes in our "pseudo"-attention knowledge distillation experiments, providing the flexibility of our proposed attention knowledge distillation module in integrating with generative adversarial networks training.

III. TECHNICAL APPROACH

We propose an end-to-end framework, **AKD-GAN** (see Fig. 4), to improve model performance of a student network and also of a lightweight student network for facial attributes translation via gradient-based attention maps as guidance. The main idea is to produce visual attention, for facial attribute translation that provides supervisory signals for the proposed attention knowledge distillation process.

We first introduce the backbone network in our **AKD-GAN** and basic training objectives of the generator and the discriminator in the following paragraphs. Then we streamline the attention generation pipeline in Section III-A and the related experimental results are shown in Section IV-D. Next, we describe the proposed attention knowledge distillation loss in Section III-B. The corresponding experiments and discussions are given in Section IV-E, Section IV-G and Section IV-H. In addition, we introduce the proposed "pseudo"-attention knowledge distillation loss in Section III-C and the corresponding experiments and discussions are carried out in Section IV-F.

A conditional generative adversarial network [25] extends GANs by adding a condition as the input to the generator and discriminator. Conditions act as prior information for the GAN to generate data, and such conditions can be in various formats, like latent vectors, images, language priors, etc. The auxiliary classifier GANs [51] further extend a classification stream in the discriminator.

The basic system of our **AKD-GAN** is composed of a teacher ***T*** network and a student ***S*** network which is the one to be optimized and evaluated. Both of them are conditional GANs for facial attributes translation and are conditioned by the label of the target facial attribute that we want to translate over the input face image to obtain a new output face image. The generator takes as input a face image and the target facial attribute and returns the translated face image with target facial attributes applied, while the discriminator takes the translated face image as input and returns an adversarial output and a facial attribute classification output.

The facial attribute generator is composed of an encoder, followed by a group of residual blocks and a decoder. The facial attribute translation of the generator can be written as:

$$x_{fake} = Gen_{fa}(x_{real}, c) = Dec(Enc(x_{real}, c)) \quad (1)$$

where x_{real} is the input face image, c is the target facial attribute for translation, Gen_{fa} is the facial attribute generator containing the encoder Enc and decoder Dec and x_{fake} is the generated (fake) face image with target facial attributes.

The facial attribute discriminator Dis_{fa} consists of a group of convolutional layers and two output streams: one is the

adversarial output telling how realistic the generated face image is by distinguishing it as real or fake, and the other one is auxiliary classification output for calculating the facial attribute classification loss. The overall learning objectives to train the student network of **AKD-GAN** for facial attribute translation follows the one of StarGAN [42] and can be written as:

$$\mathcal{L}_{Dis_{fa}} = \mathcal{L}_{adv} + \mathcal{L}_{cls} \quad (2)$$

$$\mathcal{L}_{Gen_{fa}} = \mathcal{L}_{adv} + \mathcal{L}_{cls} + \mathcal{L}_{rec} \quad (3)$$

for the discriminator and the generator. The generator is trained using adversarial loss $\mathcal{L}_{adv}$, classification loss $\mathcal{L}_{cls}^{Gen_{fa}}$ and reconstruction loss $\mathcal{L}_{rec}$, while the discriminator is trained with adversarial loss $\mathcal{L}_{adv}$ and classification loss $\mathcal{L}_{cls}^{Dis_{fa}}$.

A. Attention Generation for Facial Attributes Translation

Inspired by the fundamental framework of Grad-CAM [3], we streamlined the generation of attention map on either the student network or the teacher network for facial attribute translation. An attention map corresponding to the input face images can be obtained within each inference step so that it can be employed during training stage. Given an input face image $x \in \mathbf{x_{real}}$ and a set of target facial attributes $\mathbf{c}$, for each class $c \in \mathbf{c}$, from the ground-truth labels of target facial attributes, we compute the gradient of score y^c corresponding to the class c. We backpropagate the gradients directly from the classification output of the discriminator to the convolutional layers of the generator with feature maps $\mathbf{F} \in \mathbb{R}^{n \times h \times w}$, with n, h and w being number of channels, height and width of the feature map, respectively, obtaining facial attributes attention maps $\mathbf{A}^{fa}$ corresponding to y^c. Indeed, this represent a significant difference with respect to an ordinary classification network (the typical setup where Grad-CAM operates) where, in order to get attention, the gradients are backpropagated only to the layer before the classification output. Specifically, we calculate $\mathbf{A}^{fa}$ by using the following equation:

$$\mathbf{A}^{fa} = ReLU\left(\sum_{k=1}^{n} \alpha_k^c \mathbf{F}_k\right) \quad (4)$$

where the scalar $\alpha_k^c = \text{GAP}\left(\frac{\partial y^c}{\partial \mathbf{F}_k}\right)$ and $\mathbf{F}_k$ is the kth feature channel ($k = 1, \ldots, n$) of the feature maps $\mathbf{F}$, with $\frac{\partial y^c}{\partial \mathbf{F}^k}$ representing the gradient of the score y^c with respect to the feature maps $\mathbf{F}^k$. The global average pooling (GAP) operation is used to obtain scalar α_k^c as:

$$\alpha_k^c = \frac{1}{S} \sum_{m=1}^{h} \sum_{n=1}^{w} \left(\frac{\partial y^c}{\partial F_k^{mn}}\right) \quad (5)$$

where $S = h \times w$ and F_k^{mn} is the pixel value at location (m, n) of the $h \times w$ matrix $\mathbf{F}_k$. The attention map generation process is illustrated in Fig. 5.

Note that we took this conditional GAN for facial attributes translation as the main case study in our work, but this pipeline to generate attention maps can be applied to a wider variety of GANs, and simply requires a generator-discriminator structure with auxiliary streams integrated in the discriminator.

Example attention results are shown in Fig. 2. It can be observed that visual attention reveal how the conditional generative models perform translations on the input to generate output in a more transparent way. Particularly, these results confirm that we can use these attention maps for knowledge distillation, either in a self-supervised or weakly-supervised manner, to train our **AKG-GAN** for improved facial attributes translations, which will be introduced in following sub-sections.

B. Attention Knowledge Distillation Loss

We use the notion of attention knowledge distillation as the main part of our learning process for distilling the knowledge encoded in visual attention from the teacher to the student and propose a new learning objective $\mathcal{L}_{akd}$ (attention knowledge distillation loss, denoted as akd loss). Essentially, given the input face images x and target facial attribute c, the attention maps of the facial attribute class are computed from the teacher network and student network using the method introduced in Sect. III-A, as $\mathbf{A}_T^{fa}(x, c)$ and $\mathbf{A}_S^{fa}(x, c)$, respectively. We enforce the two attentions to be consistent (*i.e.*, the student attention must imitate the teacher one) and integrate this constraint during the training. To this end, we propose a loss function $\mathcal{L}_{akd}$ which is defined as:

$$\mathcal{L}_{akd} = \mathbb{E}_{x,c}\left[\parallel \mathbf{A}_T^{fa}(x, c) - \mathbf{A}_S^{fa}(x, c) \parallel\right] \quad (6)$$

The proposed loss is differentiable which means that it can be directly used for model training without introducing additional training units.

