AttentionGAN: Unpaired Image-to-Image Translation Using Attention-Guided Generative Adversarial Networks

Hao Tang [iD], Hong Liu [iD], Dan Xu [iD], Philip H. S. Torr, and Nicu Sebe [iD], *Senior Member, IEEE*

Abstract—State-of-the-art methods in the image-to-image translation are capable of learning a mapping from a source domain to a target domain with unpaired image data. Though the existing methods have achieved promising results, they still produce visual artifacts, being able to translate low-level information but not high-level semantics of input images. One possible reason is that generators do not have the ability to perceive the most discriminative parts between the source and target domains, thus making the generated images low quality. In this article, we propose a new Attention-Guided Generative Adversarial Networks (AttentionGAN) for the unpaired image-to-image translation task. AttentionGAN can identify the most discriminative foreground objects and minimize the change of the background. The attention-guided generators in AttentionGAN are able to produce attention masks, and then fuse the generation output with the attention masks to obtain high-quality target images. Accordingly, we also design a novel attention-guided discriminator which only considers attended regions. Extensive experiments are conducted on several generative tasks with eight public datasets, demonstrating that the proposed method is effective to generate sharper and more realistic images compared with existing competitive models. The code is available at https://github.com/Ha0Tang/AttentionGAN.

Index Terms—Attention guided, generative adversarial networks (GANs), unpaired image-to-image translation.

I. INTRODUCTION

RECENTLY, Generative Adversarial Networks (GANs) [1] in various fields such as computer vision and image processing have produced powerful translation systems with supervised settings such as Pix2pix [2], where

Fig. 1. Comparison with existing unpaired image-to-image translation methods (e.g., CycleGAN [3] and GANimorph [6]) with an example of horse-to-zebra translation. We are interested in transforming horses into zebras. In this case we should be agnostic to the background. However, methods such as CycleGAN and GANimorph will transform the background in a nonsensical way, in contrast to our attention-based method (the generated attention masks are shown in the last column).

paired training images are required. However, paired data are usually difficult or expensive to be obtained. The input-output pairs for tasks such as artistic stylization could be even more difficult to be acquired since the desired output is quite complex, typically requiring artistic authoring. To tackle this problem, CycleGAN [3], DualGAN [4] and DiscoGAN [5] provide a new insight, in which the GAN models can learn the mapping from a source domain to a target one with unpaired image data.

Despite these efforts, unpaired image-to-image translation remains a challenging problem. Most existing models change unwanted parts in the translation and can also be easily affected by background changes (see Fig. 1). In order to address these limitations, Liang *et al.* [7] propose Contrast-GAN, which uses object-mask annotations provided by the dataset to guide the generation, first cropping the unwanted parts in the image based on the masks, and then pasting them back after the translation. While the generated results are reasonable, it is hard to collect training data with object-mask annotations. Another option is to train an extra model to detect the object masks and then employ them for the mask-guided generation [8], [9]. In this case, we need to significantly increase the network capacity, which consequently raises the training complexity in both time and space.

To overcome the aforementioned issues, we propose a novel Attention-Guided Generative Adversarial Networks (AttentionGAN) for the unpaired image-to-image translation task. Fig. 1 shows a comparison with exiting image-to-image trans-
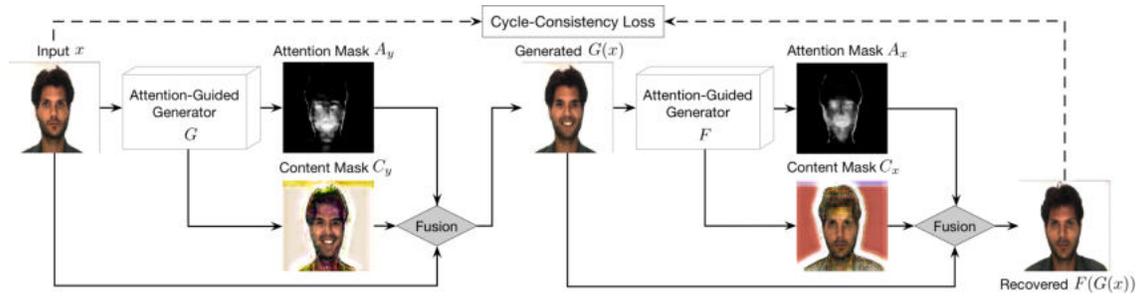
Fig. 2.   Framework of the proposed attention-guided generation Scheme I, which consists of two AGs G and F. We show one mapping in this figure, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. We also have the other mapping, i.e., $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. The AGs have a built-in attention module, which can perceive the most discriminative content between the source and target domains. We then fuse the input image, the content mask, and the attention mask to synthesize the final result.

lation methods using a horse-to-zebra translation example. The most important advantage of AttentionGAN is that the proposed generators can focus on the foreground of the target domain and preserve the background of the source domain effectively. Specifically, the proposed generator learns both foreground and background attentions. It uses the foreground attention to select from the generated output for the foreground regions, while uses the background attention to maintain the background information from the input image. In this way, the proposed AttentionGAN can focus on the most discriminative foreground and ignore the unwanted background. We observe that AttentionGAN achieves significantly better results than both GANimorph [6] and CycleGAN [3]. As shown in Fig. 1, AttentionGAN not only produces clearer results but also successfully maintains the little boy in the background and only performs the translation for the horse behind it. However, the existing holistic image-to-image translation approaches are generally interfered by irrelevant background content, thus hallucinating texture patterns of the target objects.

We propose two different attention-guided generation Schemes for the proposed AttentionGAN. The framework of the proposed Scheme I is shown in Fig. 2. The proposed generators are equipped with a built-in attention module, which can disentangle the discriminative semantic objects from the unwanted parts via producing an attention mask and a content mask. Then we fuse the attention and the content masks to obtain the final generation. Moreover, we design two novel attention-guided discriminators (ADs) which aim to consider only the attended foreground regions. The proposed attention-guided generators (AGs) and discriminators are trained in an end-to-end fashion. The proposed attention-guided generation Scheme I can achieve promising results on the facial expression translation, where the change between the source domain and the target domain is relatively minor. However, it performs unsatisfactorily on more challenging scenarios in which a more complex semantic translation is required such as horse-to-zebra translation as shown in Fig. 1. To tackle this issue, we further propose a more advanced attention-guided generation Scheme II, as depicted in Fig. 3. The improvement upon the Scheme I is mainly threefold: First, in Scheme I, the attention and the content masks are generated with the same network. To have a more powerful generation of them, we employ two separate subnetworks in Scheme II. Second, in Scheme I, we only generate the foreground attention mask to

focus on the most discriminative semantic content. However, to better learn the foreground and preserve the background simultaneously, we produce both foreground and background attention masks in Scheme II. Third, as the foreground generation is more complex, instead of learning a single content mask in Scheme I, we learn a set of several intermediate content masks, and correspondingly we also learn the same number of foreground attention masks. The generation of multiple intermediate content masks is beneficial for the network to learn a more rich generation space. The intermediate content masks are then fused with the foreground attention masks to produce the final content masks. Extensive experiments on several challenging public benchmarks demonstrate that the proposed Scheme II can produce higher-quality target images compared with existing state-of-the-art methods.

The contribution of this article is summarized as follows:
1) We propose a new AttentionGAN for the unpaired image-to-image translation task. This framework stabilizes the GANs' training and thus improves the quality of generated images through jointly approximating attention and content masks with several losses and optimization methods.
2) We design two novel attention-guided generation schemes for the proposed framework, to better perceive and generate the most discriminative foreground parts and simultaneously preserve well the unfocused objects and background. Moreover, the proposed AG and discriminator can be flexibly applied in other GANs to improve the multi domain image-to-image translation task, which we believe would also be beneficial to other related research.
3) We conduct extensive experiments on eight popular datasets and experimental results show that the proposed AttentionGAN can generate photo-realistic images with more clear details compared with existing methods. We also established new state-of-the-art results on these datasets.

Part of the material presented here appeared in [10]. The current article extends [10] in several ways. 1) We propose a more advanced attention-guided generation scheme, i.e., Scheme II, which is a more robust and general framework for both unpaired image-to-image translation and multi domain image translation tasks. 2) We present an in-depth description of the proposed approach, providing all the architectural and implementation details of the method,
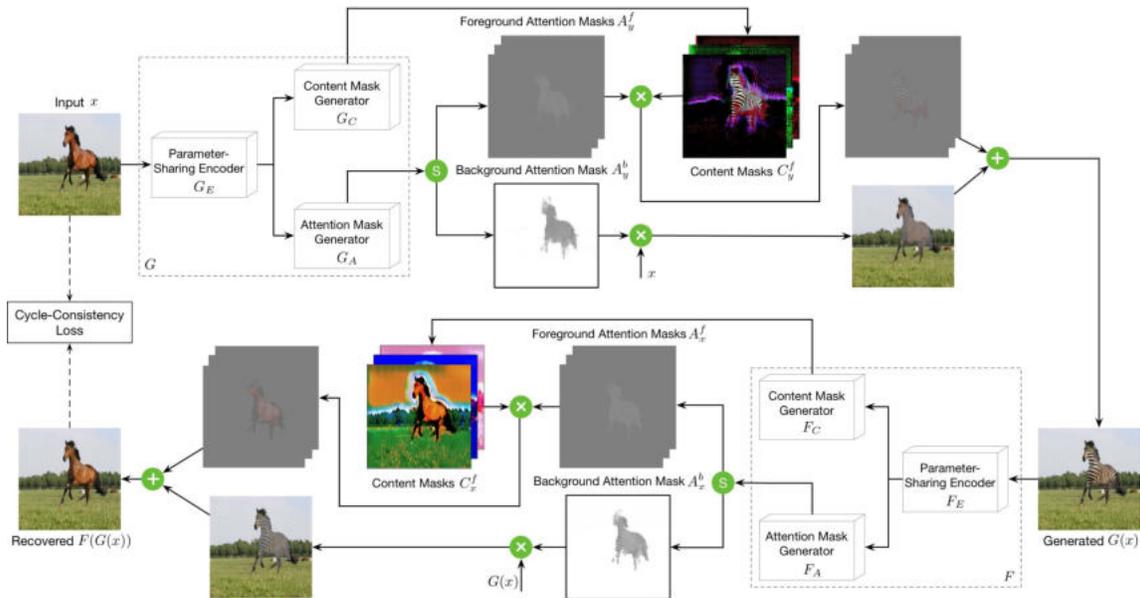
Fig. 3. Framework of the proposed attention-guided generation Scheme II, which consists of two AGs G and F. We show one mapping in this figure, i.e., $x{\rightarrow}G(x){\rightarrow}F(G(x)){\approx}x$. We also have the other mapping, i.e., $y{\rightarrow}F(y){\rightarrow}G(F(y)){\approx}y$. Each generator such as G consists of a parameter-sharing encoder G_E, an attention mask generator G_A, and a content mask generator G_C. G_A aims to produce attention masks of both foreground and background to attentively select the useful content from the corresponding content masks generated by G_C and the input image x. The proposed model is constrained by the cycle-consistency loss and trained in an end-to-end fashion so that each generator can benefit from each other. The symbols $\oplus$, $\otimes$ and $\circledS$ denote element-wise addition, multiplication, and channel-wise Softmax, respectively.

