Attention-Guided Generative Adversarial Networks for Unsupervised Image-to-Image Translation

Hao Tang\(^1\) Dan Xu\(^2\) Nicu Şebè\(^1\) Yan Yan\(^3\)

\(^1\)Department of Information Engineering and Computer Science, University of Trento, Trento, Italy
\(^2\)Department of Engineering Science, University of Oxford, Oxford, United Kingdom
\(^3\)Department of Computer Science, Texas State University, San Marcos, USA

\{hao.tang, niculae.sebe\}@unitn.it, danxu@robots.ox.ac.uk, y_y34@txstate.edu

Abstract—The state-of-the-art approaches in Generative Adversarial Networks (GANs) are able to learn a mapping function from one image domain to another with unpaired image data. However, these methods often produce artifacts and can only be able to convert low-level information, but fail to transfer high-level semantic part of images. The reason is mainly that generators do not have the ability to detect the most discriminative semantic part of images, which thus makes the generated images with low-quality. To handle the limitation, in this paper we propose a novel Attention-Guided Generative Adversarial Network (AGGAN), which can detect the most discriminative semantic object and minimize changes of unwanted part for semantic manipulation problems without using extra data and models. The attention-guided generators in AGGAN are able to produce attention masks via a built-in attention mechanism, and then fuse the input image with the attention mask to obtain a target image with high-quality. Moreover, we propose a novel attention-guided discriminator which only considers attended regions. The proposed AGGAN is trained by an end-to-end fashion with an adversarial loss, cycle-consistency loss, pixel loss and attention loss. Both qualitative and quantitative results demonstrate that our approach is effective to generate sharper and more accurate images than existing models. The code is available at https://github.com/Ha0Tang/AttentionGAN.

Index Terms—GANs, Image-to-Image Translation, Attention

I. INTRODUCTION

Recently, Generative Adversarial Networks (GANs) [8] have received considerable attention across many communities, e.g., computer vision, natural language processing, audio and video processing. GANs are generative models, which are particularly designed for image generation task. Recent works in computer vision, image processing and computer graphics have produced powerful translation systems in supervised settings such as Pix2pix [11], where the image pairs are required. However, the paired training data are usually difficult and expensive to obtain. Especially, the input-output pairs for images tasks such as artistic stylization can be even more difficult to acquire since the desired output is quite complex, typically requiring artistic authoring. To tackle this problem, CycleGAN [47], DualGAN [43] and DiscoGAN [13] provide an insight, in which the models can learn the mapping from one image domain to another one with unpaired image data.

Despite these efforts, image-to-image translation, e.g., converting a neutral expression to a happy expression, remains a challenging problem due to the fact that the facial expression changes are non-linear, unaligned and vary conditioned on the appearance of the face. Moreover, most previous models change unwanted objects during the translation stage and can also be easily affected by background changes. In order to address these limitations, Liang et al. propose the ContrastGAN [18], which uses object mask annotations from each dataset. In ContratGAN, it first crops a part in the image according to the masks, and then makes translations and finally pastes it back. Promising results have been obtained from it, however it is hard to collect training data with object masks. More importantly, we have to make an assumption that the object shape should not change after applying semantic modification. Another option is to train an extra model to detect the object masks and fit them into the generated image patches [6], [12]. In this case, we need to increase the number of parameters of our network, which consequently increases the training complexity both in time and space.

To overcome the aforementioned issues, in this paper we propose a novel Attention-Guided Generative Adversarial Network (AGGAN) for the image translation problem without using extra data and models. The proposed AGGAN comprises of two generators and two discriminators, which is similar with CycleGAN [47]. Fig. 1 illustrates the differences between previous representative works and the proposed AGGAN. Two attention-guided generators in the proposed AGGAN have built-in attention modules, which can disentangle the discriminative semantic object and the unwanted part by producing a attention mask and a content mask. Then we fuse the input image with new patches produced through the attention mask to obtain high-quality results. We also constrain generators with pixel-wise and cycle-consistency loss function, which forces the generators to reduce changes. Moreover, we propose two novel attention-guided discriminators which aims to consider only the attended regions. The proposed AGGAN is trained by an end-to-end fashion, and can produce attention mask, content mask and targeted images at the same time. Experimental results on four public available datasets demonstrate that the proposed AGGAN is able to produce higher-quality images compared with the state-of-the-art methods.