The **AKD-GAN** workflow is illustrated in Fig. 4, with $\mathcal{L}_{akd}$ integrated in different teacher-student knowledge distillation designs: the first one is teacher-student training while the second one is teacher-lightweight-student training. In addition, during each training step of the student network, we trained the student discriminator to distinguish between real and fake face images and to correctly classify the face images into multiple facial attributes. The generator is trained to fool the discriminator by producing better face images belonging to the target facial attributes. Finally, the full training objective for student network is:

$$\mathcal{L}_{Dis_{fa}}^S = \mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls} \quad (7)$$

$$\mathcal{L}_{Gen_{fa}}^S = \mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{akd}\mathcal{L}_{akd} \quad (8)$$

where $\mathcal{L}_{Dis_{fa}}^S$ is the loss function that we use to optimize the discriminator of the student network and $\mathcal{L}_{Gen_{fa}}^S$ is the loss function used to optimize the generator of the student network. During the training, only parameters of the student network are optimized and the discriminator and the generator are optimized jointly. Following StarGAN [42], we use $\lambda_{cls} = 1$, $\lambda_{rec} = 10$ in all our experiments. For λ_{akd}, we use different values and empirically set it as 10 which gives us the best results in the preliminary experiments for face synthesis with 5 facial attributes.

The intuition of $\mathcal{L}_{akd}$ is that knowledge involved in visual attentions can be distilled as supervisory signals via a teacher-student paradigm so that training of the student can be boosted.
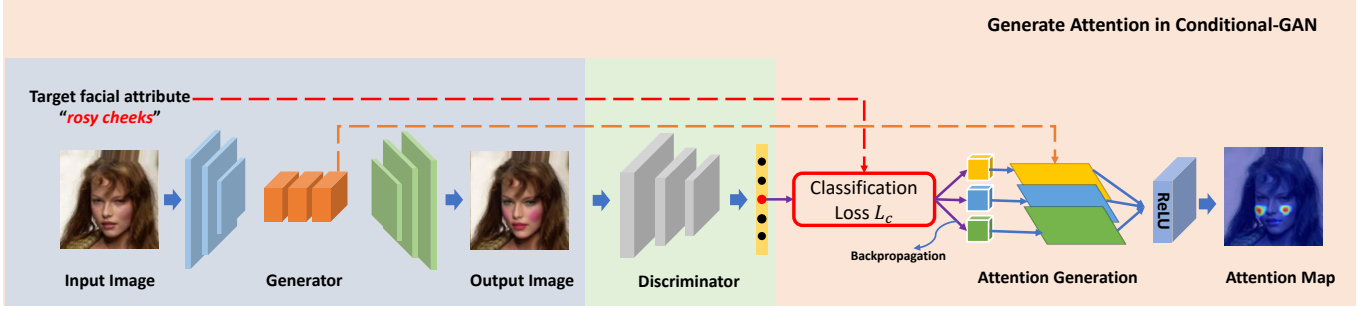
Fig. 5: Attention generation with conditional GANs for facial attribute translation.

Especially for those facial attributes that need to be translated over a small region of face images, by using $\mathcal{L}_{akd}$, our expectation is that it can help to prevent noise from other face areas so that the model is pushed to focus only on the region where the facial attribute needs to be translated.

C. Pseudo-Attention Knowledge Distillation Loss as Weak Supervision

In the previous section, we discussed how we have generated attention maps as interpretations and designed the attention distillation loss integrated with the adversarial objectives to train the model for facial attributes translation. In this section, we will discuss how to distill knowledge from a teacher T network translating input face images with a set of facial attributes to a student S network that works on a *new* set of *different* facial attributes. Our intuition is that when the input face images are translated to a different target facial attribute, the regions "looked at" by the model in the input images might be shared through different facial attributes. For example, in order to translate an input face image to the output images with specified facial attributes such as *black hair*, *blond hair* or *brown hair*, the network is expected to pay more attention to and edit the region corresponding to the facial attribute ***"hair"*** in the input image so that it can perform translations in "color". Given these observations, in this section, we start from generating the "pseudo"-attention maps for facial attribute translations and present how to design "pseudo"-attention knowledge distillation loss. Finally, we demonstrate the advantages of using "pseudo"-attentions in designing weak supervisory signals which can be integrated flexibly in training a student network with **AKD-GAN**.

*1) Training **AKD-GAN** with Pseudo-Attention Knowledge Distillation:* The objective is to distill knowledge using "pseudo"-attentions as weak supervisions.

Firstly, given a set of target facial attributes **c**, we identified a second set of facial attributes $\mathbf{c^{pse}}$ where $c^{pse} \in \mathbf{c^{pse}}$ would share spatial features on the face with a facial attribute $c \in \mathbf{c}$. Next, we trained a teacher $\mathbf{T_{pse}}$ network using the *different* facial attributes $\mathbf{c^{pse}}$. Finally, we defined a "pseudo" attention knowledge distillation module for training **AKD-GAN** using the teacher-student paradigm.

While training the student network using **c** as target facial attributes, we generated attention maps and proposed

the "pseudo" attention knowledge distillation loss to distill knowledge from the teacher network to the student network. Specifically, given an input face image $x \in \mathbf{x}$ and a target facial attribute $c \in \mathbf{c}$, we generated the attention maps using the method in Section III-A. For the teacher network trained with $\mathbf{c_{pse}}$, the class score $y^{c^{pse}}$ is backpropagated as usual to calculate the facial attribute attention map as $\mathbf{A}_{\mathbf{T}_{pse}}^{fa,c^{pse}} = ReLU\left(\sum_{k=1}^{n}\alpha_k^{c^{pse}}\mathbf{F}_k\right)$. For the student network trained to translate to different facial attributes **c**, the attention for the student network is generated as $\mathbf{A_S}^{fa,c}$.

Thus, "pseudo"-attention maps are obtained as $\mathbf{A}_{\mathbf{T}_{pse}}^{fa,c^{pse}}$ from the teacher network and $\mathbf{A_S}^{fa,c}$ from the student network, respectively. Since our goal is to train the student network to translate input face images to target facial attributes **c**, while the supervisions via the knowledge distillation are defined with different facial attributes $\mathbf{c^{pse}}$, we call the attentions obtained in this way "pseudo"-attentions. Finally, the attention knowledge distillation loss $\mathcal{L}_{akd}^{pse}$ employed to train **AKD-GAN** with "pseudo"-attentions is calculated as:

$$\mathcal{L}_{akd}^{pse} = \mathbb{E}_{x,c}[\| \mathbf{A}_{\mathbf{T}_{pse}}^{fa,c^{pse}}(x, c^{pse}) - \mathbf{A}_{\mathbf{S}}^{fa,c}(x,c) \|] \quad (9)$$

The training objective of discriminator is the same as Equation 2 and the objective of generator in student network is:

$$\mathcal{L}_{Gen_{fa}}^{S} = \mathcal{L}_{adv} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{akd}^{pse}\mathcal{L}_{akd}^{pse} \quad (10)$$

TABLE I: Number of parameters of different generators.

AKD-GAN Model Design	Number of parameters
Teacher (T), Student (S)	8.4 Million
Lightweight Student (***Lite-S***)	1.2 Million

IV. EXPERIMENTS AND RESULTS

A. Experimental Settings

We present experiments using two different settings to train the **AKD-GAN**: **(a)** we train the student network to translate a set of facial attributes and distill knowledge using the attention maps calculated from a teacher network trained on the same set of attributes, **(b)** we train the student network to translate a
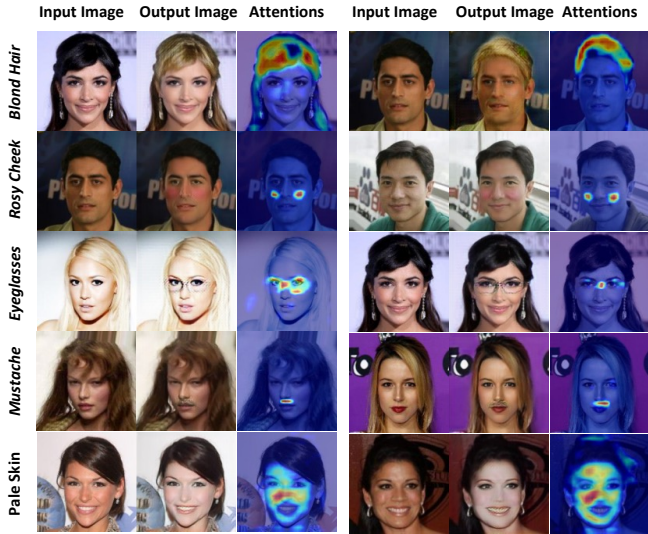
Fig. 6: Attention maps used by the proposed attention knowledge distillation for facial attributes translation.

set of facial attributes and distill knowledge using the pseudo-attention maps calculated from a teacher network trained on a set of different facial attributes.