with special emphasis on guaranteeing the reproducibility of the experiments. The source code is also available online. 3) We extend the experimental evaluation provided in [10] in several directions. First, we conduct extensive experiments on eight popular datasets, demonstrating the wide application scope of the proposed AttentionGAN. Second, we conduct exhaustive ablation studies to evaluate each component of the proposed method. Third, we investigate the influence of hyperparameters on generation performance. Fourth, we provide the visualization of the learned attention masks to better interpret our model. Fifth, we have also included more state-of-the-art baselines, e.g., UAIT [11], U-GAT-IT [12], SAT [13], MUNIT [14], DRIT [15], GANimorph [6], TransGaGa [16], DA-GAN [17], and we observe that AttentionGAN achieves better results than those. Lastly, we conduct extensive experiments on the multi domain image translation task, and experimental results show that AttentionGAN achieves much better results than existing leading methods such as StarGAN [18].

II. RELATED WORK

A. Generative Adversarial Networks (GANs)

[1] are powerful generative models, which have achieved impressive results on different computer vision tasks, e.g., image/video generation [19], [20]. To generate meaningful images that meet user requirements, Conditional GANs (CGANs) [21], [22] inject extra information to guide the image generation process, which can be discrete labels [18], [23], object keypoints [24], human skeleton [25]–[27], and semantic maps [28]–[30].

B. Image-to-Image Translation

models learn a translation function using convolutional neural networks (CNNs). Pix2pix [2] is a conditional

framework using a CGAN to learn a mapping function from input to output images. Wang *et al.* [31] propose Pix2pixHD for high-resolution photo-realistic image-to-image translation, which can be used for turning semantic label maps into photo-realistic images. Similar ideas have also been applied to many other tasks, such as hand gesture generation [25]. However, most of the tasks in the real world suffer from having few or none of the paired input-output samples available. When paired training data is not accessible, image-to-image translation becomes an ill-posed problem.

C. Unpaired Image-to-Image Translation

To overcome this limitation, the unpaired image-to-image translation task has been proposed. In this task, the approaches learn the mapping function without the requirement of paired training data. Specifically, CycleGAN [3] learns the mapping between two image domains instead of paired images. Apart from CycleGAN, many other GAN variants [4], [5], [18], [32]–[35] have been proposed to tackle the cross-domain problem. However, those models can be easily affected by unwanted content and cannot focus on the most discriminative semantic part of images during the translation stage.

D. Attention-Guided Image-to-Image Translation

To fix the aforementioned limitations, several works employ an attention mechanism to help image translation. Attention mechanisms have been successfully introduced in many applications in computer vision such as depth estimation [36], helping the models to focus on the relevant portion of the input.

Recent works use attention modules to attend to the region of interest for the image translation task in an unsupervised way, which can be divided into two categories. The first

category is to use extra data to provide attention. For instance, Liang *et al.* [7] propose ContrastGAN, which uses the object mask annotations from each dataset as extra input data. Moreover, Mo *et al.* [37] propose InstaGAN that incorporates instance information (e.g., object segmentation masks) and improves multi-instance transfiguration.

The second type is to train another segmentation or attention model to generate attention maps and fit them to the system. For example, Chen *et al.* [8] use an extra attention network to generate attention maps so that more attention can be paid to objects of interest. Kastaniotis *et al.* [9] present ATAGAN, which uses a teacher network to produce attention maps. Yang *et al.* [13] proposed to add an attention module to predict an attention map to guide the image translation process. Kim *et al.* [12] propose to use an auxiliary classifier to generate attention masks. Mejjati *et al.* [11] propose attention mechanisms that are jointly trained with the generators, discriminators, and other two attention networks.

All these methods employ extra networks or data to obtain attention masks, which increases the number of parameters, training time, and storage space of the whole system. Moreover, we still observe unsatisfactory aspects in the generated images by these methods. To fix both limitations, we propose a novel AttentionGAN, which aims to disentangle the input image into foreground and background by generating multiple foreground attention masks, a background attention mask, and multiple content masks. Then AttentionGAN learns to attend to key parts of the image and translate the salient objects/foreground to the target domain while keeping everything else unaltered, essentially avoiding undesired artifacts or changes. In this way, we do not need any extra models to obtain the attention masks of objects of interest. To the best of our knowledge, this idea has not been investigated by existing attention-guided image-to-image translation methods. Also, extensive experiments show that AttentionGAN achieves significantly better results than the existing attention-guided generation methods. Most importantly, our method can be applied to any GAN-based framework such as unpaired, paired, and multi domain image-to-image translation frameworks.

III. ATTENTION-GUIDED GANs

A. Attention-Guided Generation

GANs [1] are composed of two competing modules: the generator G and the discriminator D, which are iteratively trained competing against each other in the manner of two-player mini-max. More formally, let X and Y denote two different image domains, $x_i \in X$ and $y_j \in Y$ denote the training images in X and Y, respectively (for simplicity, we usually omit the subscript i and j). For most current image translation models, e.g., CycleGAN [3] and DualGAN [4], they include two generators G and F, and two corresponding adversarial discriminators D_X and D_Y. Generator G maps x from the source domain to the generated image $G(x)$ in the target domain Y and tries to fool the discriminator D_Y, whilst D_Y focuses on improving itself to be able to tell whether a sample is a generated sample or a real data sample. Similar to generator F and discriminator D_X.

1) Attention-Guided Generation Scheme I: For the proposed AttentionGAN, we learn two mappings between domains X and Y via two generators with built-in attention mechanism, i.e., $G : x \rightarrow [A_y, C_y] \rightarrow G(x)$ and $F : y \rightarrow [A_x, C_x] \rightarrow F(y)$, where A_x and A_y are the attention masks of images x and y, respectively; C_x and C_y are the content masks of images x and y, respectively; $G(x)$ and $F(y)$ are the generated images. The attention masks A_x and A_y define a per pixel intensity specifying to which extent each pixel for the content masks C_x and C_y will contribute to the final rendered image. In this way, the generator does not need to render static elements (it refers to background) and can focus exclusively on the pixels defining the domain content movements, leading to sharper and more realistic synthetic images. After that, we fuse the input image x, the generated attention mask A_y, and the content mask C_y to obtain the targeted image $G(x)$. In this way, we can disentangle the most discriminative semantic objects and the unwanted part of images. Take Fig. 2 as an example, the AGs focus only on those regions of the image that are responsible for generating the novel expression such as eyes and mouth, and keeping the rest of the parts of the image such as hair, glasses, and clothes untouched. The higher intensity in the attention mask means the larger contribution for changing the expression.

The input of each generator is a three-channel image, and the outputs of each generator are an attention mask and a content mask. Specifically, the input image of generator G is $x \in \mathbb{R}^{H \times W \times 3}$, and the outputs are the attention mask $A_y \in \{0, \ldots, 1\}^{H \times W}$ and content mask $C_y \in \mathbb{R}^{H \times W \times 3}$. Thus, we use the following formula to produce the final image $G(x)$:

$$G(x) = C_y * A_y + x * (1 - A_y) \tag{1}$$

where the attention mask A_y is copied to three channels for multiplication purpose. Intuitively, the attention mask A_y enables some specific areas where domain changed to get more focus and applying it to the content mask C_y can generate images with clear dynamic area and unclear static area. The static area should be similar between the generated image and the original real image. Thus, we can enhance the static area in the original real image $x * (1 - A_y)$ and merge it to $C_y * A_y$ to obtain final result $C_y * A_y + x * (1 - A_y)$. Similarly, the formulation for generator F and input image y can be expressed as $F(y) = C_x * A_x + y * (1 - A_x)$.

The proposed attention-guided generation Scheme I performs well on the tasks where the source domain and the target domain have large overlap similarities, such as facial expression-to-expression translation. However, we observe that it cannot generate photo-realistic images on complex tasks such as horse-to-zebra translation, as shown in Fig. 6. The drawbacks of Scheme I are threefold: 1) The attention and the content mask are generated by the same network, which could degrade the quality of the generated images. 2) We observe that the Scheme I only produces one attention mask to simultaneously change the foreground and preserve the background of the input images. 3) We see that Scheme I only produces one content mask to select useful content for generating the foreground content, which means the model does not have enough ability to deal with complex tasks

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TANG *et al.*: ATTENTIONGAN: UNPAIRED IMAGE-TO-IMAGE TRANSLATION

5

such as horse-to-zebra translation. To solve these limitations, we further propose a more advanced attention-guided generation Scheme II as shown in Fig. 3.

2) Attention-Guided Generation Scheme II: Scheme I adopts the same network to produce both attention and content masks, and we argue that this will degrade the generation performance. In Scheme II, the proposed generators G and F are composed of two subnets each for generating attention masks and content masks as shown in Fig. 3. For instance, G is comprised of a parameter-sharing encoder G_E, an attention mask generator G_A, and a content mask generator G_C. G_E aims at extracting both low-level and high-level deep feature representations. G_C targets to produce multiple intermediate content masks. G_A tries to generate multiple attention masks. By the way, both attention mask generation and content mask generation have their own network parameters and will not interfere with each other.

To fix the limitation (2) of Scheme I, in Scheme II, the attention mask generator G_A targets to generate both $n-1$ foreground attention masks $\{A_y^f\}_{f=1}^{n-1}$ and one background attention mask A_y^b, as shown in Fig. 3. By doing so, the proposed network can simultaneously learn the novel foreground and preserve the background of input images. The key point of success of the proposed Scheme II is the generation of both foreground and background attention masks, which allow the model to modify the foreground and simultaneously preserve the background of input images. This is exactly the goal that the unpaired image-to-image translation tasks aim to optimize.

