Attribute Consistency Guided Generative Adversarial Networks for Unsupervised Image-to-Image Translation

Fengjiang Liu
School of Computer Science and Engineering, Southeast University
Nanjing, 211189, P.R. China

Li Yao(✉)
School of Computer Science and Engineering, Southeast University
Key Laboratory of Computer Network and Information Integration
Nanjing, 211189, P.R. China
{fengjiang_liu, yao.li}@seu.edu.cn

Abstract

We propose a novel method for unsupervised image-to-image translation, which includes a generator with attribute consistency constraints and a new multi-scale discriminator. Unlike the existing methods, which cannot preserve important attributes of input images while handling large shape deformations, our method can simultaneously take both tasks. We introduce the attribute-based recalibration module into the generator to extract the input image's attribute features efficiently. On this basis, we impose attribute consistency constraints on feature space to ensure the preservation of the crucial attributes of the input images. Moreover, our multi-scale discriminator shares a backbone and only introduces the attention mechanism on the highest scale, facilitating shape deformation and improving image generation quality. We show that the proposed method outperforms the existing state-of-the-art models in various challenging applications, including selfie-to-anime, dog-to-cat, and cat-to-dog. Our code is available at `https://github.com/Nightfury12366/ACG-GAN`.

1. Introduction

Unsupervised image-to-image translation aims to learn the mapping function of images in two different domains. It has become an area that attracts a lot of attention from researchers in computer vision. Driven by the generative adversarial networks (GANs) [11], recent works [1, 14, 19, 24, 27, 30] can change the local texture and style of the image by using unpaired training data. Despite the promising results they attain, all these methods are still challenging to complete the tasks that require large shape deformation. In view of this problem, several works [7, 13, 18, 22, 28] can be successfully achieved in image translation tasks with large deformation (*e.g.*, selfie2anime and cat2dog).

Despite these advances, U-GAT-IT [13] focuses more on the regions with significant geometric differences between the two image domains while neglecting the common features between the two image domains. Fig.1(c) shows that U-GAT-IT is more likely to generate images with strange texture shapes and poor quality (*e.g.*, the cat in the generated image has no ears). StarGAN-v2 [7] introduces the style encoding guide generator to complete the image translation task between multiple domains, but it needs a reference image or style code as an additional input. CouncilGAN [22] and ACL-GAN [28] have introduced new mechanisms to avoid incomplete translation, but the generated images by these methods may have attribute differences with the input images, and some essential attributes of the input image were not preserved successfully. Fig.1(d)(e) show the inconsistency between the input images and the generated images; there is a significant difference in texture and hair color between the objects in the input image and output image. *e.g.*, in cat2dog task, the input image is a white cat, model may generate a yellow dog as output, but we expect a white dog image as output.

In the unsupervised image-to-image translation task, the common methods [10, 13, 18, 28, 30] use two generators to transform between two domains. One generator is used to complete the translation from source domain to target domain, and the other is the opposite. Even if there are geometric differences between the source and target domains, there will be common features between the two domains