In each experimental setting, firstly we train the teacher network which shares the same architecture design as StarGAN. Then, we train the student by distilling attention knowledge from the last residual block of the pre-trained teacher network to the last residual block of student networks in all experiments. More in detail, we experimented with student networks having different model complexity: **(a)** a student network (*S*) that shares the same architecture design as the teacher network, **(b)** a lightweight student (*Lite-S*) network which is a lighter and pruned network.

In the lightweight student (*Lite-S*) network, we reduced feature map dimensions and the number of residual blocks, obtaining a much lighter network: the number of feature maps in each layer is one half that of each layer of the teacher generator (*e.g.*, from 64 to 32 feature maps in the first convolutional layer of of (*Lite-S*)). Besides, there are only 3 residual blocks instead of 6. Table I shows the number of parameters of generators in the teacher (*T*), the student network (*S*) and the lightweight student network (*Lite-S*).

In all experiments, only the parameters of student network are optimized during the training and only the student network is utilized during evaluation.

B. Datasets

CelebA: The core experiments were performed using the CelebA dataset [49], which is a large-scale facial attributes dataset with more than 200,000 images and 40 attributes. Facial attributes were also well suited to prove the efficacy of our system, since, in order to correctly translate the input face images to outputs, the network needs to learn spatial information on the human face precisely and visual attentions

can acquire this kind of information, especially for small, localized attribute.

RaFD: This dataset [50] includes 8,000 images divided in 8 emotional expressions.

Following [42], in all experiments, input images are cropped and resized to 128×128.

C. Metrics

The classification accuracy of translation is used as the main evaluation metric to evaluate the performance of translating target facial attributes over input face images. To elaborate, we use the training set of each dataset to train a deep attribute classification model (ResNet50 [52]) for the facial attributes to be translated and take the generated face image with target facial attributes as the input to the classifier and calculate classification results. It is noted that the classifier we used for evaluation only saw training samples and never saw testing samples. On the **CelebA** dataset, the classification model achieves an accuracy of 91.3% for all facial attributes on the test set. On the **RaFD** dataset, the classification model achieves an accuracy of 98.9% for all facial expressions. In addition, we resort to use another alternative measure Fréchet Inception Distance [53] to evaluate the image quality.

D. Visualizing Attentions in Conditional GANs

Examples of visual attention maps for facial attributes translation on the **CelebA** dataset are shown in Fig. 6. The generated attention maps are utilized to develop our attention knowledge distillation loss (see Section III-B and Equation 6) and "pseudo"-attention knowledge distillation loss (see Section III-C and Equation 9) for face synthesis with facial attributes translation. As an example of a comparison, we generate attention maps from style transfer models (see Fig. 1 in the supplementary materials) which target on different tasks from ours and these differences are discussed in Section II. In Fig. 6, each triplet of images consists of an input face image, a translated face image and the corresponding attention maps. These attention maps indicate spatial features where the conditional generative models focus on when performing the facial attribute translation. For example, "face" is highlighted precisely when the facial attribute *Pale Skin* is used as the target for translation. In addition, for small attributes, like *Mustache* and *Eyeglasses*, attention maps can localize their spatial areas over the input face images accurately.

It is observed that there are green color artifacts in the generated face images (see Fig. 6). They represent a limitation of the state-of-the-art networks in generating realistic texture. We adopted the network designs of StarGAN and we visualized interpretations with attention maps that are obtained directly from StarGAN's generated images. They may be caused by the group of deconvolutional layers in the decoder of the generator. Another reason can lie in the low-resolution data used for this task. Face images are 128×128 and contain rich facial attributes. This poses challenges in synthesizing face images with certain facial attributes translated while maintaining the rest of the face untouched. In the future work, a solution to limit these artifacts would be to replace deconvolutional

TABLE II: Quantitative comparisons in classification accuracy between the proposed method (**AKD-GAN**) and state-of-art methods for generated face images over various facial attributes. The bold results represent the best results. Inference in AKD-GAN is performed using only the generator of the student network.

Methods	Interpretations	Blond Hair ↑	Goatee ↑	Eyeglasses ↑	Heavy Makeup ↑	Pale Skin ↑
ACGAN [51]	✗	73.2%	53.2%	95.6%	60.8%	83%
RelGAN [45]	✗	57.31%	65.35%	97.87%	47.54%	52.56%
AttGAN [43]	✗	35.53%	56.90%	98.23%	55.63%	80.22%
STGAN [46]	✗	75.38%	68.22%	95.83%	34.44%	71.78%
StarGAN [42](baseline)	✗	80.94%	64.54%	99.10%	86.27%	79.66%
StarGAN [42](double iterations)	✗	83.68%	71.22%	98.82%	65.57%	83.06%
AKD-GAN (Ours)	✓	**86.02%**	**74.45%**	**99.48%**	**91.47%**	**88.32%**
	Interpretations	Brown Hair ↑	Bushy Eyebrows ↑	Wear Hat ↑	Wear Lipstick ↑	Pointy Nose ↑
ACGAN [51]	✗	76.5%	90.2%	74.4%	61.9%	77.6%
RelGAN [45]	✗	35.20%	62.86%	48.78%	49.29%	20.68%
AttGAN [43]	✗	48.38%	72.37%	20.23%	65.35%	50.99%
STGAN [46]	✗	87.33%	88.25%	35.61%	41.09%	59.61%
StarGAN [42](baseline)	✗	75.10%	95.60%	79.66%	95.92%	82.44%
StarGAN[42](double iterations)	✗	75.22%	87.30%	92.89%	98.00%	88.24%
AKD-GAN (Ours)	✓	**92.71%**	**97.04%**	**92.89%**	**98.00%**	**88.24%**

layers with convolutional layers and interpolation operations. Another empirical solution would be to design the model and training schemes by introducing ideas of the progressive GANs [54].

E. Attention Knowledge Distillation for Face Synthesis with Facial Attributes Translation

We first conducted experiments by training the proposed **AKD-GAN** with a standard teacher-student paradigm, expecting improved performance of student network on facial attributes translation. Moreover, we trained our **AKD-GAN** by using a light-weight student network where model complexity is reduced significantly, to further evaluate the capability of the attention knowledge distillation module in improving an even smaller model performance. Experiments are conducted on the **CelebA** dataset.

1) AKD-GAN with Teacher and Student Networks: In the first set of experiments the objective is to use visual attention to distill knowledge from a teacher T network to a student S network (see top two rows in Fig. 4) to improve the performance of student model for facial attribute translation.

We trained the **AKD-GAN** to translate various facial attributes including *blond hair*, *goatee*, *eyeglasses*, *pale skin*, etc. We present both quantitative and qualitative results in the following paragraphs. As introduced in Sec. IV-A, our AKD-GAN is built upon StarGAN which, for this reason, is the baseline method. We also tested and compared results with ACGAN [51], AttGAN [43] RelGAN [45] and STGAN [46]. **Quantitative Results.** Quantitative comparisons of facial attributes classification accuracy are shown in Table II. We used the synthesised face images outputted from the generator in the student network to calculate the classification accuracy of translated images of each facial attribute to evaluate the

TABLE III: Quantitative comparisons in FID score between the proposed **AKD-GAN** and state-of-the-art methods for generated face images over various facial attributes. The bold and underlined results represent the best and the second best results respectively.