Moreover, we observe that in some generation tasks such as horse-to-zebra translation, the foreground generation is very difficult if we only produce one content mask as done in Scheme I. To fix this limitation, we use the content mask generator G_C to produce $n-1$ content masks $\{C_y^f\}_{f=1}^{n-1}$, as shown in Fig. 3. Then take the input image x into consideration, we obtain n intermediate content masks. In this way, a three-channel generation space can be enlarged to a $3n$-channel generation space, which is more suitable for learning a good mapping for complex image translation.

Specifically, the feature map m extracted from G_E is first fed into the generator G_C to generate $n-1$ content masks $\{C_y^f\}_{f=1}^{n-1}$, followed by a Tanh($\cdot$) activation function. This process can be expressed as

$$C_y^f = \text{Tanh}\left(mW_C^f + b_C^f\right), \quad \text{for } f = 1, \ldots, n-1 \quad (2)$$

where a convolution operation is performed with $n-1$ convolutional filters $\{W_C^f, b_C^f\}_{f=1}^{n-1}$. Thus, the $n-1$ content masks and the input image x can be regarded as the candidate image pool.

Meanwhile, the feature map m is fed into a group of filters $\{W_A^f, b_A^f\}_{f=1}^{n}$ to generate the corresponding n attention masks

$$A_y^f = \text{Softmax}\left(mW_A^f + b_A^f\right), \quad \text{for } f = 1, \ldots, n \quad (3)$$

where Softmax($\cdot$) is a channel-wise Softmax function used for the normalization. We then split $\{A_y^f\}_{f=1}^{n}$ into $n-1$ foreground attention masks $\{A_y^f\}_{f=1}^{n-1}$ and one background attention mask A_y^b along the channel dimension. Note that the generated one background attention mask and $n-1$ foreground

attention masks are complementary, but the generated $n-1$ foreground attention masks are not complementary to each other. By doing so, the proposed model has a more flexible ability to learn and translate the foreground content.

Finally, the attention masks are multiplied by the corresponding content masks to obtain the final target result as shown in Fig. 3. Formally, this is written as

$$G(x) = \sum_{f=1}^{n-1}(C_y^f * A_y^f) + x * A_y^b. \quad (4)$$

In this way, we can preserve the background of the input image x, i.e., $x*A_y^b$, and simultaneously generate the novel foreground content for the input image, i.e., $\sum_{f=1}^{n-1}(C_y^f*A_y^f)$. Then, we merge the generate foreground $\sum_{f=1}^{n-1}(C_y^f*A_y^f)$ to the background $x*A_y^b$ for obtaining the final result $G(x)$. The formulation of generator F and input image y can be expressed as $F(y) = \sum_{f=1}^{n-1}(C_x^f * A_x^f) + y * A_x^b$, where n attention masks $[\{A_x^f\}_{f=1}^{n-1}, A_x^b]$ are also produced by a channel-wise Softmax activation function for the normalization.

B. Attention-Guided Cycle

To further regularize the mappings, CycleGAN [3] adopts two cycles in the generation process. The motivation of cycle consistency is that if we translate from one domain to the other and back again, we should arrive at where we started. Specifically, for each image x in domain X, the image translation cycle should be able to bring x back to the original one, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. Similarly, for the image y, we have another cycle, i.e., $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. These behaviors can be achieved by using a cycle-consistency loss

$$\mathcal{L}_{\text{cycle}}(G, F) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] \\ + \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1] \quad (5)$$

where the reconstructed image $F(G(x))$ is closely matched to the input image x, and is similar to the generated image $G(F(y))$ and the input image y. This could lead to generators to further reduce the space of possible mappings.

We also adopt the cycle-consistency loss in the proposed attention-guided generation Schemes I and II. However, we have modified it for the proposed models.

1) Attention-Guided Generation Cycle I: For the proposed attention-guided generation Scheme I, we should push back the generated image $G(x)$ in (1) to the original domain. Thus, we introduce another generator F, which has a similar structure to the generator G (see Fig. 2). Different from CycleGAN, the proposed F tries to generate a content mask C_x and attention mask A_x. Therefore, we fuse both masks and the generated image $G(x)$ to reconstruct the original input image x. This process can be formulated as

$$F(G(x)) = C_x * A_x + G(x) * (1 - A_x) \quad (6)$$

where the reconstructed image $F(G(x))$ should be very close to the original one x. For image y, we can reconstruct it by using $G(F(y)) = C_y * A_y + F(y) * (1 - A_y)$, and the recovered image $G(F(y))$ should be very close to y.

2) Attention-Guided Generation Cycle II: For the proposed attention-guided generation Scheme II, after generating the result $G(x)$ by the generator G in (4), we should push $G(x)$ back to the original domain to reduce the space of possible mappings. Thus, we have another generator F, which is very different from the one in Scheme I. F has a similar structure to the generator G and also consists of three subnets, i.e., a parameter-sharing encoder F_E, an attention mask generator F_A, and a content mask generator F_C (see Fig. 3). F_C tries to generate $n - 1$ content masks (i.e., $\{C_x^f\}_{f=1}^{n-1}$) and F_A tries to generate n attention masks of both background and foreground (i.e., A_x^b and $\{A_x^f\}_{f=1}^{n-1}$). Then we fuse these masks and the generated image $G(x)$ to reconstruct the original input image x. This process can be formulated as

$$F(G(x)) = \sum_{f=1}^{n-1}\left(C_x^f * A_x^f\right) + G(x) * A_x^b \qquad (7)$$

where the reconstructed image $F(G(x))$ should be very close to the original one x. For image y, we have the cycle $G(F(y)) = \sum_{f=1}^{n-1}(C_y^f * A_y^f) + F(y) * A_y^b$, and the recovered image $G(F(y))$ should be very close to y.

C. Attention-Guided Discriminator

Equation (1) constrains the generators to act only on the attended regions. However, the discriminators currently consider the whole image. More specifically, the vanilla discriminator D_Y takes the generated image $G(x)$ or the real image y as input and tries to distinguish them, this adversarial loss can be formulated as follows:

$$\mathcal{L}_{\mathrm{GAN}}(G, D_Y) = \mathbb{E}_{y \sim p_{\mathrm{data}}(y)}\left[\log D_Y(y)\right]$$
$$+ \mathbb{E}_{x \sim p_{\mathrm{data}}(x)}\left[\log(1 - D_Y(G(x)))\right] \qquad (8)$$

where G tries to minimize the adversarial loss $\mathcal{L}_{\mathrm{GAN}}(G, D_Y)$ while D_Y tries to maximize it. The target of G is to generate an image $G(x)$ that looks similar to the images from domain Y, while D_Y aims to distinguish between the generated image $G(x)$ and the real one y. A similar adversarial loss of the generator F and its discriminator D_X is defined as

$$\mathcal{L}_{\mathrm{GAN}}(F, D_X) = \mathbb{E}_{x \sim p_{\mathrm{data}}(x)}\left[\log D_X(x)\right]$$
$$+ \mathbb{E}_{y \sim p_{\mathrm{data}}(y)}\left[\log(1 - D_X(F(y)))\right] \qquad (9)$$

where D_X tries to distinguish between the generated image $F(y)$ and the real one x.

To add an attention mechanism to the discriminator, we propose two ADs. The AD is structurally the same as the vanilla discriminator but it also takes the attention mask as input. The AD D_{YA}, tries to distinguish between the fake image pairs $[A_y, G(x)]$ and the real image pairs $[A_y, y]$. Moreover, we propose the attention-guided adversarial loss for training the ADs. The min-max game between the AD D_{YA} and the generator G is performed through the following objective functions:

$$\mathcal{L}_{\mathrm{AGAN}}(G, D_{\mathrm{YA}}) = \mathbb{E}_{y \sim p_{\mathrm{data}}(y)}\left[\log D_{\mathrm{YA}}([A_y, y])\right]$$
$$+ \mathbb{E}_{x \sim p_{\mathrm{data}}(x)}\left[\log\left(1 - D_{\mathrm{YA}}([A_y, G(x)])\right)\right] \qquad (10)$$

where D_{YA} aims to distinguish between the generated image pairs $[A_y, G(x)]$ and the real image pairs $[A_y, y]$. We also have another loss $\mathcal{L}_{\mathrm{AGAN}}(F, D_{\mathrm{XA}})$ for discriminator D_{XA} and generator F, where D_{XA} tries to distinguish the fake image pairs $[A_x, F(y)]$ and the real image pairs $[A_x, x]$. In this way, the discriminators can focus on the most discriminative content and ignore the unrelated content.

Note that the proposed AD is only used in Scheme I. In preliminary experiments, we also used the proposed AD in Scheme II but did not observe improved performance. The reason could be that Scheme II has enough ability to learn the most discriminative content between the source and target domains.

D. Optimization Objective

The optimization objective of the proposed attention-guided generation Scheme II can be expressed as

$$\mathcal{L} = \mathcal{L}_{\mathrm{GAN}} + \lambda_{\mathrm{cycle}} * \mathcal{L}_{\mathrm{cycle}} + \lambda_{\mathrm{id}} * \mathcal{L}_{\mathrm{id}} \qquad (11)$$

where $\mathcal{L}_{\mathrm{GAN}}$, $\mathcal{L}_{\mathrm{cycle}}$ and $\mathcal{L}_{\mathrm{id}}$ are GAN, cycle-consistency, and identity preserving loss [38], respectively. λ_{cycle} and λ_{id} are parameters controlling the relative relation of each term. The optimization objective of the proposed attention-guided generation Scheme I can be expressed

$$\mathcal{L} = \lambda_{\mathrm{cycle}} * \mathcal{L}_{\mathrm{cycle}} + \lambda_{\mathrm{pixel}} * \mathcal{L}_{\mathrm{pixel}}$$
$$+ \lambda_{\mathrm{gan}} * (\mathcal{L}_{\mathrm{GAN}} + \mathcal{L}_{\mathrm{AGAN}}) + \lambda_{\mathrm{tv}} * \mathcal{L}_{\mathrm{tv}} \qquad (12)$$

where $\mathcal{L}_{\mathrm{GAN}}$, $\mathcal{L}_{\mathrm{AGAN}}$, $\mathcal{L}_{\mathrm{cycle}}$, $\mathcal{L}_{\mathrm{tv}}$ and $\mathcal{L}_{\mathrm{pixel}}$ are GAN, attention-guided GAN, cycle-consistency, attention, and pixel loss (PL), respectively. λ_{gan}, λ_{cycle}, λ_{pixel} and λ_{tv} are parameters controlling the relative relation of each term. In the following, we will introduce attention loss (AL) and PL. Note that both losses are only used in Scheme I since the generator needs stronger constraints than those in Scheme II.