The contributions of this paper are summarized as follows:

- We propose a novel Attention-Guided Generative Adversarial Network (AGGAN) for unsupervised image-to-image translation.
- We propose a novel generator architecture with built-in at-
tation mechanism, which can detect the most discriminative semantic part of images in different domains.

- We propose a novel attention-guided discriminator which only consider the attended regions. Moreover, the proposed attention-guided generator and discriminator can be easily used to other GAN models.
- Extensive results demonstrate that the proposed AGGAN can generate sharper faces with clearer details and more realistic expressions compared with baseline models.

II. RELATED WORK

Generative Adversarial Networks (GANs) [8] are powerful generative models, which have achieved impressive results on different computer vision tasks, e.g., image generation [5], [27], [9], image editing [34], [35] and image inpainting [17], [10]. In order to generate meaningful images that meet user requirement, Conditional GAN (CGAN) [26] is proposed where the conditioned information is employed to guide the image generation process. The conditioned information can be discrete labels [28], text [22], [31], object keypoints [32], human skeleton [39] and reference images [11]. CGANs using a reference images as conditional images as conditional information have tackled a lot of problems, e.g., text-to-image translation [22], image-to-image translation [11] and video-to-video translation [42].

Image-to-Image Translation models learns a translation function using CNNs. Pix2pix [11] is a conditional framework using a CGAN to learn a mapping function from input to output images. Similar ideas have also been applied to many other tasks, such as generating photographs from sketches [33] or vice versa [38]. However, most of the tasks in the real world suffer from the constraint of having few or none of the paired input-output samples available. To overcome this limitation, unpaired image-to-image translation task has been proposed. Different from the prior works, unpaired image translation task try to learn the mapping function without the requirement of paired training data. Specifically, CycleGAN [47] learns the mappings between two image domains (i.e., a source domain X to a target domain Y) instead of the paired images. Apart from CycleGAN, many other GAN variants are proposed to tackle the cross-domain problem. For example, to learn a common representation across domains, CoupledGAN [19] uses a weight-sharing strategy. The work of [37] utilizes some certain shared content features between input and output even though they may differ in style. Kim et al. [13] propose a method based on GANs that learns to discover relations between different domains. A model which can learn object transfiguration from two unpaired sets of images is presented in [46]. Tang et al. [41] propose $G^2$GAN, which is a robust and scalable approach allowing to perform unpaired image-to-image translation for multiple domains. However, those models can be easily affected by unwanted content and cannot focus on the most discriminative semantic part of images during translation stage.

Attention-Guided Image-to-Image Translation. In order to fix the aforementioned limitations, Liang et al. propose ContrastGAN [18], which uses the object mask annotations from each dataset as extra input data. In this method, we have to make an assumption that after applying semantic changes an object shape does not change. Another method is to train another segmentation or attention model and fit it to the system. For instance, Mejjati et al. [25] propose an attention mechanisms that are jointly trained with the generators and discriminators. Chen et al. propose AttentionGAN [6], which uses an extra attention network to generate attention maps, so that major attention can be paid to objects of interests. Kastaniotis et al. [12] present ATAGAN, which use a teacher network to produce attention maps. Zhang et al. [45] propose the Self-Attention Generative Adversarial Networks (SAGAN) for image generation task. Qian et al. [30] employ a recurrent network to generate visual attention first and then transform a raindrop degraded image into a clean one. Tang et al. [40] propose a novel Multi-Channel Attention Selection GAN for the challenging cross-view image translation task. Sun et al. [36] generate a facial mask by using FCN [21] for face attribute manipulation.