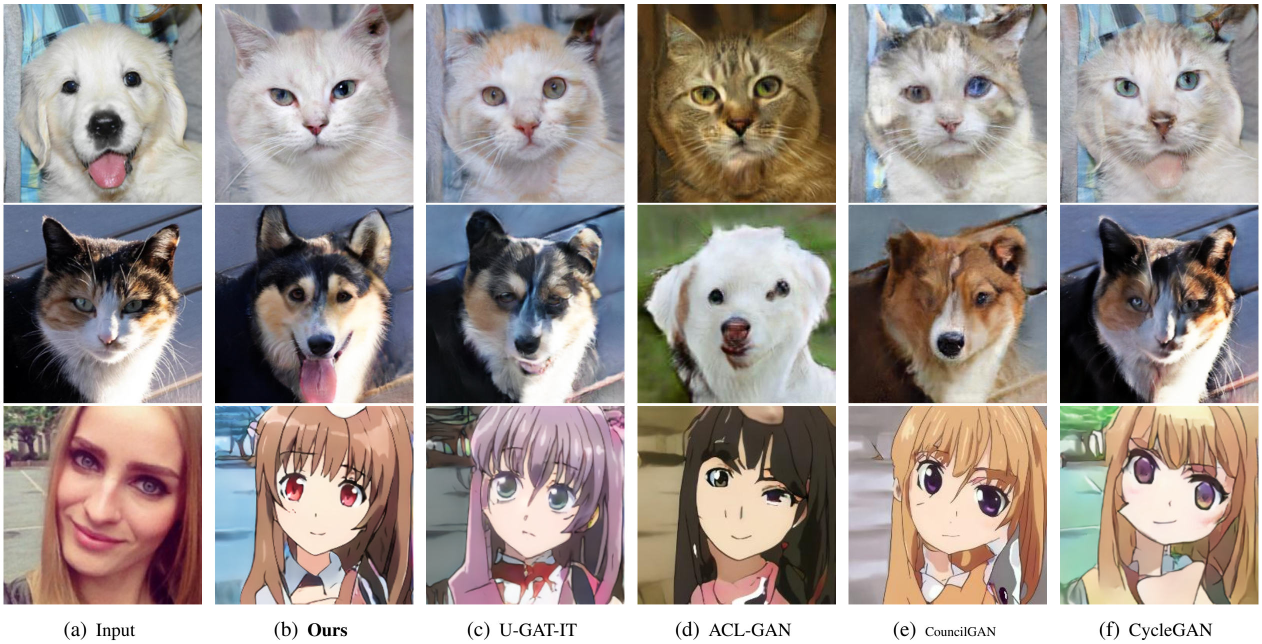


Figure 1. **Example results of our method and baselines.** Our method can better preserve essential attributes of the input images by introducing the attribute-based recalibration module and attribute consistency constraints. Our multi-scale discriminator can guide the generator to deal with deformation better and generate higher-quality images.

(*e.g*., in the cat2dog task, cats and dogs have similar hair textures). Using two independent generators is not conducive to the model to learn the common features between the source and target domains, reduce the model's stability, and generated image quality. Our goal is to propose an improved scheme for unsupervised image-to-image translation tasks involving geometric changes between source and target domains, dealing with the geometric deformation from the source domain to the target domain, and generating higher quality images. At the same time, the generated image should retain the crucial attributes of the input image.

In this work, we propose a novel unsupervised image-to-image translation method, which includes a shared backbone generator with the attribute recalibration module and a new multi-scale discriminator. We only use one generator instead of two to improve the stability of the model and the quality of output images. Besides, we introduce the attribute-based recalibration module into the generator to help the generator better extract the image attribute features, and we impose attribute consistency constraints on the feature space. As a result, the generated image can retain the essential attributes of the input image. In addition, we propose an improved multi-scale discriminator, each scale shares a backbone, and we introduce the attention mechanism on the discriminator of the highest scale. This improvement allows the discriminator to simultaneously focus on the texture details of the image and the shape difference between the two domains. The low-scale discriminators guide the generator to generate clear and high-quality images, while the highest-scale discriminator distinguishes the largest difference regions between two domain images, promoting the generator deal with deformation. Our main contributions can be summarized as follows:

- We introduce the attribute-based recalibration module into the generator to better extract the attribute features of the input images.
- We impose attribute consistency constraints on attribute features to ensure the attribute consistency between input images and generated images.
- We facilitate the shape deformation and improve generated images quality by improving the structure of multi-scale discriminator.

2. Related Work

Generative adversarial networks (GANs). Since the introduction of the GAN framework [11], it has been demonstrated to achieve impressive results in image generation. In this framework, a generator aims to fool a discriminator by generating realistic images, whereas the latter attempts to distinguish between the generated images and real images. Conditional GAN-based standard framework Pix2Pix [12] promotes the study on image-to-image translation. Image-to-image translation aims to learn a mapping from a source domain to a target domain. Several works have proposed super-resolution [26] and video generation [25] frameworks by extending Pix2Pix. However, all these approaches need paired data for training, which limits their practical usage.