When training AttentionGAN, we do not have ground-truth annotation for the attention masks. They are learned from the resulting gradients of both AGs and discriminators and the rest of the losses. However, the attention masks can easily saturate to 1 causing the AGs to have no effect. To prevent this situation, we perform a Total Variation regularization over attention masks A_x and A_y. The AL of A_x thus can be defined as

$$\mathcal{L}_{\mathrm{tv}} = \sum_{w,h=1}^{W,H} |A_x(w + 1, h, c) - A_x(w, h, c)|$$
$$+ |A_x(w, h + 1, c) - A_x(w, h, c)| \qquad (13)$$

where W and H are the width and height of A_x.

Moreover, to reduce the changes and constrain the generator in Scheme I, we adopt the PL between the input images and the generated images. This loss can be regarded as another form of the identity preserving loss. We express this loss as

$$\mathcal{L}_{\mathrm{pixel}}(G, F) = \mathbb{E}_{x \sim p_{\mathrm{data}}(x)}[\|G(x) - x\|_1]$$
$$+ \mathbb{E}_{y \sim p_{\mathrm{data}}(y)}[\|F(y) - y\|_1]. \qquad (14)$$

We adopt $L1$ distance as loss measurement in PL.

E. Implementation Details

1) Network Architecture: For a fair comparison, we use the generator architecture from CycleGAN [3]. We have slightly modified it for our task. Scheme I takes a three-channel RGB image as input and outputs a one-channel attention mask and a three-channel content mask. Scheme II takes a three-channel RGB image as input and outputs n attention masks and $n-1$ content masks, thus we fuse all of these masks and the input image to produce the final results. Specifically, the attention generator in Fig. 3 consists of $u128$, $u64$, and $c7s1-10$, where $U.K.$ denotes a 3×3 fractional-strided-Convolution-InstanceNorm-rectified linear unit (ReLU) layer with k filters and stride $1/2$; $c7s1-k$ denote a 7×7 Convolution-InstanceNorm-ReLU layer with k filters and stride 1. The output of the attention generator is represented by ten attention masks; we choose the first one as the background attention mask, and the rest as the foreground attention masks. We set $n = 10$ in our experiments. For the vanilla discriminator, we employ the discriminator architecture from [3]. We employ the same architecture as the proposed AD except the AD takes an attention mask and an image as input, while the vanilla discriminator only takes an image as input.

2) Training Strategy: Our innovation lies in the structure of the proposed model rather than the training approach. Thus, we follow the standard optimization method from [1], [3] to optimize the proposed AttentionGAN for fair comparisons, i.e., we alternate between one gradient descent step on the generators, then one step on discriminators. To slow down the rate of discriminators relative to generators, we follow CycleGAN [3] and divide the objective by 2 while optimizing the discriminators. We also use a least square loss to stabilize our model during the training procedure, which is more stable than the negative log-likelihood objective. Moreover, we use a history of the generated images to update discriminators similar to CycleGAN.

IV. EXPERIMENTS

A. Experimental Setups

1) Datasets: We follow previous works [3], [10], [12], [18] and employ eight popular datasets to evaluate the proposed AttentionGAN, including four face image datasets (i.e., CelebA [39], Radboud faces database (RaFD) [40], AR Face [41], and Selfie2Anime [12]) and four natural image datasets (i.e., Horse2Zebra [3], Apple2Orange [3], Maps [3], and Style Transfer [3]). More details about these datasets can be found in the corresponding articles. Note that these eight datasets are widespread for evaluating unsupervised image translation models since they contain diverse scenarios, including natural images, face images, artistic paintings, and aerial photographs. Therefore, it is challenging to build a general and versatile model to handle all of these tasks on various datasets.

2) Parameter Settings: For all datasets, images are rescaled to 256×256. We do left-right flip and random crop for data augmentation. We set the number of image buffer to 50 similar in [3]. We use the Adam optimizer [42] with the momentum terms $\beta_1 = 0.5$ and $\beta_2 = 0.999$. For fair comparisons, we follow [10] and set $\lambda_{\text{cycle}} = 10$, $\lambda_{\text{gan}} = 0.5$, $\lambda_{\text{pixel}} = 1$ and $\lambda_{\text{tv}} = 1e - 6$ in (12). Moreover, we follow [3] and set $\lambda_{\text{cycle}} = 10$, $\lambda_{\text{id}} = 0.5$ in (11). Previous works [3], [10] have verified the effectiveness of these parameters in the unsupervised image-to-image translation task.

3) Competing Models: We consider several state-of-the-art image translation models as our baselines. 1) Unpaired image translation methods: CycleGAN [3], DualGAN [4], DIAT [43], DiscoGAN [5], DistanceGAN [32], Dist.+Cycle [32], Self Dist. [32], ComboGAN [33], UNIT [44], MUNIT [14], DRIT [15], GANimorph [6], CoGAN [45], SimGAN [46], Feature loss+GAN [46] (a variant of SimGAN). 2) Paired image translation methods: BicycleGAN [38], Pix2pix [2], Encoder-Decoder [2]. 3) Class label, object mask, or attention-guided image translation methods: IcGAN [23], StarGAN [18], ContrastGAN [7], GANimation [47], RA [48], UAIT [11], U-GAT-IT [12], SAT [13], TransGaGa [16], DA-GAN [17]. 4) CGANs methods: BiGAN/ALI [49], [50]. Note that the fully supervised Pix2pix, Encoder-Decoder (Enc.-Decoder), and BicycleGAN are trained on paired data. Since BicycleGAN can generate several different outputs with one single input image, we randomly select one output from them for fair comparisons. To reimplement ContrastGAN, we use OpenFace [51] to obtain the face masks as extra input data.

4) Evaluation Metrics: We follow CycleGAN [3] and adopt Amazon Mechanical Turk (AMT) perceptual studies to evaluate the generated images. Moreover, to seek a quantitative measure that does not require human participation, Peak Signal-to-Noise Ratio (PSNR), Kernel Inception Distance (KID) [52] and Fréchet Inception Distance (FID) [53] are employed according to different translation tasks. Specifically, PSNR measures the pixel-level similarity between the generated image and the real one. Both KID and FID evaluate the generated images from a high-level feature space. These three metrics are the most mainstream methods to measure the images generated by unsupervised image translation models.

B. Experimental Results

1) Ablation Study: We first conduct extensive ablation studies to evaluate the effectiveness of the proposed methods.

a) Analysis of Model Component: To evaluate the components of our AttentionGAN, we first conduct extensive ablation studies. We gradually remove components of the proposed AttentionGAN, i.e., AD, AG, AL, and PL. The results of AMT and PSNR on AR Face are shown in Table I. We find that removing one of them substantially degrades the results, which means all of them are critical to our results. We conduct more experiments on the more challenging Horse2Zebra dataset to validate the effectiveness of each component of the proposed AttentionGAN. The results are shown in Table I, we can see that the proposed method achieves similar results both on Horse2Zebra and on AR Face datasets. We also provide qualitative results of AR Face in Fig. 4. Note that without AG we cannot generate both attention and content masks.

We conduct more ablation studies to verify how much of the improvement is due to the three designs: 1) employing two

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

8   IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS

TABLE I

ABLATION STUDY OF ATTENTIONGAN ON AR FACE AND HORSE2ZEBRA

Setup	AR Face		Horse2Zebra
	AMT $\uparrow$	PSNR $\uparrow$	FID $\downarrow$
Full	**12.8**	**14.9187**	**68.55**
Full - AD	10.2	14.6352	73.63
Full - AD - AG	3.2	14.4646	90.78
Full - AD - PL	8.9	14.5128	78.55
Full - AD - AL	6.3	14.6129	82.35
Full - AD - PL - AL	5.2	14.3287	86.14

Fig. 4.   Ablation study of AttentionGAN on AR Face.

Fig. 5.   Ablation study of AttentionGAN on Horse2Zebra.

Fig. 6.   Comparison results of the proposed attention-guided generation Schemes I and II.

TABLE II

ABLATION STUDY OF ATTENTIONGAN ON HORSE2ZEBRA

Setup	Full	Full-(1)	Full-(2)	Full-(3)
FID $\downarrow$	**68.55**	101.80	77.29	79.33
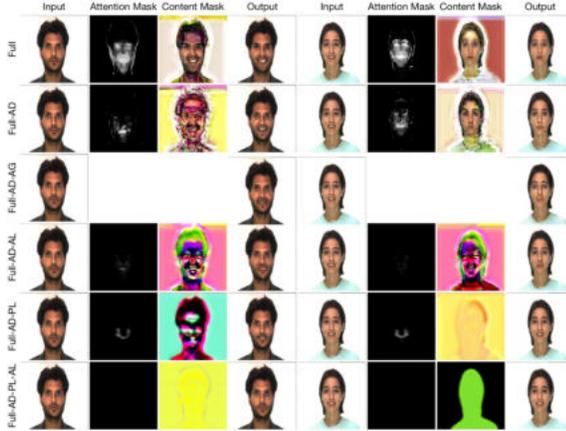
separate subnetworks to generate content mask and attention mask. 2) producing both foreground mask and background mask. 3) learning a set of intermediate content masks. Results are shown in Table II and Fig. 5, where "Full" denotes our full model, "Full-(1)' denotes that we use a shared network to generate both content mask and attention mask, "Full-(2)" denotes that we only produce foreground masks, "Full-(3)" denotes that we learn only one content mask. As can be seen in Table II that removing one of these three designs significantly decreases the generation performance, validating all of them are critical to learn a better unsupervised image-to-image translation mapping. Moreover, "Full" generates much better results than others (see Fig. 5). Specifically, "Full-(3)" generates a lot of visual artifacts (shown in the red boxes), but "Full" does not. "Full-(1)" and "Full-(2)" change the background while translating, but "Full" does not. In addition, both "Full-(1)" and "Full-(2)" also produce some artifacts, which make the generated zebras look blurry (see the red boxes near the zebras of both "Full-(1)" and "Full-(2)").