Methods	Interpretations	Blond Hair ↓	Goatee ↓	Heavy Makeup ↓	Brown Hair ↓
ACGAN [51]	✗	42.77	71.29	33.62	29.19
RelGAN [45]	✗	**29.04**	**28.34**	40.25	**18.1**
AttGAN [43]	✗	36.13	61.48	<u>29.27</u>	<u>19.13</u>
STGAN [46]	✗	34.5	<u>49.04</u>	36.8	19.53
StarGAN [42] (baseline)	✗	34.87	52.89	62.44	23.51
AKD-GAN (Ours)	✓	<u>30.03</u>	50.87	**21.72**	19.63
Methods	Interpretations	Bushy Eyebrows ↓	Wear Hat ↓	Wear Lipstick ↓	Pointy Nose ↓
ACGAN [51]	✗	33.78	100.47	33.95	27.93
RelGAN [45]	✗	<u>25.52</u>	90.18	34.17	<u>18.53</u>
AttGAN [43]	✗	27	100.15	25.43	20.83
STGAN [46]	✗	**23.35**	118.65	36.06	**16.83**
StarGAN [42] (baseline)	✗	30.92	90.337	<u>23.81</u>	21.53
AKD-GAN (Ours)	✓	28.36	**79.24**	**21.29**	19.48

quality of translations that are made to the input face images. From the Table II, it can be observed that the proposed **AKD-GAN** model trained with attention knowledge distillation loss outperforms all other methods in translating facial attributes in terms of classification accuracy, which proves the effectiveness of our model in translating facial attributes over input faces.

Next, we calculated the FID score using translated face images of each attribute and the results are shown in Table III. Compared with the baseline method, StarGAN, our framework can consistently obtain better quality of face images after translating facial attributes over inputs, which validates that visual attention can help to improve facial attributes

Fig. 7: Comparisons of qualitative results between the baseline method (StarGAN [42]) and the proposed **AKD-GAN**.

Fig. 8: Comparisons of qualitative results between the STGAN [46] and the proposed **AKD-GAN**.

translations as well as image generation. Compared with the state-of-the-art methods, our method can achieve competitive performance in synthesizing face images with target facial attributes. In particular, for facial attributes *heavy_makeup*, *wear_hat* and *wear_lipstick*, our method can obtain the best quality of generated face images by large margins.

Given the above observations, it is shown that our method with knowledge distillation loss, which is derived from attention interpretations, can achieve high performance in trans-

lating facial attributes over the face images. Meanwhile, our method can generate face images by translating facial attributes with competitive performance.

Qualitative Results. Collections of qualitative results between the proposed framework and the baseline method StarGAN are shown in Fig. 7. Each triplet presents the input face image, the generated result from StarGAN [42] and the result from the proposed method, **AKD-GAN**. Indeed, in face images that are generated from the proposed AKD-GAN, target facial attributes are applied much more strongly over the input. In addition, for the facial attributes that require precise spatial localizations for translations, (*e.g. eyeglasses* and *smiling*), **AKD-GAN** gives more convincing results. Furthermore, it is observed that results from the **AKD-GAN** are less noisy and with fewer artifacts. For example, in the first triplet of the first row of Fig. 7, the attribute *blond hair* is applied very convincingly while still maintaining the original face characteristics; in the second triplet of the last row of Fig. 7, the attribute *pale skin* is applied well with fewer artifacts.

In addition, Fig. 8 presents collections of qualitative comparisons between the proposed method and the STGAN [46]. It can be observed that our method generates face images with target facial attributes applied better than STGAN, in particular, for facial attributes *mustache*, *eyeglasses*, etc.

2) Empirical Observations on Bias Removal in Facial Attributes Translation: An observation that often occurs in facial attributes translation given a certain dataset, *e.g.*, CelebA, is to find some kind of correlation between attributes (a phenomenon often increased by datasets bias). This means that when translating a certain attribute, other undesired attributes will be translated as well. A clear example is represented by
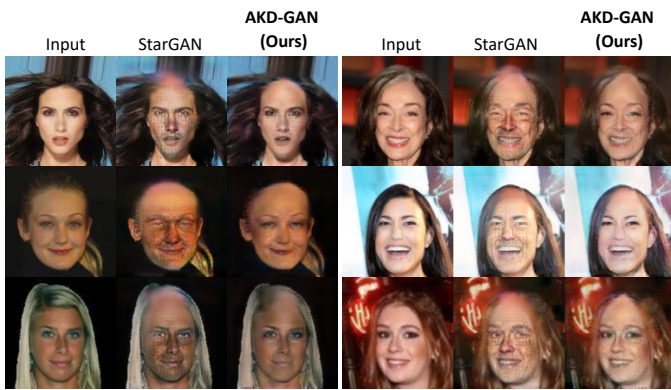
Fig. 9: Generated face images of the facial attribute *Bald* using StarGAN and our method **AKD-GAN**. AKD-GAN does not change the gender of the input image.

TABLE IV: Percentage of gender **misclassification** for generated face images with target attributes *Bald* and *Lipstick*.

Methods	Female misclassified as Male (Attribute: Bald) ↓	Male misclassified as Female (Attribute: Lipstick) ↓
StarGAN (Baseline)	97.72%	88.11%
AKD-GAN (Ours)	**50.97%**	**83.60%**

TABLE V: Distributions of facial attributes *Bald* and *Wearing Lipstick* in the CelebA dataset.

|	Bald	Lipstick		Male	Female
% in CelebA	2.2%	47.24%	% in CelebA	41.6%	58.4%
Male/ Female	99.6%/0.4%	99.4%/0.6%	Bald/ Lipstick	99.6%/0.4%	99.4%/0.6%

Fig. 10: Collection of comparisons between qualitative results obtained from the lightweight student (***Lite-S***) without using the loss $\mathcal{L}_{akd}$ and results obtained when training (***Lite-S***) with the $\mathcal{L}_{akd}$ proposed in **AKD-GAN**.

the *Bald* attribute which can only be found in face images of the gender-"male" in the dataset. This means that a facial attribute translation model (like StarGAN) will likely turn each input image into a face image with the gender-"male" when applying the *Bald* facial attribute since it has learned to correlate that attribute with a specific gender.