Unsupervised image-to-image translation. Unsupervised image-to-image translation aims to learn a mapping from a source domain to a target domain with unpaired training data. CycleGAN [30], DiscoGAN [14], DualGAN [27] stabilize GANs for unsupervised image translation by using a cycle-consistency loss. Some recent works consider the problem of generating multiple output images for a given source image: MUNIT [10] and DRIT [18] decompose the latent space of images into a domain-invariant content space and a domain-specific style space to get diverse outputs. Other works aim to achieve simultaneous translation between multiple(more than two) domains, such as StarGAN [6] and StarGAN-v2 [7]. A more functional line of research focuses on the transformation between domains with a significant geometric difference. U-GAT-IT [13] resort to CAM modules [29] for feature selection and use the Adaptive Layer-Instance Normalization (AdaLIN) function to control the amount of change. Council-GAN [22] considers distribution matching among multiple generators and no longer uses cycle-consistency loss to avoid incomplete conversion. ACL-GAN [28] considers distribution matching between the cycle output and the identity mapping output. Lu et al. [20] improved the performance of CycleGAN [30] by adding additional consistency constraints. MaskGAN [16] implements high-fidelity face semantic editing by introducing mask supervision. DeepFaceEditing [4] decouples the geometric features and appearance features of the face to edit the detailed features of the face.

Multi-scale discriminator. The discriminator in the original GAN framework [11] is simply a binary classifier. For image-to-image translation tasks, Pix2Pix [12] propose the PatchGAN discriminator to classify if each image patch is real or fake. The final output of PatchGAN discriminator is the average results of all the patches. PatchGAN is initially designed for improving the high-frequency part of the generated image; this structure has been extended to multiple scales [26] to cover the low-frequency part as well and has been widely adopted by the latest unsupervised image translation networks [3, 8, 10, 13, 22, 28]. Among these methods, U-GAT-IT uses two scales discriminator, and each scale incorporates a Class Activation Map (CAM) [29] based attention module. MUNIT [10] and ACL-GAN [28] collect the outputs from three scales. The multi-scale discriminators of the above methods have an independent network for each scale, NICE-GAN [3] lets each scale share a backbone.

3. Method

Our goal is to translate images from one domain to another only using unpaired samples drawn from each domain. Let X_s and X_t be the source and target domains, $x_s \in X_s$ and $x_t \in X_t$ be the images of different domains, X be the union set of X_s and X_t ($i.e., X = X_s \cup X_t$), $x \in X$ be a single image. z_s and z_t are latent vector spaces corresponding to the source and target domains. Our framework consists of one generator G and two discriminators D_s and D_t. G and D play an adversarial game, in which D aims to distinguish the generated images from the real images, while G aims to fool the discriminator.

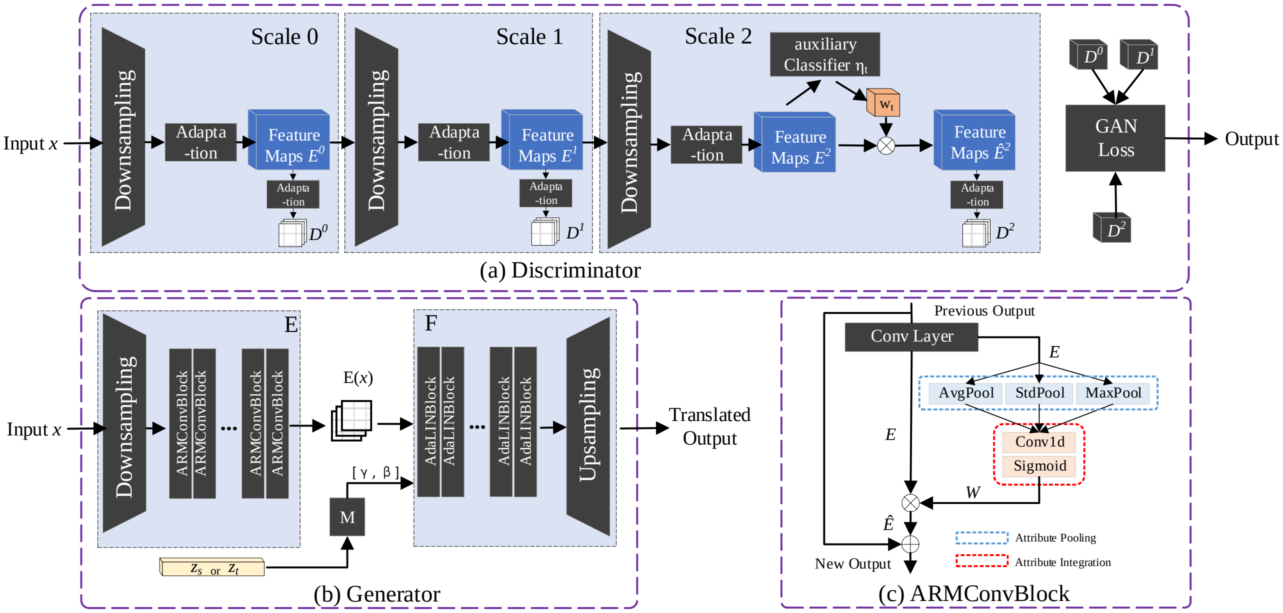
Unlike these methods [10, 13, 18, 28, 30], we only use one generator instead of two to complete the translation between two domains. Let x be the input to the generator and receive samples from both domains. z is used to control the translation direction of the generator(*e.g.*, $G(x, z_t)$: transform x to the target domain). We integrate the attribute-based recalibration module into the generator and introduce attribute consistency losses. Each scale shares a backbone in our multi-scale discriminator, and only the highest scale introduces an attention mechanism.