Lastly, we observe that "Full-(1)" in Table II performs much worse "Full-AD-AG" in Table I. "Full-(1)" denotes that we use a shared network to generate both content mask and attention mask. Since there is a big difference between attention mask (i.e., grayscale image) and content mask (i.e., RGB image), if we use one network to learn both, the two of them will pollute each other, causing both to learn poorly. In this case, the result of "Full-(1)" will be worse than that of "Full-AD-AG." This means that the design "(1)" is a prerequisite for both designs "(2)" and "(3)," and the design "(1)" is also the most important design among the three. In the end, the performance will be the best when only the three are used together (see Table II).
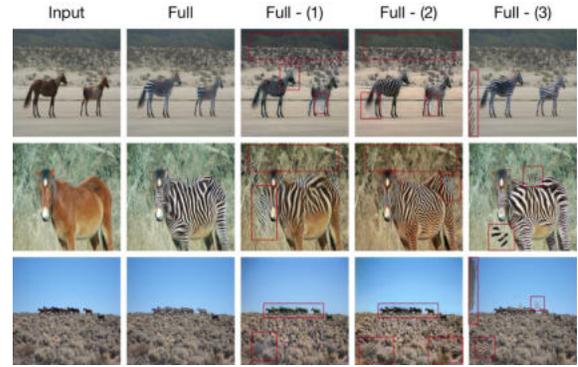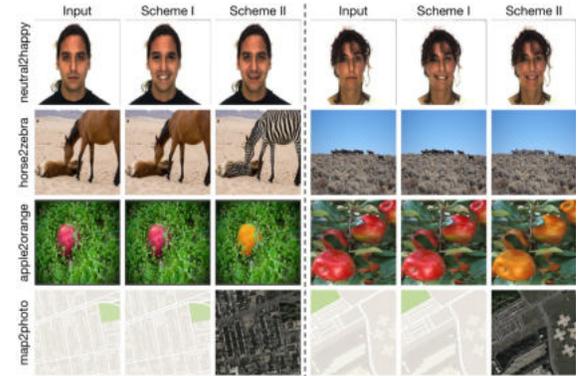
b) Attention-Guided Generation Scheme I versus II: Moreover, we present the comparison results of the proposed attention-guided generation Schemes I and II. Scheme I is used in our conference paper [10]. Scheme II is a refined version proposed in this article. Comparison results are shown in Fig. 6. We observe that Scheme I generates good results on facial expression transfer, however, it generates identical images with the inputs on other tasks, e.g., horse-to-zebra translation, apple to orange translation, and map to aerial photo translation. However, the proposed attention-guided generation Scheme II can handle all of these tasks, which validates the effectiveness of the proposed Scheme II.

c) Analysis of Hyperparameters: We investigate the influence of λ_{cycle} and λ_{id} in (11) on the performance of our method. As can be seen in Tables III and IV, when $\lambda_{\text{cycle}} = 10$ and $\lambda_{\text{id}} = 0.5$, the proposed method achieves the best FID score on Horse2Zebra, validating the effectiveness of our setting of hyperparameters. Moreover, to show that the proposed framework can stabilize GAN training, we illustrate the convergence loss of the proposed method in (11) on Horse2Zebra (see Fig. 7). We observe that the proposed method ensures a very fast yet stable convergence (around the 50th epoch).

2) Experiments on Face Images: We conduct extensive experiments on four datasets to validate AttentionGAN.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TANG *et al.*: ATTENTIONGAN: UNPAIRED IMAGE-TO-IMAGE TRANSLATION

9
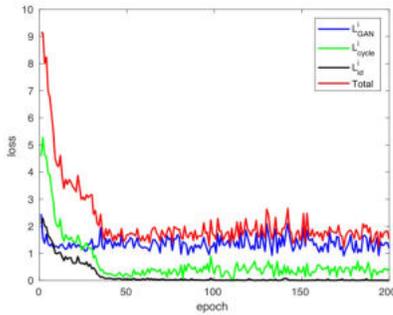
Fig. 7.   Convergence loss in (11) on Horse2Zebra.

TABLE III
INFLUENCE OF λ_{cycle} ON HORSE2ZEBRA

λ_{cycle}	0.1	1	10	100	1000
FID ↓	137.74	79.10	**68.55**	244.27	244.22

TABLE IV
INFLUENCE OF λ_{id} ON HORSE2ZEBRA

λ_{id}	0.01	0.1	0.5	1	10
FID ↓	106.76	83.67	**68.55**	117.97	244.48

a) Results on AR Face: The results of neutral ↔ happy expression translation on AR Face are shown in Fig. 8. Clearly, the results of Dist.+Cycle and Self Dist. cannot even generate human faces. DiscoGAN produces identical results regardless of the input faces suffering from mode collapse. The results of DualGAN, DistanceGAN, StarGAN, Pix2pix, Encoder–Decoder, and BicycleGAN tend to be blurry, while ComboGAN and ContrastGAN produce the same identity but without expression changing. CycleGAN generates sharper images, but the details of the generated faces are not convincing. Compared with all the baselines, the results of AttentionGAN are correct, smoother, and with more details.

b) Results on CelebA: We conduct both facial expression translation and facial attribute transfer tasks on this dataset. The facial expression translation task on this dataset is more challenging than the AR Face dataset since the background of this dataset is very complicated. Note that this dataset does not provide paired data, thus we cannot conduct experiments with supervised methods, i.e., Pix2pix, BicycleGAN, and Encoder-Decoder. The results compared with other baselines are shown in Fig. 9. We observe that only the proposed AttentionGAN produces photo-realistic faces with correct expressions. The reason could be that methods without attention cannot learn the most discriminative part and the unwanted part. All existing methods failed to generate novel expressions, which means they treat the whole image as the unwanted part, while the proposed AttentionGAN can learn novel expressions by distinguishing the discriminative part from the unwanted part.

Moreover, our model can be easily extended to solve multi domain image-to-image translation problems. To control multiple domains in one single model we employ the domain classification loss proposed in StarGAN. Thus, we follow StarGAN and conduct facial attribute transfer on this dataset to evaluate the proposed AttentionGAN. The results compared with StarGAN are shown in Fig. 10. We observe that the

Fig. 8.   Results of facial expression transfer on AR Face.

Fig. 9.   Results of facial expression transfer on CelebA.
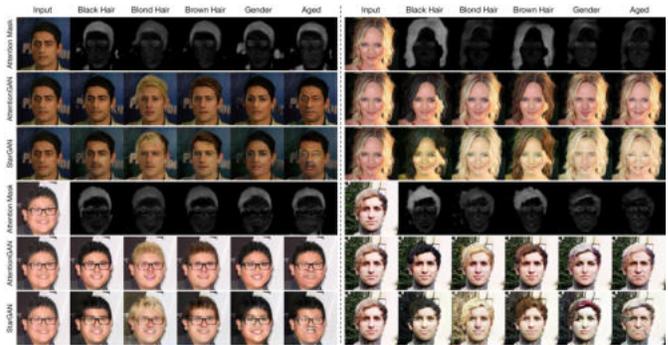
Fig. 10.   Results of facial attribute transfer on CelebA.

proposed AttentionGAN achieves visually better results than StarGAN without changing the background.

c) Results on RaFD: We follow StarGAN and conduct diverse facial expression translation on this dataset. The results compared to the baselines DIAT, CycleGAN, IcGAN, StarGAN, and GANimation are shown in Fig. 11. We observe that the proposed AttentionGAN achieves better results than DIAT, CycleGAN, StarGAN, and IcGAN. For GANimation, we follow the authors' instructions and use OpenFace [51] to obtain the action units of each face as extra input data. Note that the proposed method generates competitive results compared to GANimation but GANimation needs action units annotations as extra training data limiting its practical application. More importantly, GANimation cannot handle other generative tasks such as facial attribute transfer as shown in Fig. 10.

d) Results of Selfie2Anime: We follow U-GAT-IT [12] and conduct selfie to anime translation on the Selfie2Anime dataset. The results compared with state-of-the-art methods are shown in Fig. 12. We observe that the proposed AttentionGAN achieves better results than other baselines.

We conclude that even though the subjects in these four datasets have different races, poses, styles, skin colors, illumination conditions, occlusions, and complex backgrounds, our method consistently generates sharper images with correct expressions/attributes than existing models. We also observe that our AttentionGAN performs better than other baselines when training data are limited (see Fig. 9), which also shows that our method is very robust.

e) Quantitative Comparison: We also provide quantitative results of these tasks. As shown in Table V, AttentionGAN achieves the best results compared with competing models including fully supervised methods (e.g., Pix2pix,
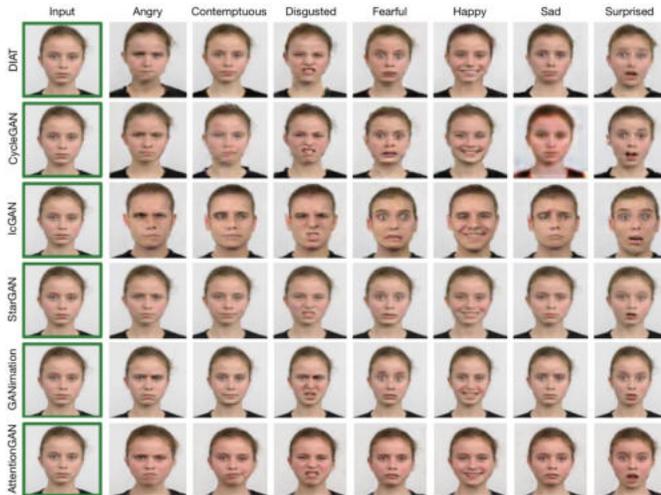
Fig. 11.   Results of facial expression transfer on RaFD.

Fig. 12.   Different methods for mapping selfie to anime.