None of existing works [42], [43], [45], [46] has discussed this phenomenon, while, we argue that it is an important problem and attention should be paid to it. We conduct experiments to evaluate the model performance in mitigating biases on facial attribute translations. Firstly, we use face images in the training set to fine-tune a classifier for face images with attributes *Male* and *Female*. Secondly, we use only face images with the attribute *Female* in **AKD-GAN** and StarGAN to translate input face images to the output with the attribute *Bald*. Finally, to fairly evaluate the model performance in generating unbiased face images, the aforementioned classifier is used to distinguish if the translated face images have maintained the correct *Female* attribute during the translation. Table IV shows the classification errors of *Female* images being classified as *Male*. It demonstrates that existing state-of-the-art method StarGAN [42] caused an error of over 97%, which means genders of generated face images have largely changed and the facial attribute translation process is completely biased. The classification error can be reduced to around 51% by using the proposed **AKD-GAN**, which proves the capability of the proposed method in eliminating biases for facial attribute translation (by concentrating only on the image parts of interest). Furthermore, a visual comparison is presented in Fig. 9 where it is clear how our method does not change the gender of the input face images during the translation. In addition, a dual experiment has been conducted
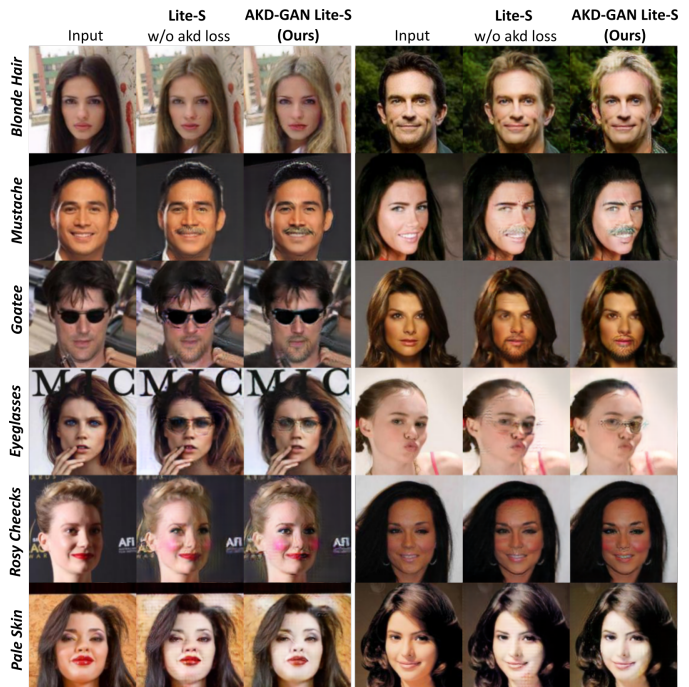
in order to generate males with attribute *Lipstick* which is uncommon in the dataset. The results (presented in last column of Table IV) further confirm the efficacy of our method.

Nevertheless, improvements of *Bald* and *Wearing Lipstick* are different and we argue that they are related to the significant different distributions of face images with these facial attributes in the dataset. Table V shows that, (i) over 99% of images with the *Bald* facial attribute are males (ii) over 99% of images with the *Wearing Lipstick* are females (iii) the *Bald* class is very underrepresented in the dataset. In addition, Table V shows how (i) a large portion of female face images have the attribute *Wearing Lipstick* internally, while, (ii) only a small portion of male images have the attribute *Bald*.

As a consequence of all these observations, for the *Bald* attribute, even if for the generator it is difficult to apply it over female images, its entanglement with the attribute *Male* is less strong due to a higher number of male images that are not bald. For this reason, localizing the facial attribute for translation using the proposed attention knowledge distillation greatly boosts the performance of generating face images of the *Bald* facial attribute effectively and empirical observations showed the bias mitigation during face image synthesis. Since the percentage of images with the *Wearing Lipstick* facial

TABLE VI: Classification results and overall FID scores over different facial attributes. The bold results represent the best results between the lightweight student network (*Lite-S*) trained without and with attention distillation loss, respectively.

Methods	Blond Hair ↑	Mustache ↑	Goatee ↑	Eyeglasses ↑	Rosy Cheeks ↑	Pale Skin ↑	Heavy Makeup ↑	Mean ↑	Overall FID ↓
Lite-S w/o akd loss (Baseline)	75.46%	24.71%	**51.38%**	56.48%	30.59%	67.14%	94.78%	57.22%	21.48
Lite-S with akd loss (Ours)	**77.81%**	**24.87%**	48.46%	**62.65%**	**35.40%**	**70.90%**	**96.25%**	**59.47%**	**21.02**

attribute is much higher in the dataset, and almost every female image has this attribute, the attribute is much more entangled with the gender and, therefore, the bias is more difficult to mitigate.

We have observed that biases exist as part of the dataset itself. We proposed two potential ways to further tackle this issue in the future. First, one possible way would be to bring external data in order to help mitigate biases in training the model. The external data could be from an extra dataset or data generated by data augmentation tools. Second, another possible way would be to design an online training strategy which enables the model to reject samples with strong biases, especially at the beginning of the training, and accept samples with biases progressively after the model has been trained stably. Our expectation is that the model can be trained without introducing biased data at the beginning, and then trained with introducing biased data to increase the diversity of the data and improve model generalizability.
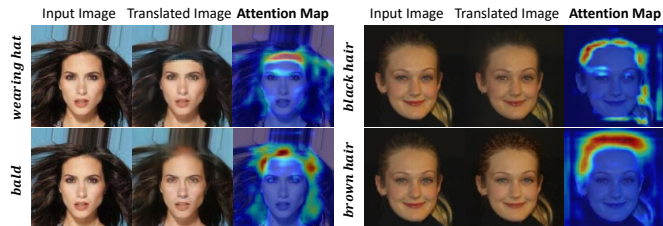

Fig. 11: "Pseudo"-attention maps generated for facial attribute translation targeting two different sets of attributes.

3) AKD-GAN with the Teacher Network and Lightweight Student Network: In this set of experiments the objective is to use visual attentions to distill attention knowledge between a teacher *T* and a smaller, lightweight student (*Lite-S*) network. Our expectation is that reduced complexity of model can be remedied by attention knowledge distillation. The models were trained to translate different facial attributes: *goatee*, *rosy cheeks*, *eyeglasses*, *mustache*, *blond hair*, etc. We present both quantitative and qualitative results in following paragraphs.

Qualitative Results. Qualitative results are presented in Fig. 10. Thanks to the contribution of the proposed attention knowledge distillation loss in **AKD-GAN**, target facial attributes are applied much more strongly to the input face images. This is particularly true for *blond hair*, *rosy checks*, *mustache* and *pale skin* whose application was unsatisfactory in the lightweight student without attention knowledge distillation loss. Moreover, target facial attributes of *eyeglasses* and *heavy makeup* are applied more convincingly to the input images. Finally, some undesired changes that can happen during the translation of certain facial attributes are less frequent. More
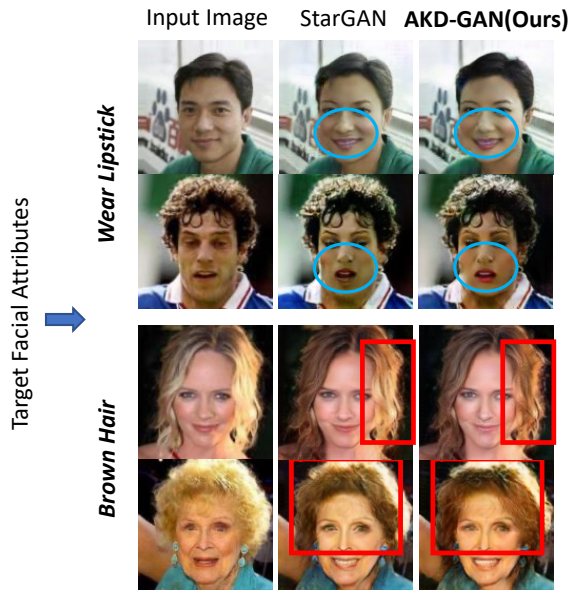

Fig. 12: Qualitative results of AKD-GAN with "pseudo"-attention maps.

specifically, when translating some facial attributes related to one particular gender, *i.e. goatee*, it can happen that gender is also translated in the output image, while, as discussed in Sec. IV-E2, this effect is greatly reduced thanks to attention knowledge distillation.

Quantitative Results. Quantitative results are shown in Table VI. First of all, the classification accuracy of synthesized face images of each facial attribute is calculated using the pretrained deep attribute classifier. Then, we also calculated the overall FID score for the synthesized face images for image quality evaluation.