All these aforementioned 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. In this work, we propose the Attention-Guided Generative Adversarial Network (AGGAN), which can produce attention masks by the generators. For this purpose, we embed an attention method to the vanilla generator which means that we do not need any extra models to obtain the attention masks of objects of interests.
Fig. 2: The framework of the proposed AGGAN. Because of the space limitation, we only show one mapping in this figure, i.e., $x \rightarrow [M_y, R_y, G_y] \rightarrow C_x \approx x$. We also have the other mapping, i.e., $y \rightarrow [M_x, R_x, G_x] \rightarrow C_y \approx y$. The attention-guided generators have built-in attention mechanism, which can detect the most discriminative part of images. After that we mix the input image, content mask and the attention mask to synthesize the targeted image. Moreover, to distinguish only the most discriminative content, we also propose a attention-guided discriminator $D_{YA}$. Note that our systems do not require supervision, i.e., no pairs of images of the same person with different expressions.

III. METHOD

We start with the attention-guided generator and discriminator of the proposed Attention-Guided Generative Adversarial Network (AGGAN), and then introduce the loss function for better optimization of the model. Finally we present the implementation details of the whole model including network architecture and training procedure.

A. Attention-Guided Generator

GANs [8] are composed of two competing modules, i.e., the generator $G_{X \rightarrow Y}$ and the discriminator $D_Y$ (where $X$ and $Y$ denote two different image domains), which are iteratively trained competing against each other in the manner of two-player minimax. More formally, let $x_i \in X$ and $y_j \in Y$ denote the training images in source and target domain, respectively (for simplicity, we usually omit the subscript $i$ and $j$). For most current image translation models, e.g., CycleGAN [47] and DualGAN [43], they include two mappings $G_{X \rightarrow Y}: x \rightarrow G_y$ and $G_{Y \rightarrow X}: y \rightarrow G_x$, and two corresponding adversarial discriminators $D_X$ and $D_Y$. The generator $G_{X \rightarrow Y}$ maps $x$ from the source domain to the generated image $G_y$ in the target domain $Y$ and tries to fool the discriminator $D_Y$, whilst the $D_Y$ focuses on improving itself in order to be able to tell whether a sample is a generated sample or a real data sample. Similar to $G_{Y \rightarrow X}$ and $D_X$.

While for the proposed AGGAN, we intend to learn two mappings between domains $X$ and $Y$ via two generators with built-in attention mechanism, i.e., $G_{X \rightarrow Y}: x \rightarrow [M_y, R_y, G_y]$ and $G_{Y \rightarrow X}: y \rightarrow [M_x, R_x, G_x]$, where $M_x$ and $M_y$ are the attention masks of images $x$ and $y$, respectively; $R_x$ and $R_y$ are the content masks of images $x$ and $y$, respectively; $G_x$ and $G_y$ are the generated images. The attention masks $M_x$ and $M_y$ define a per pixel intensity specifying to which extend each pixel of the content masks $R_x$ and $R_y$ will contribute in the final rendered image. In this way, the generator does not need to render static elements, and can focus exclusively on the pixels defining the facial movements, leading to sharper and more realistic synthetic images. After that, we fuse input image $x$ and the generated attention mask $M_y$, and the content mask $R_y$ to obtain the targeted image $G_y$. Through this way, we can disentangle the most discriminative semantic object and unwanted part of images. In Fig. 2, the attention-guided generators focus only on those regions of the image that are responsible of generating the novel expression such as eyes and mouth, and keep the rest parts of the image such as hair, glasses, clothes untouched. The higher intensity in the attention mask means the larger contribution for changing the expression.

We focus on the discriminative semantic parts in two different domains, we specifically designed two generators with built-in attention mechanism. By using this mechanism, generators can generate attention masks in two different domains. The input of each generator is a three-channel image, and the outputs of each generator are a attention mask and a content mask. Specifically, the input image of $G_{X \rightarrow Y}$ is $x \in \mathbb{R}^{H \times W \times 3}$, and the outputs are the attention mask $M_y \in \{0, \ldots, 1\}^{H \times W}$ and content mask $R_y \in \mathbb{R}^{H \times W \times 3}$. Thus, we use the following formulation to calculate the final image $G_y$.

$$G_y = R_y \ast M_y + x \ast (1 - M_y),$$

where attention mask $M_y$ is copied into three-channel for multiplication purpose. The formulation for generator $G_{Y \rightarrow X}$ and input image $y$ is $G_x = R_x \ast M_x + y \ast (1 - M_x)$. Intuitively, attention mask $M_y$ enables some specific areas where facial muscle changed to get more focus, applying it to the content mask $R_y$ can generate images with clear dynamic area and unclear static area. After that, what left is to enhance the static area, which should be similar between the generated image and the original real image. Therefore we can enhance the static area (basically it refers to background area) in the original
real image \((1-M_y)\times x\) and merge it to \(R_y\times M_y\) to obtain final result \(R_y\times M_y + x\times(1-M_y)\).