3.1. Generator

Let $x \in \{X_s, X_t\}$ represent a sample from the source and the target domain. Our translation model G consists of an encoder E, a latent mapping network M, and a decoder F. Fig. 2 (b) shows the structure of the generator.

The encoder E consists of a down-sampling module and a feature manipulate module. Most unsupervised image-to-image translation methods [10, 13, 18, 28, 30] use several residual blocks as feature manipulate modules. Beyond them, we introduce an auxiliary unit into each residual block to represent attribute features effectively in feature space. Inspired by style-based recalibration module [17], an attention mechanism, which adaptively recalibrates intermediate feature maps by exploiting their styles. On this basis, we propose an attribute-based recalibration module(ARM) to better exploit the attribute features by extracting more statistical features. This module guides the generator to efficiently extract the input image's attribute features and focus on regions that are critical to generating a realistic image. We introduce the ARM into the residual block and note that ARMConvBlock represents the residual block that integrates ARM. Fig. 2 (c) shows the structure of the ARMConvBlock. The ARM is comprised of two parts: *attribute pooling* and *attribute integration*. In *attribute pooling*, the average mean, standard deviation, and maximum value of the feature map are selected as attribute features. Let $E^k(x) \in \mathbb{R}^{C\times H\times W}$ be the k-th feature map of the input feature maps and $E^{k_{ij}}(x)$ be the value at (i, j), where C is the total number of channels, and H, W denote spatial dimensions. The attribute feature $T \in \mathbb{R}^{C\times 3}$ can be computed in each channel by:

$$\mu_k = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} E^{k_{ij}}(x), \tag{1}$$



(a) Discriminator

(b) Generator

(c) ARMConvBlock

Figure 2. **The model architecture of our method.** The detailed notations are described in Section Method.

$$\sigma_k = \sqrt{\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}(E^{k_{ij}}(x) - \mu_k)^2}, \qquad (2)$$

$$\alpha_k = \max_{1\leq i\leq H, 1\leq j\leq W} E^{k_{ij}}(x), \qquad (3)$$

$$t_k = [\mu_k, \sigma_k, \alpha_k]. \qquad (4)$$

The attribute vector $t_k \in \mathbb{R}^3$ serves as a summary description of the attribute information for channel k. In *attribute integration*, the attribute features $T \in R^{C\times 3}$ are converted into channel-wise attribute weights $W \in \mathbb{R}^{C\times 1}$ by using a 1-d convolutional layer and softmax activation function. Finally, the input feature E is recalibrated by the channel-wise attribute integration weights to produce the output $\hat{E} \in \mathbb{R}^{C\times H\times W}$ as: $\hat{E} = W \cdot E$.

The latent mapping network M is an eight-layer MLP network [23], which guides the generator to translate the image to the source domain or target domain, making our model integrate the two generators into one. The input of M is a latent vector, $\mathcal{Z}$, is represented by a n element vector, we generate $z_s, z_t \in \mathcal{Z}$ as:

$$z_s = \mathbf{1}_s + v, \quad z_t = \mathbf{1}_t + v, \quad v \sim \mathcal{N}(0, 0.2). \qquad (5)$$

where $\mathbf{1}_s$ is a n element vector that contains ones on elements 0 through $\frac{n}{2}$ and zeros elsewhere, and $\mathbf{1}_t$ is set the opposite way. The decoder F equips the residual blocks with AdaLIN [13] whose parameters, γ and β are dynamically computed by latent mapping network M.

The structure of the decoder F is similar to the U-GAT-IT [13], but we changed the way it gets γ and β , it uses the γ and β output by M to determine whether to generate a image in the source domain or the target domain.