Looking at the classification results, the lightweight student model (*Lite-S*) trained with attention distillation loss in **AKD-GAN** outperforms the one trained without attention distillation loss demonstrating the effectiveness of our approach. Regardless, distilling the attention knowledge with attention maps has shown its advantages in improving the performance of a lightweight network without altering the network design, demonstrating that visual attention serves more purpose than just visualization and can be used during training with success.

Regarding the overall FID score, the lightweight student model trained with attention distillation loss shows a slight boost in visual quality over the one trained without attention distillation loss but with a superior ability in translating the facial attributes.

TABLE VII: Classification accuracy and the overall FID scores over a group of different facial attributes using **AKD-GAN** with "pseudo"-attention knowledge distillation loss. The bold results represent the best results.

Methods	Teacher Attributes:	Brown Hair	Rosy Cheeks	Pointy Nose	Arched Eyebrows	Big Lips	Big Nose	Overall FID ↓
	Student Attributes:	Black Hair ↑	High Cheekbones ↑	Big Nose ↑	Bushy Eyebrows ↑	Wear Lipstick ↑	Pointy Nose ↑	
AKD-GAN (Ours)	**"pseudo"-attention knowledge distillation**	88.30%	**95.92%**	**94.66%**	**96.97%**	**97.60%**	**84.8%**	**14.34**
StarGAN [42] (Baseline)	No distillation	**94.28%**	94.92%	91.43%	95.6%	95.92%	82.44%	18.43

F. Pseudo-Attention Knowledge Distillation for Face Synthesis with Facial Attributes Translation

We present "pseudo"-attention knowledge distillation experimental results under different teacher-student designs in the proposed framework **AKD-GAN**.

TABLE VIII: Classification accuracy over different "hair color" attributes using **AKD-GAN** with attention knowledge distillation loss. The bold results represent the best results.

Methods	Teacher Attribute:	Blond Hair		
	Student Attributes:	Blond Hair ↑	Brown Hair ↑	Gray Hair ↑
AKD-GAN (Ours)	**"pseudo"-attention knowledge distillation**	**86.02%**	**88.59%**	**89.46%**
	Teacher Attribute:	Black Hair		
AKD-GAN (Ours)	**"pseudo"-attention knowledge distillation**	84.89%	89.59%	77.33%
StarGAN [42] (Baseline)	No distillation	80.94%	86.50%	76.94%

1) "Pseudo"-attention Knowledge Distillation with the Teacher and Student: We first visualize attention maps for facial attribute translation from the models (T^{pse} and T) trained for different attributes (c^{pse} and c) in Fig. 11. The top row shows attention maps generated from the model trained on a set of facial attributes c^{pse} (*wearing hat* and *black hair* for instance), while the bottom row shows attention maps generated from another model trained on a set of facial attributes c (*bald* and *brown hair*). It is evident that attention maps (last column) of each pair of facial attributes (*wearing hat* vs. *bald* and *black hair* vs. *brown hair*) share some spatial features in the input face images. This provides a strong evidence that it is possible to distill knowledge by defining and using "pseudo"-attentions from a teacher network to a student network for learning a *new* set of facial attributes.

We conducted experiments with **AKD-GAN** for facial attributes translation by using the proposed "pseudo"-attention knowledge distillation module. Firstly, we train one teacher model (T^{pse}) to translate input face images to facial attributes (c^{pse}) that are *black hair*, *arched eyebrows*, *big lips*, and *big nose*, etc. Secondly, we trained the student network (S) with **AKD-GAN** to translate input face images to different facial attributes (c) that are *brown hair*, *bushy eyebrows*, *wear lipstick* and *pointy nose*, etc., with the proposed "pseudo"-attention knowledge distillation loss. The attention maps generated on the student network are the so-called "pseudo"-attentions and "pseudo"-attention knowledge distillation loss is calculated as a weak supervisory signals for training the student network.

Quantitative Results. Quantitative results are presented in Table VII. Classification accuracy of synthesized face images of each attribute is calculated to evaluate the performance of the translation over face images with each target facial attributes and overall FID score is calculated to evaluate the quality of synthesized face images. From the results in Table VII, the proposed **AKD-GAN** trained with attention knowledge distillation loss using "pseudo"-attentions can consistently improve the translation of facial attributes and the quality of generated face images with target facial attributes.

It is observed that, when facial attributes are related to hair colors, *e.g.*, *blond hair*, *brown hair*, etc. our teacher model can be trained on just one of them and be used to distill attention knowledge for all the others. To further show that, in Table. VIII, we use the proposed pseudo-attention knowledge distillation to train two additional experiments: (a) to distill attention knowledge from the teacher trained on facial attribute *"Blond_Hair"* to different attributes, *i.e.*, *"Blond_Hair"*, *"Brown_Hair"* and *"Gray_Hair"*, respectively. (b) do the same things as in (a) but training the teacher using *"Black_Hair"*. It can be seen that "pseudo"-attention knowledge distillation module can consistently improve model performance by distilling attention knowledge from one facial attribute to a group of different attributes.
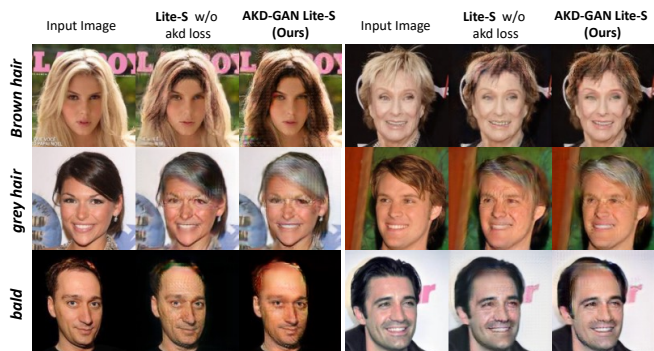

Fig. 13: Qualitative results obtained from the lightweight student (**Lite-S**) trained with and without the proposed "pseudo"-attention distillation loss.

Qualitative Results. Sample synthesized face images are presented in Fig. 12. We can see that the results of the proposed **AKD-GAN** with "pseudo"-attention knowledge distillation loss can generate better facial attributes than the state-of-art method StarGAN [42]. This is particularly evident when the

translation appears in only a small region of the input face image.

TABLE IX: Classification results and the overall FID scores over different facial attributes for generated face images from the lightweight student network (*Lite-S*) trained using "pseudo"-attention distillation loss.

Target Attributes: Methods	Brown Hair ↑	Gray Hair ↑	Bald ↑	Straight Hair ↑	Avg. Accuracy ↑	Overall FID ↓
Lite-S w/o akd loss (Baseline)	79.16%	32.46%	8.03%	56.48%	44.03%	27.02
Lite-S with akd loss (Ours)	**82.77%**	**32.66%**	**12.43%**	**86.13%**	**53.5%**	**24.72**

2) "Pseudo"-attention Knowledge Distillation with the Teacher and Lightweight Student: Following the similar experimental settings as in Sec. (IV-E3), the objective is to use "pseudo"-attentions defined in Sec. III-C to distill knowledge between a teacher T network targeting on a set of facial attributes and a lightweight student (*Lite-S*) network targeting on a different set of facial attributes. Firstly, we trained one teacher model (T^{pse}) to translate input images to facial attributes (c^{pse}) that are *black hair*, *blond hair*, *wearing hat* and *wavy hair*. Secondly, we trained the lightweight student network (*Lite-S*) with **AKD-GAN** to translate input images to different facial attributes (c) that are *brown hair*, *gray hair*, *bald* and *straight hair* with the attention knowledge distillation loss. Finally, "pseudo"-attentions are used to distill knowledge between the teacher network and the lightweight student network (*Lite-S*).