### B. Attention-Guided Discriminator

Eq. (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 \(G_y\) or the real image \(y\) as input and tries to distinguish them. Similar to discriminator \(D_X\), which tries to distinguish the generated image \(G_x\) and the real image \(x\). To add attention mechanism to the discriminator so that it only considers attended regions. We propose two attention-guided discriminators. The attention-guided discriminator is structurally the same with the vanilla discriminator but it also takes the attention mask as input. For attention-guided discriminator \(D_Y^A\), which tries to distinguish the fake image pairs \([M_y, G_y]\) and the real image pairs \([M_y, y]\). Similar to \(D_X^A\), which tries to distinguish the fake image pairs \([M_x, G_x]\) and the real image pairs \([M_x, x]\). In this way, discriminators can focus on the most discriminative content.

### C. Optimization Objective

#### Vanilla Adversarial Loss \(L_{GAN}(G_{X\rightarrow Y}, D_Y)\) [8] can be formulated as follows:

\[
L_{GAN}(G_{X\rightarrow Y}, D_Y) = \mathbb{E}_{y\sim \text{p\text{-data}(y)}} [\log D_Y(y)] + \mathbb{E}_{x\sim \text{p\text{-data}(x)}} [\log(1 - D_Y(G_{X\rightarrow Y}(x)))].
\]  

\(G_{X\rightarrow Y}\) tries to minimize the adversarial loss objective \(L_{GAN}(G_{X\rightarrow Y}, D_Y)\) while \(D_Y\) tries to maximize it. The target of \(G_{X\rightarrow Y}\) is to generate an image \(G_y=G_{X\rightarrow Y}(x)\) that looks similar to the images from domain \(Y\), while \(D_Y\) aims to distinguish between the generated images \(G_{X\rightarrow Y}(x)\) and the real images \(y\). A similar adversarial loss of Eq. (2) for mapping function \(G_{Y\rightarrow X}\) and its discriminator \(D_X\) is defined as \(L_{GAN}(G_{Y\rightarrow X}, D_X) = \mathbb{E}_{x\sim \text{p\text{-data}(x)}} [\log D_X(x)] + \mathbb{E}_{y\sim \text{p\text{-data}(y)}} [\log(1 - D_X(G_{Y\rightarrow X}(y)))]\).

#### Attention-Guided Adversarial Loss

We propose the attention-guided adversarial loss for training the attention-guided discriminators. The min-max game between the attention-guided discriminator \(D_Y^A\) and the generator \(G_{X\rightarrow Y}\) is performed through the following objective functions:

\[
L_{AGAN}(G_{X\rightarrow Y}, D_{YA}) = \mathbb{E}_{y\sim \text{p\text{-data}(y)}} [\log D_{YA}([M_y, y])] + \mathbb{E}_{x\sim \text{p\text{-data}(x)}} [\log(1 - D_{YA}([M_y, G_{X\rightarrow Y}(x)]))],
\]

where \(D_{YA}\) aims to distinguish between the generated image pairs \([M_y, G_{X\rightarrow Y}(x)]\) and the real image pairs \([M_y, y]\). We also have another loss \(L_{AGAN}(G_{Y\rightarrow X}, D_A)\) for discriminator \(D_X^A\) and the generator \(G_{Y\rightarrow X}\).