TABLE V

QUANTITATIVE COMPARISON ON THE FACIAL EXPRESSION TRANSLATION
TASK. FOR BOTH AMT AND PSNR, HIGH IS BETTER

| Method | AR Face | | CelebA |
	AMT ↑	PSNR ↑	AMT ↑
CycleGAN [3]	10.2	14.8142	34.6
DualGAN [4]	1.3	14.7458	3.2
DiscoGAN [5]	0.1	13.1547	1.2
ComboGAN [33]	1.5	14.7465	9.6
DistanceGAN [32]	0.3	11.4983	1.9
Dist.+Cycle [32]	0.1	3.8632	1.3
Self Dist. [32]	0.1	3.8674	1.2
StarGAN [18]	1.6	13.5757	14.8
ContrastGAN [7]	8.3	14.8495	25.1
Pix2pix [2]	2.6	14.6118	-
Enc.-Decoder [2]	0.1	12.6660	-
BicycleGAN [38]	1.5	14.7914	-
AttentionGAN	**12.8**	**14.9187**	**38.9**

Encoder-Decoder, and BicycleGAN) and mask-conditional methods (e.g., ContrastGAN). Next, following StarGAN, we perform a user study using AMT to assess attribute transfer tasks on CelebA. The results comparing the state-of-the-art methods are shown in Table VI. We observe that AttentionGAN achieves significantly better results than all the baselines. Moreover, we follow U-GAT-IT [12] and adopt KID to evaluate the generated images on a selfie to anime translation. The results are shown in Table VII, we observe that our AttentionGAN achieves the best results compared with baselines except for U-GAT-IT. However, U-GAT-IT needs

TABLE VI

AMT RESULTS OF THE FACIAL ATTRIBUTE TRANSFER TASK ON CELEBA.
FOR THIS METRIC, HIGHER IS BETTER

Method	Hair Color	Gender	Aged
DIAT [43]	3.5	21.1	3.2
CycleGAN [3]	9.8	8.2	9.4
IcGAN [23]	1.3	6.3	5.7
StarGAN [18]	24.8	28.8	30.8
AttentionGAN	**60.6**	**35.6**	**50.9**

TABLE VII

KID × 100 ± STD. × 100 OF THE SELFIE TO ANIME TRANSLATION TASK.
FOR THIS METRIC, LOWER IS BETTER

Method	Selfie to Anime
U-GAT-IT [12]	**11.61 ± 0.57**
CycleGAN [3]	13.08 ± 0.49
UNIT [44]	14.71 ± 0.59
MUNIT [14]	13.85 ± 0.41
DRIT [15]	15.08 ± 0.62
AttentionGAN	12.14 ± 0.43

TABLE VIII

OVERALL MODEL CAPACITY ON RAFD ($m = 8$)

Method	# Models	# Parameters
Pix2pix [2]	m(m-1)	57.2M×56
Encoder-Decoder [2]	m(m-1)	41.9M×56
BicycleGAN [38]	m(m-1)	64.3M×56
CycleGAN [3]	m(m-1)/2	52.6M×28
DualGAN [4]	m(m-1)/2	178.7M×28
DiscoGAN [5]	m(m-1)/2	16.6M×28
DistanceGAN [32]	m(m-1)/2	52.6M×28
Dist.+Cycle [32]	m(m-1)/2	52.6M×28
Self Dist. [32]	m(m-1)/2	52.6M×28
U-GAT-IT [12]	m(m-1)/2	134.0M×28
ComboGAN [33]	m	14.4M×8
StarGAN [18]	1	53.2M×1
ContrastGAN [7]	1	52.6M×1
AttentionGAN (Scheme I) [10]	1	52.6M×1
AttentionGAN (Scheme II)	1	53.5M×1

to adopt two auxiliary classifiers to obtain attention makes. Moreover, U-GAT-IT uses four discriminators (i.e., two global discriminators and two local discriminators), while we only use two discriminators. Both designs of U-GAT-IT significantly increase the number of network parameters and training time (see Table VIII). Moreover, compared with Scheme I, Scheme II only adds a little amount of parameters (0.9M), but the performance is significantly improved (see Fig. 6). This indicates that the performance improvement of Scheme II does not come from simply increasing network parameters.

f) Visualization of Learned Attention and Content Masks: Instead of regressing a full image, our generator outputs two masks, a content mask and an attention mask. We also visualize both masks on RaFD and CelebA in Figs. 13 and 14, respectively. In Fig. 13, we observe that different expressions generate different attention masks and content masks. The proposed method makes the generator focus only on those discriminative regions of the image that are responsible for synthesizing the novel expression. The attention masks mainly focus on the eyes and mouth, which means these parts are important for generating novel expressions. The proposed method also keeps the other elements of the image or unwanted parts untouched. In Fig. 13, the unwanted part
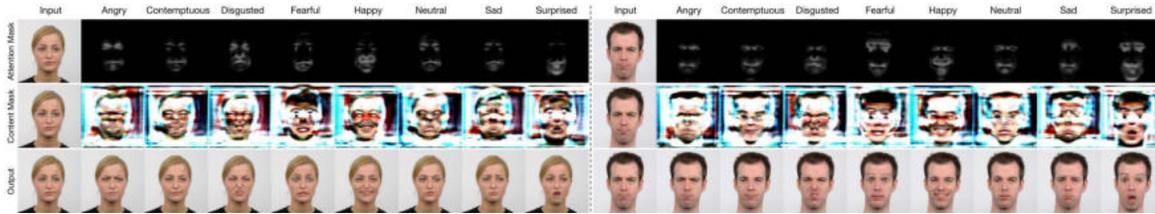
Fig. 13. Attention and content masks on RaFD.

Fig. 14. Attention and content masks on CelebA.
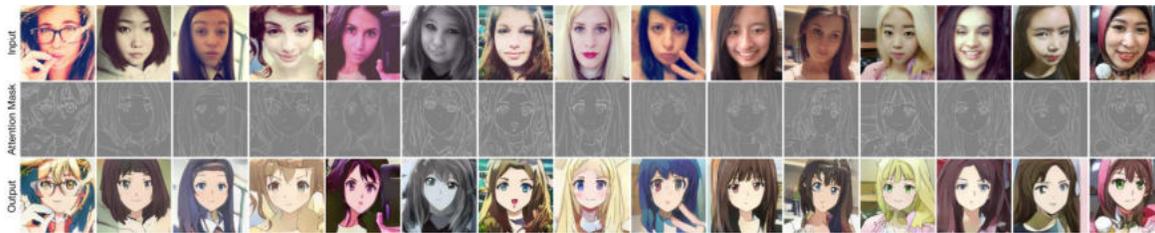
Fig. 15. Attention mask on the selfie to anime translation task.
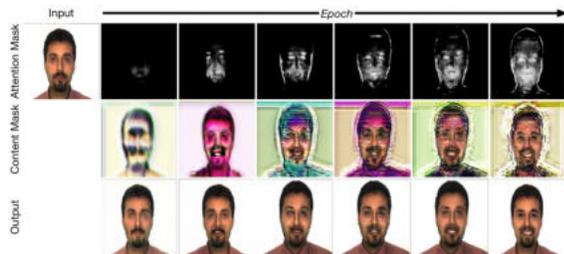
Fig. 16. Evolution of attention masks and content masks.

Fig. 17. Different methods for mapping horse to zebra.

are hair, cheeks, clothes, and background, which means these parts have no contribution in generating novel expressions. In Fig. 14, we observe that different facial attributes also generate different attention masks and content masks, which further validates our initial motivations. More attention masks generated by AttentionGAN on the facial attribute transfer task are shown in Fig. 10. Note that the proposed AttentionGAN can handle the geometric changes between source and target domains, such as selfie to anime translation. Therefore, we show the learned attention masks on a selfie to anime translation task to interpret the generation process in Fig. 15.

We also present the generation of both attention and content masks on the AR Face dataset epoch-by-epoch in Fig. 16. We see that with the number of training epochs increases, the attention mask and the result become better, and the attention masks correlate well with image quality, which demonstrates the proposed AttentionGAN is effective.

g) Comparison of Model Parameters: The number of models for m image domains and the number of model parameters on RaFD are shown in Table VIII. Note that our generation performance is much better than these baselines and the number of parameters is also comparable with
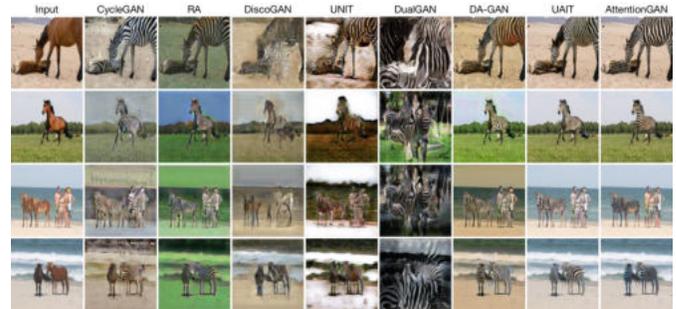
ContrastGAN, while ContrastGAN requires object masks as extra data.

3) Experiments on Natural Images: We also conduct experiments on four datasets to evaluate AttentionGAN.

a) Results of Horse2Zebra: The results of horse-to-zebra translation compared with CycleGAN, RA, DiscoGAN, UNIT, DualGAN, DA-GAN, and UAIT are shown in Fig. 17. We observe that DiscoGAN, UNIT, DualGAN generate blurred results. CycleGAN, RA, and DA-GAN can generate the corresponding zebras, however, the background of images produced by these models has also been changed. Both UAIT and the proposed AttentionGAN generate the corresponding zebras without changing the background. By carefully examining the translated images from both UAIT and the proposed AttentionGAN, we observe that AttentionGAN achieves slightly better results than UAIT as shown in the first and the third rows of Fig. 17. Our method produces better stripes on the body of the lying horse than UAIT as shown in the first row. In the third row, the proposed method generates fewer stripes on the body of the people than UAIT.
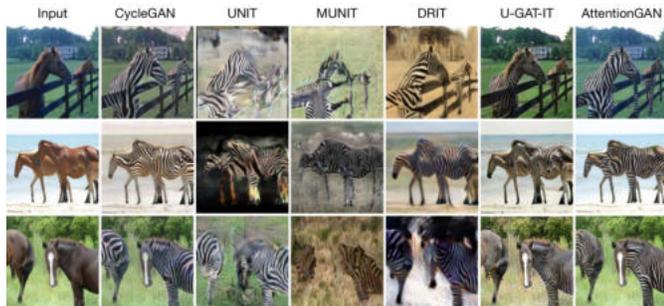
Fig. 18.   Different methods for mapping horse to zebra.

Moreover, we compare the proposed method with Cycle-GAN, UNIT, MUNIT, DRIT, and U-GAT-IT in Fig. 18. We can see that UNIT, MUNIT and DRIT generate blurred images with many visual artifacts. CycleGAN can produce the corresponding zebras, however, the background of the images has also been changed. U-GAT-IT and AttentionGAN can produce better results than other approaches. However, if we look closely at the results generated by both methods, we observe that U-GAT-IT slightly changes the background, while the proposed AttentionGAN perfectly keeps the background unchanged. For instance, as can be seen from the results of the first line, U-GAT-IT produces a darker background than the background of the input image in Fig. 18. While the background color of the generated images by U-GAT-IT is lighter than the input images as shown in the second and third rows in Fig. 18. Lastly, our method behaves the same as CycleGAN and U-GAT-IT, which greatly changes the wooden fence in the first row. The reason is that part of the horse is occluded by the fence. Our model tries to learn and translate a complete horse, thus it will change the fence. A feasible solution is to introduce an additional object mask [7] or instance mask [37] to alleviate this phenomenon, which will be further investigated in our future work.