Qualitative Results. We present translated face images in Fig. 13. We can see that results of the lightweight student network (*Lite-S*) trained with the proposed attention distillation loss are more convincing than the outputs from the same model that does not employ attention distillation loss. This is particularly evident for facial attributes of *brown hair*, *grey hair* and *bald*.

Quantitative Results. Furthermore, the classification accuracy of translated face images of each attribute and the overall FID are calculated in Table IX. From results in Table IX, the lightweight student network (*Lite-S*) trained with attention distillation loss using "pseudo"-attentions outperforms the network trained without it, which proves the effectiveness of our approach.

G. Attention Knowledge Distillation for Face Synthesis with Human Facial Expressions Translation

We conducted experiments by using the proposed method **AKD-GAN** for human facial expressions translation on the Radboud Faces Database (**RaFD**) [50] which consists of different human facial expressions. We follow the proposed experimental settings as in Sec. IV-E and use two different designs for *teacher-student* and *teacher-lightweight-student*. We use classification accuracy to evaluate the performance of the proposed method for face synthesis with facial expressions translation and FID scores is not calculated as an evaluation metric since the test images in the dataset are too few.

The first experiment is conducted on 8 facial expressions translation. Classification accuracy of synthesized face images with different human facial expressions is presented in Table X. Improved classification accuracy can be observed in

TABLE X: Comparisons of classification accuracy over different human expressions using the proposed **AKD-GAN**.

Methods	Happy ↑	Angry ↑	Sad ↑	Contemptuous ↑	Disgusted ↑
StarGAN [42] (Baseline)	77.25%	71.37%	66.31%	84.00%	80.00%
AKD-GAN (Ours)	**79.31%**	**75.63%**	**73.81%**	**84.88%**	**86.75%**
	Neutral ↑	Fearful ↑	Surprised ↑	Mean ↑	
StarGAN [42] (Baseline)	67.93%	85.18%	81.56%	76.70%	
AKD-GAN (Ours)	**71.44%**	**86.88%**	**92.82%**	**81.44%**	

most facial expressions by using the proposed **AKD-GAN**, especially for *sad*, *disgusted*, and *surprised*, which proves the effectiveness of the proposed method on face synthesis with human facial expression translation.

Some qualitative results of human facial expression translation on *angry*, *disgusted*, *neutral*, and *surprised*, are presented in Fig. 14. We can see that better human facial expression translation images are obtained by using the proposed **AKD-GAN**, which validates the effectiveness of our method.
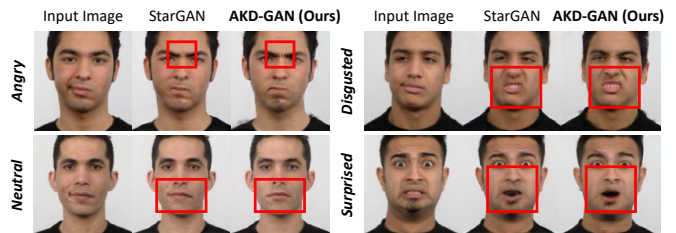

Fig. 14: Qualitative results of human facial expression translation using **AKD-GAN**. *Note: better views are obtained with zooming in.*

TABLE XI: Comparisons of classification results over 4 different human expressions using lightweight student network (*Lite-S*) in **AKD-GAN**.

Methods	Disgusted ↑	Fearful ↑	Happy ↑	Sad ↑	Mean ↑
Lite-S w/o akd loss	100.00%	95.31%	96.85%	70.57%	90.68%
Lite-S with akd loss (Ours)	**100.00%**	**97.13%**	**97.91%**	**72.13%**	**91.79%**

Furthermore, we conducted experiments using lightweight student network (*Lite-S*). Classification results on 4 different facial expressions, *disgusted*, *fearful*, *happy* and *sad*, on the synthesized face images are presented in Table XI.

H. Ablation Studies

In this sub-section, we explain the design choice of using last residual block for generating attention maps from two perspectives: empirical observations and additional ablation experiments. As introduced in Section I, exploiting visual interpretations in image-to-image conditional adversarial networks is a fundamental step in order to improve upon them.

First, we investigate the architecture of a typical image-conditioned cGANs, *e.g.*, StarGAN. It consists of an encoder followed by residual blocks, then a decoder to decode features to images with high resolution. It is known that convolutional layers naturally retain spatial information, so we can expect

TABLE XII: Ablation results by using our AKD-GAN with attention maps generated from the ***second last*** and ***last*** residual blocks for attention knowledge distillation during training.

Methods	Attention Knowledge Distillation	Interpretations	Blond Hair ↑	Goatee ↑	Eyeglasses ↑	Heavy Makeup ↑	Pale Skin ↑
AKD-GAN (Ours)	✓(second last block)	✓	81.22%	68.53%	98.35%	86.00%	83.11%
AKD-GAN (Ours)	✓(last block)	✓	**86.02%**	**74.45%**	**99.48%**	**91.47%**	**88.32%**

the last convolutional layers to have the best representation between high-level semantics and detailed spatial information. Therefore, in StarGAN, the encoder and residual blocks process the input images and target attributes step by step, then the neurons in convolutional layers of residual blocks encode semantic and spatial information of the input images and target attributes, outputting feature representations for the decoder. We streamline the attention generation by using the gradient information flowing into the last residual blocks of the generator for a particular facial attributes translation of interest.

Second, this design choice is supported by preliminary experimental observations shown in Fig. 2, where we present visual attention maps extracted from different residual blocks. The produced attention maps highlight the spatial information of features, where the generator focuses on a certain facial attribute and allow us to identify which layers in the generator are more devoted to the facial attributes translation task. The most refined attention can be seen in the last residual block (highlighted in red). This is a crucial observation for the development of our system, since it motivates us to derive the proposed attention knowledge distillation with the attention maps extracted from the last residual block. We argue that the attention maps extracted from previous layers are more noisy and deriving attention knowledge distillation using these attention maps would mislead the model training. The designed experiments given in previous sections and described below validate our proposed system.

We present ablation experimental results on CelebA dataset in Table XII by using the proposed **AKD-GAN** with the attention knowledge distillation but attention maps are generated from the ***second last*** residual blocks and compare them with the **AKD-GAN** trained with attention knowledge distillation and attention maps generated from the ***last*** residual blocks. From Table XII, it can be observed that using attention maps generated from the ***second last*** residual blocks for attention knowledge distillation gives worse performance than using attention maps generated from the last residual blocks. Given the attention maps shown in Fig. 2, we argue that the main reason for this is that the attention maps generated from the ***second last*** residual blocks contain inaccurate spatial information which would mislead the model training if these attention maps are used for attention knowledge distillation.

V. CONCLUSIONS

We streamlined the generation of visual interpretations by means of gradient-based attention mechanisms for facial attribute translation on the conditional generative adversarial networks and showed how they can be used during training to improve the performance of generative models. We developed a novel system called **AKD-GAN** where a teacher network distills its knowledge via visual attentions to a student network by proposing an attention knowledge distillation loss in order to train the student network to generate better face images.

We experimented with the proposed novel method with various system design implementations, including the teacher-student paradigms with different model complexities in student networks and attention knowledge distillation using full self-supervision and weak supervision for model training. We showed the effectiveness of the system in removing biases in the attribute generation. Finally, we evaluated the proposed framework on the widely used public face image datasets CelebA and human expression dataset RaFD, demonstrating with both quantitative and qualitative results the effectiveness of the approach for diverse applications.