#### Cycle-Consistency Loss

Note that CycleGAN [47] and DualGAN [43] are different from Pix2pix [11] as the training data in those models are unpaired. The cycle-consistency loss can be used to enforce forward and backward consistency. The cycle-consistency loss can be regarded as “pseudo” pairs of training data even though we do not have the corresponding data in the target domain which corresponds to the input data from the source domain. Thus, the loss function of cycle-consistency can be defined as:

\[
L_{cycle}(G_{X\rightarrow Y}, G_{Y\rightarrow X}) = \mathbb{E}_{x\sim \text{p\text{-data}(x)}} [||G_{Y\rightarrow X}(G_{X\rightarrow Y}(x)) - x||_1] + \mathbb{E}_{y\sim \text{p\text{-data}(y)}} [||G_{Y\rightarrow X}(G_{Y\rightarrow X}(y)) - y||_1].
\]

The reconstructed images \(C_x=G_{Y\rightarrow X}(G_{X\rightarrow Y}(x))\) are closely matched to the input image \(x\), and similar to \(G_{X\rightarrow Y}(G_{Y\rightarrow X}(y))\) and image \(y\).

#### Pixel Loss

To reduce changes and restrain generators, we adopt pixel loss between the input images and the generated images. We express this loss as:

\[
L_{\text{pixel}}(G_{X\rightarrow Y}, G_{Y\rightarrow X}) = \mathbb{E}_{x\sim \text{p\text{-data}(x)}} [||G_{X\rightarrow Y}(x) - x||_1] + \mathbb{E}_{y\sim \text{p\text{-data}(y)}} [||G_{Y\rightarrow X}(y) - y||_1].
\]

We adopt \(L1\) distance as loss measurement in pixel loss. Note that the pixel loss usually used in the paired image-to-image translation models such as Pix2pix [11]. While we use it in our AGGAN for unpaired image-to-image translation task.

#### Attention Loss

When training our AGGAN we do not have ground-truth annotation for the attention masks. They are learned from the resulting gradients of the attention-guided discriminators and the rest of the losses. However, the attention masks can easily saturate to 1 which makes the attention-guided generator has no effect as indicated in GANimation [29]. To prevent this situation, we perform a Total Variation Regularization over attention masks \(M_y\) and \(M_x\).

The attention loss of mask \(M_x\) therefore can be defined as:

\[
L_{tv}(M_x) = \sum_{w,h} |M_x(w + 1, h, c) - M_x(w, h, c)| + |M_x(w, h + 1, c) - M_x(w, h, c)|,
\]

where \(W\) and \(H\) are the width and height of \(M_x\).

#### Full Objective

Thus, the complete objective loss of AGGAN can be formulated as follows:

\[
L(G_{X\rightarrow Y}, G_{Y\rightarrow X}, D_X, D_Y, D_{XA}, D_{YA}) = \lambda_{gan} \times [L_{GAN}(G_{X\rightarrow Y}, D_Y) + L_{GAN}(G_{Y\rightarrow X}, D_X)] + \lambda_{AGAN} [L_{AGAN}(G_{X\rightarrow Y}, D_A) + L_{AGAN}(G_{Y\rightarrow X}, D_A)] + \lambda_{cycle} \times L_{cycle}(G_{X\rightarrow Y}, G_{Y\rightarrow X}) + \lambda_{pix} \times L_{\text{pixel}}(G_{X\rightarrow Y}, G_{Y\rightarrow X}) + \lambda_{tv} [L_{tv}(M_x) + L_{tv}(M_y)],
\]

where \(\lambda_{gan}, \lambda_{cycle}, \lambda_{pix}\) and \(\lambda_{tv}\) are parameters controlling the relative relation of objectives terms. We aim to solve:

\[
G_{X\rightarrow Y}^*, G_{Y\rightarrow X}^* = \arg \min_{G_{X\rightarrow Y}, D_{YA}, D_Y, D_{XA}, D_{YA}} L(G_{X\rightarrow Y}, G_{Y\rightarrow X}, D_X, D_Y, D_{XA}, D_{YA}).
\]