We also compare the proposed AttentionGAN with GANimorph and CycleGAN in Fig. 1. We see that the proposed AttentionGAN demonstrates a significant qualitative improvement over both methods. Moreover, in the last row of Fig. 1, there is some blur in the bottom right corner of our result. The reason is that our generator regards that location as foreground content and translates it to the target domain. This phenomenon rarely occurs in other samples because this sample is more difficult to learn and translate.

The results of zebra to horse translation are shown in Fig. 19. We note that our method generates better results than all the baselines. In summary, AttentionGAN is able to better alter the object of interest than existing methods by modeling attention masks in unpaired image-to-image translation tasks without changing the background at the same time.

b) Results of Apple2Orange: The results compared with CycleGAN, RA, DiscoGAN, UNIT, DualGAN, TransGaGa, and UAIT are shown in Figs. 20 and 21. We see that RA, DiscoGAN, UNIT, and DualGAN generate blurred results with lots of visual artifacts. CycleGAN generates better results, however, we can see that the background and other unwanted objects have also been changed, e.g., the banana in the second

Fig. 19.   Different methods for mapping zebra to horse.
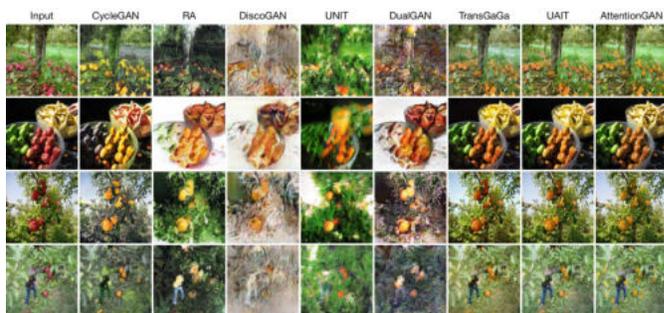
Fig. 20.   Different methods for mapping apple to orange.

Fig. 21.   Different methods for mapping orange to apple.

row of Fig. 20. Both UAIT and our AttentionGAN can generate much better results than other baselines. However, UAIT adds an attention network before each generator to achieve the translation of the relevant parts, which increases the number of network parameters.

c) Results of Maps: Qualitative results of both translation directions compared with existing methods are shown in Figs. 22 and 23, respectively. We note that BiGAN, CoGAN, SimGAN, Feature loss+GAN only generate blurred results with lots of visual artifacts. The results generated by our method are better than those generated by CycleGAN. Moreover, we compare AttentionGAN with the fully supervised Pix2pix, we see that AttentionGAN achieves comparable or even better results than Pix2pix as indicated in the black boxes in Fig. 23.

d) Results of Style Transfer: We also show the generation results of our AttentionGAN on the style transfer task. The results compared with the leading method, i.e., CycleGAN, are shown in Fig. 24. We observe that the proposed AttentionGAN generates much sharper and diverse results than CycleGAN.
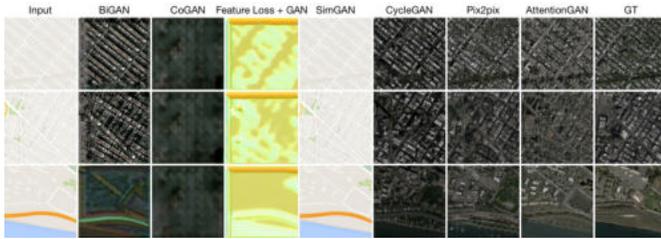
Fig. 22.   Different methods for mapping map to aerial photo.

Fig. 23.   Different methods for mapping aerial photo to map.

TABLE IX

KID $\times$ 100 $\pm$ Std. $\times$ 100 for Different Methods. For This Metric, Lower Is Better. Abbreviations: (H)orse, (Z)ebra (A)pple, (O)range

Method	H $\rightarrow$ Z	Z $\rightarrow$ H	A $\rightarrow$ O	O $\rightarrow$ A
DiscoGAN [5]	13.68 $\pm$ 0.28	16.60 $\pm$ 0.50	18.34 $\pm$ 0.75	21.56 $\pm$ 0.80
RA [48]	10.16 $\pm$ 0.12	10.97 $\pm$ 0.26	12.75 $\pm$ 0.49	13.84 $\pm$ 0.78
DualGAN [4]	10.38 $\pm$ 0.31	12.86 $\pm$ 0.50	13.04 $\pm$ 0.72	12.42 $\pm$ 0.88
UNIT [44]	11.22 $\pm$ 0.24	13.63 $\pm$ 0.34	11.68 $\pm$ 0.43	11.76 $\pm$ 0.51
CycleGAN [3]	10.25 $\pm$ 0.25	11.44 $\pm$ 0.38	8.48 $\pm$ 0.53	9.82 $\pm$ 0.51
UAIT [11]	6.93 $\pm$ 0.27	8.87 $\pm$ 0.26	**6.44 $\pm$ 0.69**	5.32 $\pm$ 0.48
AttentionGAN	**2.03 $\pm$ 0.64**	**6.48 $\pm$ 0.51**	10.03 $\pm$ 0.66	**4.38 $\pm$ 0.42**

TABLE X

Preference Score on Both Horse to Zebra and Apple to Orange Translation Tasks. For This Metric, Higher Is Better. Abbreviations: (H)orse, (Z)ebra (A)pple, (O)range

Method	H $\rightarrow$ Z	Z $\rightarrow$ H	A $\rightarrow$ O	O $\rightarrow$ A
UNIT [44]	1.83	3.57	2.67	2.91
MUNIT [14]	3.86	5.61	6.23	4.98
DRIT [15]	1.27	2.14	1.09	1.97
CycleGAN [3]	22.12	21.65	26.76	26.14
U-GAT-IT [12]	33.17	31.39	30.05	29.89
AttentionGAN	**37.75**	**35.64**	**33.20**	**34.11**

e) Quantitative Comparison: We follow UAIT [11] and adopt KID [52] to evaluate the generated images by different methods. The results of horse $\leftrightarrow$ zebra and apple $\leftrightarrow$ orange are shown in Table IX. We observe that AttentionGAN achieves the lowest KID on H $\rightarrow$ Z, Z $\rightarrow$ H and O $\rightarrow$ A translation tasks. We note that both UAIT and CycleGAN produce a lower KID score on apple to orange translation (A $\rightarrow$ O) but have poor quality image generation as shown in Fig. 20.

Moreover, following U-GAT-IT [12], we conducted a perceptual study to evaluate the generated images on Horse2Zebra and Apple2Orange. Specifically, 50 participants are shown the generated images from different methods including AttentionGAN with the source image, and asked to select the best generated image to the target domain. The results are shown in Table X. We observe that the proposed method outperforms other baselines including U-GAT-IT on these four tasks.

Next, we follow SAT [13] and adopt FID [53] to measure the distance between generated samples and target samples. We compute FID for horse-to-zebra translation and the results

TABLE XI

Results of FID on the Horse to Zebra Translation Task. For This Metric, Lower Is Better

Method	Horse to Zebra
UNIT [44]	241.13
CycleGAN [3]	109.36
DA-GAN [17]	103.42
TransGaGa [16]	95.81
SAT (Before Attention) [13]	98.90
SAT (After Attention) [13]	128.32
AttentionGAN	**68.55**

TABLE XII

AMT "Real Versus Fake" Results on Maps $\leftrightarrow$ Aerial Photos. For This Metric, Higher Is Better

Method	Map to Photo	Photo to Map
CoGAN [45]	0.8 $\pm$ 0.7	1.3 $\pm$ 0.8
BiGAN/ALI [49], [50]	3.2 $\pm$ 1.5	2.9 $\pm$ 1.2
SimGAN [46]	0.4 $\pm$ 0.3	2.2 $\pm$ 0.7
Feature loss + GAN [46]	1.1 $\pm$ 0.8	0.5 $\pm$ 0.3
CycleGAN [3]	27.9 $\pm$ 3.2	25.1 $\pm$ 2.9
Pix2pix [2]	33.7 $\pm$ 2.6.	29.4 $\pm$ 3.2
AttentionGAN	**35.18 $\pm$ 2.9**	**32.4 $\pm$ 2.5**
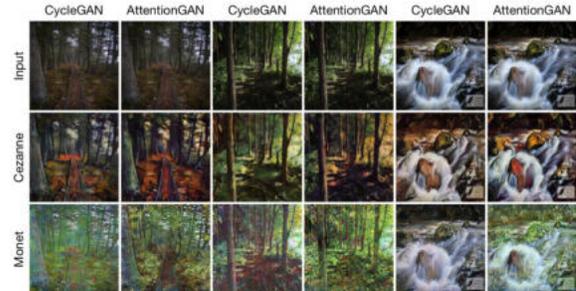
Fig. 24.   Different methods for style transfer.

Fig. 25.   Attention masks on horse $\leftrightarrow$ zebra translation.

Fig. 26.   Attention masks on apple $\leftrightarrow$ orange translation.

compared with SAT, CycleGAN, DA-GAN, TransGaGa, and UNIT are shown in Table XI. We observe that the proposed model achieves significantly better FID than all baselines. We note that SAT with attention has worse FID than SAT without attention, which means using attention might have a negative effect on FID because there might be some correlations between foreground and background in the target domain when computing FID. While we did not observe such a negative effect on AttentionGAN. Qualitative comparison with SAT is shown in Fig. 27. We observe that the proposed AttentionGAN achieves better results than SAT.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

14   IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS
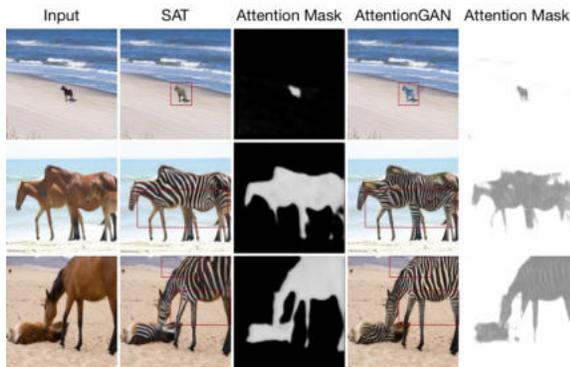
Fig. 27. Attention masks compared with SAT [13] on horse-to-zebra translation.

Fig. 28. Attention masks on aerial photo ↔ map translation.