The proposed face synthesis with facial attributes translation can be applied in a wide range of computer vision and biometrics tasks, such as attribute-based recognition and identification, etc. Also, it can be applied as a powerful data augmentation tool by using synthesized face images with various facial attributes for multi-purposes usage, including large-scale training for deep models, adversarial attacks. etc. Besides, face images synthesized with various facial attributes can be potentially applied for disguised and concealed identity recognition. In addition, generating better face images with target attributes will have the potential impact for the entertainment industry and also for digital art. Finally, synthesizing facial attributes in videos for video generation in real time will be an interesting future work.

REFERENCES

[1] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.
[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.
[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.
[4] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *WACV*, 2018.
[5] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. Fu, "Tell me where to look: Guided attention inference network," in *CVPR*, 2018.

[6] P. Dhar, R. V. Singh, K.-C. Peng, Z. Wu, and R. Chellappa, "Learning without memorizing," in *CVPR*, 2019.

[7] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps, "Towards visually explaining variational autoencoders," in *CVPR*, 2020.

[8] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *CVPR*, 2015.

[9] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *CVPR*, 2019.

[10] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.

[11] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *ICML*, 2019.

[12] K. Zhang, Y. Su, X. Guo, L. Qi, and Z. Zhao, "Mu-gan: Facial attribute editing based on multi-attention mechanism," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 9, pp. 1614–1626, 2020.

[13] Y. Lin, Y. Wang, Y. Li, Y. Gao, Z. Wang, and L. Khan, "Attention-based spatial guidance for image-to-image translation," in *WACV*, 2021.

[14] D. Kim, M. A. Khan, and J. Choo, "Not just compete, but collaborate: Local image-to-image translation via cooperative mask prediction," in *CVPR*, 2021.

[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[16] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *NeurIPS*, 2017.

[17] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *CVPR*, 2019.

[18] A. Polino, R. Pascanu, and D. Alistarh, "Model compression via distillation and quantization," *arXiv preprint arXiv:1802.05668*, 2018.

[19] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[20] A. Aguinaldo, P.-Y. Chiang, A. Gain, A. Patil, K. Pearson, and S. Feizi, "Compressing GANs using knowledge distillation," *arXiv preprint arXiv:1902.00159*, 2019.

[21] M. Li, J. Lin, Y. Ding, Z. Liu, J.-Y. Zhu, and S. Han, "Gan compression: Efficient architectures for interactive conditional GANs," in *CVPR*, 2020.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.

[23] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.

[24] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of styleGAN," in *CVPR*, 2020.

[25] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[26] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[27] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *ICCV*, 2017.

[28] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *T-PAMI*, vol. 41, no. 8, pp. 1947–1962, 2018.

[29] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[30] X. Di and V. M. Patel, "Multimodal face synthesis from visual attributes," *T-BIOM*, vol. 3, no. 3, pp. 427–439, 2021.

[31] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *CVPR*, 2017.

[32] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays, "TextureGAN: Controlling deep image synthesis with texture patches," in *CVPR*, 2018.

[33] L. Yang, K. Pang, H. Zhang, and Y.-Z. Song, "Sketchaa: Abstract representation for abstract sketches," in *ICCV*, 2021.

[34] S.-Y. Wang, D. Bau, and J.-Y. Zhu, "Sketch your own gan," in *ICCV*, 2021.

[35] X. Di and V. M. Patel, "Facial synthesis from visual attributes via sketch using multiscale generators," *T-BIOM*, vol. 2, no. 1, pp. 55–67, 2020.

[36] X. Di, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel, "Multi-scale thermal to visible face verification via attribute guided synthesis," *T-BIOM*, vol. 3, no. 2, pp. 266–280, 2021.

[37] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *NeurIPS*, 2017.

[38] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[39] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018.

[40] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *ICCV*, 2019.

[41] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017.

[42] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *CVPR*, 2018.

[43] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *T-IP*, vol. 28, no. 11, pp. 5464–5478, 2019.

[44] M. M. R. Siddiquee, Z. Zhou, N. Tajbakhsh, R. Feng, M. B. Gotway, Y. Bengio, and J. Liang, "Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization," in *ICCV*, 2019.

[45] P.-W. Wu, Y.-J. Lin, C.-H. Chang, E. Y. Chang, and S.-W. Liao, "RelGAN: Multi-domain image-to-image translation via relative attributes," in *ICCV*, 2019.

[46] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, "Stgan: A unified selective transfer network for arbitrary image attribute editing," in *CVPR*, 2019.

[47] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Starganv2: Diverse image synthesis for multiple domains," in *CVPR*, 2020.

[48] T. Fontanini, E. Iotti, L. Donati, and A. Prati, "MetalGAN: Multi-domain label-less image synthesis using cGANs and meta-learning," *Neural Networks*, vol. 131, pp. 185–200, 2020.

[49] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015.

[50] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[51] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," 2017.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[53] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *NeurIPS*, 2017.

[54] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *ICLR*, 2018.

Runze Li received the B.E degree in information engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2014 and M.S. in information technology (Distributed Computing) from the University of Melbourne, VIC, Australia. He is currently pursuing the Ph.D. degree at computer science and engineering in the University of California, Riverside, CA, USA.

Tomaso Fontanini graduated in Computer Engineering at the University of Parma in 2017 and got his PhD (funded by Adidas AG) in Information Engineering in 2020 from the same University. He is currently doing his first year as a postdoc researcher in the IMPLab laboratory at the University of Parma. His research interest are mainly Deep Learning and Generative Models.

Andrea Prati graduated in Computer Engineering from the University of Modena and Reggio Emilia in 1998. He got his PhD in 2002 from the same university. He served as Assistant Professor from 2005 to 2011, and Associate Professor at the University IUAV of Venice since 2015. In December 2015 he moved to the University of Parma and got promoted to full professorship in 2019. He is the head of IMPLab research group and his research interests are related to computer vision and image processing, deep learning and generative models. He is the author of 9 book chapters, 40+ papers in international referred journals and 100+ papers in proceedings of international conferences and workshops. To date, his h-index on Google Scholar is 41, with a total of 8773 citations. On Scopus, his h-index is 27, with a total of 4580 citations. Andrea Prati is Senior Member of IEEE, Fellow of IAPR, and a member of CVPL.

Bir Bhanu received B.S. (with Hons.) from IIT-BHU; M.E (with Distinction) from BITS (Pilani); S.M. and E.E. in electrical engineering and computer science from Massachusetts Institute of Technology, Cambridge, MA; Ph.D. in electrical engineering from the University of Southern California, Los Angeles, CA and M.B.A. from the University of California, Irvine, CA. He is the Bourns endowed University of California Presidential Chair in Engineering, the Distinguished Professor of electrical and computer engineering and the Founding Director of the interdisciplinary Center for Research in Intelligent Systems (1998-2019) and the Visualization and Intelligent Systems Laboratory (1991-) at the University of California, Riverside (UCR), CA. He is the Founding Professor of electrical engineering with UCR and served as its first Chair (1991-94). He has been the cooperative Professor of computer science and engineering (since 1991), bioengineering (since 2006) and mechanical engineering (since 2008). Recently he served as the Interim Chair of the Department of Bioengineering from 2014-16. He also served as the Director of the National Science Foundation graduate research and training program in video bioinformatics with UCR. Prior to joining UCR in 1991, he was a Senior Honeywell Fellow with Honeywell Inc. He has published extensively and has received university, industry awards for research excellence, outstanding contributions and team efforts and journal/conference best paper awards. He received the Faculty Research Lecturer Award from UCR in 2019 commencement. His research interests include computer vision, pattern recognition and data mining, machine learning, artificial intelligence, image processing, image and video database, graphics and visualization, robotics, human-computer interactions, and biological, medical, military and intelligence applications. Dr. Bhanu is a Fellow of IEEE, AAAS, IAPR, SPIE, NAI and AIMBE.