### D. Implementation Details

#### Network Architecture

For fair comparison, we use the generator architecture from CycleGAN [47]. We have slightly modified it for our task and the network architecture of the proposed generators is,

\[c7s1_64, d128, d256, R256, R256, R256, R256, R256, R256, R256, R256,\]
proposed AGGAN. These datasets contains the faces with different races and they have different illumination, occlusion, pose conditions and backgrounds. (i) Large-scale Celeb Faces Attributes (CelebA) dataset [20] has more than 200K celebrity images with complex backgrounds. To evaluate the performance of the proposed method under the situation where training data is limited. We randomly select 1,000 neutral images and 1,000 smile images as training data, and another 1,000 neutral and 1,000 smile images as testing data. (ii) RaFD dataset [15] consists of 4,824 images collected from 67 participants. Each participant have eight facial expressions in three gaze directions, which are captured from three different angles. We employed all of the images for diversity expression generation task. (iii) AR Face [24] contains over 4,000 color images in which only 1,018 images have 4 different facial expressions, i.e., smile, anger, fear and neutral expression. We employ the images with the expression labels of smile and neutral to evaluate our method. (iv) Bu3df [44] is a 3D facial expression dataset including 100 subjects with 7 different expression categories. We employed the images with smile and neutral expressions as training and testing data.

Parameter Setting. For all datasets, images are rescaled to $256 \times 256 \times 3$ and we do left-right flip for data augmentation. For optimization, all baselines and the proposed AGGAN are trained with batch size of 1. All models were trained for 200 epochs on AR Face, Bu2df and RaFD datasets, 80 epochs on CelebA dataset. For all the experiments, we set $\lambda_{cycle}=10, \lambda_{gan}=0.5, \lambda_{pixel}=1$ and $\lambda_{tv}=1e^{-6}$. We set the number of image buffer to 50 similar in [47]. We use the Adam optimizer [14] with the momentum terms $\beta_1=0.5$ and $\beta_2=0.999$. The initial learning rate for Adam optimizer is 0.0001. After 100 and 40 epochs for different datasets, the learning rate starts linearly decaying to 0. For $r$ in Eq. (9), we set it to 0.01 at the first 10 epochs. After curriculum period, we set $r$ to 0.5. The proposed AGGAN is implemented using public PyTorch framework. To speed up both training and testing processes, we use a Nvidia TITAN Xp GPU.

Competing Models. We consider several state-of-the-art cross-domain image generation models as our baselines. (i) Unpaired image translation method: CycleGAN [47], DIAT [16], DiscoGAN [13], DistanceGAN [4], Dist. + Cycle [4], Self Dist. [4], ComboGAN [1], StarGAN [7]; (ii) Paired image translation method: BicycleGAN [48], Pix2pix [11], Encoder-Decoder [11] and (iii) Label-, mask- or attention-guided image translation method: IcGAN [28], ContrastGAN [18] and GANimation [29]. Note that the fully supervised Pix2pix, Encoder-Decoder (Enc.-Decoder) and BicycleGAN are trained with paired data on AR Face and Bu3df datasets. Since BicycleGAN can generate several different outputs with one single input image, we randomly select one output from them for fair comparison. To re-implement ContrastGAN, we use OpenFace [3] to obtain the face masks as extra input data. For a fair comparison, we implement all the baselines using the same setups as our approach.

Evaluation Metrics. We adopt Amazon Mechanical Turk (AMT) perceptual studies to evaluate the generated images. We gather data from 50 participants per algorithm we tested.
Participants were shown a sequence of pairs of images, one real image and one fake (generated by our algorithm or a baseline), and asked to click on the image they thought was real. To seek a quantitative measure that does not require human participation, Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) are employed.

B. Comparison with the State-of-the-Art

Bu3dfe&AR Face Dataset. Results of Bu3dfe and AR Face datasets are shown in Fig. 3. It is clear that the results of Dist.+Cycle and Self Dist. cannot even generate human faces. DiscoGAN produce identical results regardless of the input faces, which suffers from mode collapse. While the results of DualGAN, DistanceGAN, StarGAN, Pix2pix, Encoder-Decoder and BicycleGAN tend to be blurry. While ComboGAN and ContrastGAN can produce the same identity but without expression changing. CycleGAN can generate sharper images, but the details of the generated faces are not convincing. Compared with all the baselines, the results of our AGGAN are more smooth, correct and with more details.