Finally, we follow CycleGAN and adopt the AMT score to evaluate the generated images on the map ↔ aerial photo translation task. Participants were shown a sequence of pairs of images, one real image and one fake generated by our method or exiting methods, and asked to click on the image they thought was real. Comparison results of both translation directions are shown in Table XII. We observe that the proposed AttentionGAN generates the best results compared with the leading methods and can fool participants on around 1/3 of trials in both translation directions.

f) Visualization of Learned Attention Masks: The results of both horse ↔ zebra and apple ↔ orange translations are shown in Figs. 25 and 26, respectively. We see that our AttentionGAN is able to learn relevant image regions and ignore the background and other irrelevant objects. Moreover, we also compare with the most recent method, SAT [13], on the learned attention masks. Results are shown in Fig. 27. We observe that the attention masks learned by our method are much more accurate than those generated by SAT, especially in the boundary of attended objects. Thus our method generates a more photo-realistic object boundary than SAT in the translated images, as indicated in the red boxes in Fig. 27.

The results of map ↔ aerial photo translation are shown in Fig. 28. Note that although images of the source and target domains differ greatly in appearance, the images of both domains are structurally identical. Thus the learned attention masks highlight the shared layout and structure of both source and target domains. Thus, we can conclude that AttentionGAN can handle both images requiring large shape changes and images requiring holistic changes.

V. CONCLUSION

We propose a novel AttentionGAN for both unpaired image-to-image translation and multi domain image-to-image

translation tasks. The generators in AttentionGAN have a built-in attention mechanism, which can preserve the background of the input images and discover the most discriminative content between the source and target domains by producing attention masks and content masks. Then the attention masks, content masks, and the input images are combined to generate the target images with high quality. Extensive experimental results on several challenging tasks demonstrate that the proposed AttentionGAN can generate better results with more convincing details than numerous state-of-the-art methods.

REFERENCES

[1] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NeurIPS*, 2014.
[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.
[4] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2849–2857.
[5] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. ICML*, 2017, pp. 1857–1865.
[6] A. Gokaslan, V. Ramanujan, D. Ritchie, K. I. Kim, and J. Tompkin, "Improving shape deformation in unsupervised image-to-image translation," in *Proc. ECCV*, 2018, pp. 649–665.
[7] X. Liang, H. Zhang, L. Lin, and E. Xing, "Generative semantic manipulation with mask-contrasting GAN," in *Proc. ECCV*, 2018, pp. 558–573.
[8] X. Chen, C. Xu, X. Yang, and D. Tao, "Attention-GAN for object transfiguration in wild images," in *Proc. ECCV*, 2018, pp. 164–180.
[9] D. Kastaniotis, I. Ntinou, D. Tsourounis, G. Economou, and S. Fotopoulos, "Attention-aware generative adversarial networks (ATA-GANs)," in *Proc. IEEE 13th Image, Video, Multidimensional Signal Process. Workshop (IVMSP)*, Jun. 2018, pp. 1–5.
[10] H. Tang, D. Xu, N. Sebe, and Y. Yan, "Attention-guided generative adversarial networks for unsupervised image-to-image translation," in *Proc. IJCNN*, 2019, pp. 1–8.
[11] Y. A. Mejjati, C. Richardt, J. Tompkin, D. Cosker, and K. I. Kim, "Unsupervised attention-guided image to image translation," in *Proc. NeurIPS*, 2018, pp. 3693–3703.
[12] J. Kim, M. Kim, H. Kang, and K. Lee, "U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *Proc. ICLR*, 2020.
[13] C. Yang, T. Kim, R. Wang, H. Peng, and C.-C.-J. Kuo, "Show, attend, and translate: Unsupervised image translation with self-regularization and attention," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4845–4856, Oct. 2019.
[14] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. ECCV*, 2018, pp. 172–189.
[15] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. ECCV*, 2018, pp. 35–51.
[16] W. Wu, K. Cao, C. Li, C. Qian, and C. C. Loy, "TransGaGa: Geometry-aware unsupervised image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8012–8021.
[17] S. Ma, J. Fu, C. W. Chen, and T. Mei, "DA-GAN: Instance-level image translation by deep attention generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5657–5666.
[18] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.
[19] G. Liu, H. Tang, H. Latapie, J. Corso, and Y. Yan, "Cross-view exocentric to egocentric video synthesis," in *Proc. ACM MM*, 2021.

TANG *et al.*: ATTENTIONGAN: UNPAIRED IMAGE-TO-IMAGE TRANSLATION
15

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[20] H. Tang and N. Sebe, "Total generate: Cycle in cycle generative adversarial networks for generating human faces, hands, bodies, and natural scenes," *IEEE Trans. Multimedia*, early access, Jun. 24, 2021, doi: 10.1109/TMM.2021.3091847.

[21] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[22] H. Tang, H. Liu, and N. Sebe, "Unified generative adversarial networks for controllable image-to-image translation," *IEEE Trans. Image Process.*, vol. 29, pp. 8916–8929, 2020.

[23] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional GANs for image editing," in *Proc. NeurIPS Workshop*, 2016.

[24] H. Tang, D. Xu, G. Liu, W. Wang, N. Sebe, and Y. Yan, "Cycle in cycle generative adversarial networks for keypoint-guided image generation," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2052–2060.

[25] H. Tang, W. Wang, D. Xu, Y. Yan, and N. Sebe, "GestureGAN for hand gesture-to-gesture translation in the wild," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 774–782.

[26] H. Tang, S. Bai, P. H. Torr, and N. Sebe, "Bipartite graph reasoning gans for person image generation," in *Proc. BMVC*, 2020.

[27] H. Tang, S. Bai, L. Zhang, P. H. Torr, and N. Sebe, "XingGAN for person image generation," in *Proc. ECCV*, 2020, pp. 717–734.

[28] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, "Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2417–2426.

[29] H. Tang, S. Bai, and N. Sebe, "Dual attention GANs for semantic image synthesis," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1994–2002.

[30] H. Tang, D. Xu, Y. Yan, P. H. S. Torr, and N. Sebe, "Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7870–7879.

[31] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[32] S. Benaim and L. Wolf, "One-sided unsupervised domain mapping," in *Proc. NeurIPS*, 2017, pp. 752–762.

[33] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool, "Combo-GAN: Unrestrained scalability for image domain translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 783–790.

[34] H. Tang, D. Xu, W. Wang, Y. Yan, and N. Sebe, "Dual generator generative adversarial networks for multi-domain image-to-image translation," in *Proc. ACCV*, 2018, pp. 3–21.

[35] Y. Wang, J. van de Weijer, and L. Herranz, "Mix and match networks: Encoder-decoder alignment for zero-pair image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5467–5476.

[36] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3917–3925.

[37] S. Mo, M. Cho, and J. Shin, "InstaGAN: Instance-aware image-to-image translation," in *Proc. ICLR*, 2019.

[38] J.-Y. Zhu *et al.*, "Toward multimodal image-to-image translation," in *Proc. NeurIPS*, 2017, pp. 465–476.

[39] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

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

[41] A. M. Martinez, "The AR face database," Tech. Rep. CVC TR 24, 1998. [Online]. Available: https://www2.ece.ohio-state.edu/~aleix/ARdatabase.html

[42] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[43] M. Li, W. Zuo, and D. Zhang, "Deep identity-aware transfer of facial attributes," 2016, *arXiv:1610.05586*. [Online]. Available: http://arxiv.org/abs/1610.05586

[44] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. NeurIPS*, 2017, pp. 700–708.

[45] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Proc. NeurIPS*, 2016, pp. 469–477.

[46] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2107–2116.

[47] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proc. ECCV*, 2018, pp. 818–833.

[48] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[49] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," in *Proc. ICLR*, 2017.

[50] V. Dumoulin *et al.*, "Adversarially learned inference," in *Proc. ICLR*, 2017.

[51] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[52] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *Proc. ICLR*, 2018.

[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 *Proc. NeurIPS*, 2017, pp. 6629–6640.

Hao Tang received the master's degree from the School of Electronics and Computer Engineering, Peking University, Beijing, China, in 2016, and the Ph.D. degree from the Multimedia and Human Understanding Group, University of Trento, Trento, Italy, in 2021.

He was a Visiting Scholar with the Department of Engineering Science, University of Oxford, Oxford, U.K. He is currently a Postdoctoral with the Computer Vision Laboratory, ETH Zurich, Zurich, Switzerland. His research interests are deep learning, machine learning, and their applications to computer vision.

Hong Liu received the Ph.D. degree in mechanical electronics and automation from the School of Electronics Engineering (EE) and Computer Science (CS), Peking University (PKU), Beijing, China, in 1996.

He serves as a Full Professor with the School of EE&CS, PKU. He has been selected as the Chinese Innovation Leading Talent supported by National High-level Talents Special Support Plan since 2013. He is also the Director of Open Lab on Human Robot Interaction, PKU. His research fields include computer vision and robotics, image processing, and pattern recognition.

Dan Xu received the Ph.D. degree from the Department of Computer Science, University of Trento, Trento, Italy, in 2018.

He is an Assistant Professor with the Department of Computer Science and Engineering, HKUST. He was a Postdoctoral Research Fellow with VGG, University of Oxford, Oxford, U.K. He was also a Research Assistant of the Multimedia (MM) Lab at The Chinese University of Hong Kong, Hong Kong. He served as the Senior Program Committee or the Area Chair of the Association for the Advancement of Artificial Intelligence (AAAI) 2022, the Association for Computing Machinery (ACM) MM 2021, ACM MM 2020, International Conference on Pattern Recognition (ICPR) 2020, and Winter Conference on Applications of Computer Vision (WACV) 2021.

Philip H. S. Torr is a Professor with the University of Oxford, Oxford, U.K. After working for another three years at Oxford, he worked for six years for Microsoft Research, first in Redmond, then in Cambridge, founding the vision side of the Machine Learning and Perception Group. He has won awards from top vision conferences, including International Conference on Computer Vision (ICCV), Conference on Computer Vision and Pattern Recognition (CVPR), European Conference on Computer Vision (ECCV), Neural Information Processing Systems (NIPS), and British Machine Vision Conference (BMVC). He is a Royal Society Wolfson Research Merit Award holder.

Nicu Sebe (Senior Member, IEEE) is a Professor with the University of Trento, Trento, Italy, where he is leading the research in the areas of multimedia analysis and human behavior understanding.

Dr. Sebe was the General Co-Chair of the IEEE FG 2008 and the Association for Computing Machinery (ACM) Multimedia 2013. He was a Program Chair of the ACM Multimedia 2011 and 2007, European Conference on Computer Vision (ECCV) 2016, International Conference on Computer Vision (ICCV) 2017, and International Conference on Pattern Recognition (ICPR) 2020. He is a General Chair of ACM Multimedia 2022 and a Program Chair of ECCV 2024. He is a fellow of the International Association of Pattern Recognition (IAPR).