CelebA Dataset. Since CelebA dataset do not provide paired data, thus we cannot conduct experiments on supervised methods. The results of CelebA dataset are shown in Fig. 4. We can see that only the proposed AGGAN can produce photo-realistic faces with correct expressions. Thus, we can conclude that even though the subjects in the three datasets have different races, poses, skin colors, illumination conditions, occlusions and complex backgrounds, our method consistently generates more sharper images with correct expressions than the baseline models. Moreover, we observe that our AGGAN preforms better than other baselines when training data is limited, which also shows that our method is very robust.

RaFD Dataset. Our model can be easily extended to generate facial diversity expressions (e.g., sad, happy and fearful). To generate diversity expressions in one single model we employ the domain classification loss proposed in StarGAN.
TABLE I: Quantitative comparison with different models. For all metrics except MSE, high is better.

<table>
<thead>
<tr>
<th>Model</th>
<th>AR Face</th>
<th>Bu3dfe</th>
<th>CelebA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AMT</td>
<td>PSNR</td>
<td>MSE</td>
</tr>
<tr>
<td>CycleGAN</td>
<td>10.2</td>
<td>14.8142</td>
<td>2538.4</td>
</tr>
<tr>
<td>DialGAN</td>
<td>1.3</td>
<td>14.7458</td>
<td>2545.7</td>
</tr>
<tr>
<td>DiscoGAN</td>
<td>0.1</td>
<td>13.1547</td>
<td>3321.9</td>
</tr>
<tr>
<td>ComBiGAN</td>
<td>1.5</td>
<td>14.7405</td>
<td>2550.6</td>
</tr>
<tr>
<td>DistanceGAN</td>
<td>0.3</td>
<td>13.9055</td>
<td>3018.5</td>
</tr>
<tr>
<td>Dist. + Cycle</td>
<td>0.1</td>
<td>3.8032</td>
<td>27516.2</td>
</tr>
<tr>
<td>Self Dist.</td>
<td>0.1</td>
<td>3.8674</td>
<td>26775.4</td>
</tr>
<tr>
<td>StarGAN</td>
<td>1.6</td>
<td>13.7551</td>
<td>3560.2</td>
</tr>
<tr>
<td>ContrastGAN</td>
<td>8.3</td>
<td>14.8395</td>
<td>2511.1</td>
</tr>
<tr>
<td>Pix2pix</td>
<td>2.6</td>
<td>14.6118</td>
<td>2601.3</td>
</tr>
<tr>
<td>Enc.-Decoder</td>
<td>0.1</td>
<td>12.6660</td>
<td>3755.4</td>
</tr>
<tr>
<td>BicycleGAN</td>
<td>1.5</td>
<td>13.9114</td>
<td>2541.8</td>
</tr>
<tr>
<td>AGGAN</td>
<td>12.8</td>
<td>14.9187</td>
<td>2508.6</td>
</tr>
</tbody>
</table>

TABLE II: Ablation study of AGGAN.

<table>
<thead>
<tr>
<th>Component</th>
<th>AMT</th>
<th>PSNR</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>12.8</td>
<td>14.9187</td>
<td>2508.6</td>
</tr>
<tr>
<td>Full - AD</td>
<td>10.2</td>
<td>14.6352</td>
<td>2569.2</td>
</tr>
<tr>
<td>Full - AD - AG</td>
<td>3.2</td>
<td>14.4646</td>
<td>2636.7</td>
</tr>
<tr>
<td>Full - AD - PL</td>
<td>8.9</td>
<td>14.5128</td>
<td>2619.8</td>
</tr>
<tr>
<td>Full - AD - AL</td>
<td>6.3</td>
<td>14.6129</td>
<td>2652.6</td>
</tr>
<tr>
<td>Full - AD - PL - AL</td>
<td>5.2</td>
<td>14.3287</td>
<td>2787.3</td>
</tr>
</tbody>
</table>

The results compared against the baselines DIAT, CycleGAN, IcGAN, StarGAN and GANimation are shown in Fig. 5. For GANimation, we follow the authors’ instruction and use OpenFace [3] to obtain the action units of each face as extra input data. We observe that the proposed AGGAN achieves competitive results compared to GANimation.

Quantitative Comparison on All Datasets. We also provide quantitative results on AR Face, Bu3dfe and CelebA datasets. As shown in Table I, we can see that the proposed AGGAN achieves best results on these datasets compared with competing models including fully-supervised methods Pix2pix, Encoder-Decoder, BicycleGAN, and mask-, label- conditional method, i.e., ContrastGAN.

C. Model Analysis

Analysis of Model Component. In Table II and Fig. 6 we run ablation studies of our model on the AR Face dataset. We gradually remove components of the proposed AGGAN, i.e., Attention-guided Discriminator (AD), Attention-guided Generator (AG), Attention Loss (AL), Pixel Loss (PL). We find that removing one of them substantially degrades results, which means all of them are critical to our results. Note that without AG we cannot generate the attention mask and content mask, as shown in Fig. 6.

Attention/Content Mask Visualization. Instead of regressing a full image, our generator outputs two masks, a content mask and an attention mask. We also visualize the generation of both masks on RaFD dataset in Fig. 7. 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 of synthesizing the novel expression. The attention mask mainly focuses on the eyes and mouth, which means these parts are important for generating the novel expression.

Comparison of the Number of Parameters. The number of models for different \( m \) image domains and the number of model parameters on RaFD is shown in Table III. Note that our performance is much better than these baselines and the number of parameters is comparable with ContrastGAN, while this model requires the object mask as extra data.
### TABLE III: Comparison of the overall model capacity on RaFD Dataset (m=8).

<table>
<thead>
<tr>
<th>Method</th>
<th># of Models</th>
<th># of Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Px2Pxp</td>
<td>m(m-1)</td>
<td>57.2M × 56</td>
</tr>
<tr>
<td>Encoder-Decoder</td>
<td>m(m-1)</td>
<td>41.9M × 56</td>
</tr>
<tr>
<td>BicycleGAN</td>
<td>m(m-1)</td>
<td>64.3M × 56</td>
</tr>
<tr>
<td>CycleGAN</td>
<td>m(m-1)/2</td>
<td>52.6M × 28</td>
</tr>
<tr>
<td>DualGAN</td>
<td>m(m-1)/2</td>
<td>178.7M × 28</td>
</tr>
<tr>
<td>DiscoGAN</td>
<td>m(m-1)/2</td>
<td>16.6M × 28</td>
</tr>
<tr>
<td>DistanceGAN</td>
<td>m(m-1)/2</td>
<td>52.6M × 28</td>
</tr>
<tr>
<td>Dist. + Cycle</td>
<td>m(m-1)/2</td>
<td>52.6M × 28</td>
</tr>
<tr>
<td>Self Dist.</td>
<td>m(m-1)/2</td>
<td>52.6M × 28</td>
</tr>
<tr>
<td>CombGAN</td>
<td>m</td>
<td>14.4M × 8</td>
</tr>
<tr>
<td>StarGAN</td>
<td>1</td>
<td>33.2M × 1</td>
</tr>
<tr>
<td>Conv2GAN</td>
<td>1</td>
<td>52.6M × 1</td>
</tr>
<tr>
<td>AGGAN(Ours)</td>
<td>1</td>
<td>52.6M × 1</td>
</tr>
</tbody>
</table>

### V. CONCLUSION

We propose a novel AGGAN for unsupervised image-to-image translation. The generators in AGGAN have the built-in attention mechanism, which can detect the most discriminative part of images and produce the attention mask. Then the attention mask and the input image are combined to generate the targeted images with high-quality. We also propose a novel attention-guided discriminator, which only focus on the attended content. The proposed AGGAN is trained by an end-to-end fashion. Experimental results on four datasets demonstrate that AGGAN and the training strategy can generate compelling results with more convincing details and correct expressions than the state-of-the-art models.

**Acknowledgements:** We acknowledge the National Institute of Standards and Technology Grant 60NANB17D191 for funding this research. We also acknowledge the gift donation from Cisco, Inc for this research.

## References