3.2. Discriminator

The discriminators in our framework consist of D_s and D_t. We only discuss D_t here, the formulation of D_s is similar to D_t. Unlike the multi-scale discriminator proposed by the previous method, our multi-scale discriminator shares a backbone and only introduces the attention mechanism on the highest scale. Fig.2 (a) shows the structure of the discriminator.

The discriminator D_t contains two parts: n scales shared backbone discriminators D_t^0 to D_t^{n-1} (n is 3 in Fig.2 (a)), and an auxiliary classifier η_t. Each scale discriminators are consists of several down-sampling-convolution layers, and auxiliary classifier is CAM module [29]. D_t^0 is directly connected to the output of the downsampling layer, D_t^1 is connected to the output of D_t^0, and so on. We introduce the attention mechanism on the highest scale discriminator: suppose E_t^{n-1} is the output feature map of D_t^{n-1}, w_t is the weight of feature map got from the auxiliary classifier η_t, the output of the highest scale discriminator can calculated as: $\hat{E}_t^{n-1} = w_t * E_t^{n-1}$. The highest scale discriminator has a global receptive field, and the auxiliary classifier is easier to distinguish the true and fake images on this scale, which is helpful for the discriminator to distinguish the shape differences between the source domain and the target domain and guide the generator to deal with the deformation. We do not introduce the attention mechanism on the low-scale discriminator, making our method pay more attention to the local texture features and help guide the generator to produce a well-textured image.

3.3. Attribute Consistency Losses

3.3.1 Cycle-consistency loss.

In the task of unsupervised image-to-image translation, the assumption of cycle-consistency [30] is used to ensure the translated images contain enough information from the original input images and alleviate mode collapse problem. It aims to ensure the similarity between each pixel of the input image and the reconstructed image. The cycle-consistency loss is defined as:

$$\mathcal{L}_{cyc}(G_{s\to t}) = \mathbb{E}_{x\sim X_s}[\|G(G(x,z_t),z_s) - x\|_1]. \tag{6}$$

Here, x is the input image from source domain X_s, $G(x,z_t)$ is the generated image at target domain X_t, $G(G(x,z_t),z_s)$ is the reconstructed image at source domain X_s, we only discuss $G_{s\to t}$ here, the formulation of $G_{t\to s}$ is similar to $G_{s\to t}$.

However, the cycle-consistency loss is insufficient to ensure the attribute consistency between the input image and the generated image. We define the process of generating images to the target domain as $\mathcal{G}$ and the process of generating to the source domain as $\mathcal{F}$ (i.e. $\mathcal{G}(x) = G(x,z_t), \mathcal{F}(y) = G(y,z_s)$). The model can be viewed as two auto-encoders: $\mathcal{G}\circ\mathcal{F} : X_s \to X_s$ and $\mathcal{F}\circ\mathcal{G} : X_t \to X_t$ [5], where the translated image $\mathcal{G}(x)$ and $\mathcal{F}(y)$ can be viewed as intermediate representations. Therefore, the image can be coded as any representation so long as it can be decoded back to the original, which does not guarantee attribute consistency before and after translation.

Besides, the cycle-consistency assumes the images contain all the input images' information to reconstruct the input images. Therefore, the generated images may retain too much input image information, resulting in incomplete translation.

We proposed attribute consistency constraints to solve these problems by introducing attribute consistency loss and attribute cycle-consistency loss.

3.3.2 Attribute consistency loss.

Here, E is the encoder module in our generator. The attribute consistency loss ensures the similarity between the input image and the generated image in the attribute feature space.

$$\mathcal{L}_{attri}(G_{s\to t}) = \mathbb{E}_{x\sim X_s}[\|E(G(x,z_t)) - E(x)\|_1]. \tag{7}$$

3.3.3 Attribute cycle-consistency loss.

The attribute cycle-consistency loss ensures the similarity between the input image and the reconstructed image in the attribute feature space rather than in the image pixel space. The reconstructed image only needs to retain the attribute features of the input image, which reduces the constraints on the generator and helps to alleviate the problem of incomplete translation.

$$\mathcal{L}_{attri\_cyc}(G_{s\to t}) = \mathbb{E}_{x\sim X_s}[\|E(G(G(x,z_t),z_s)) - E(x)\|_1]. \tag{8}$$

3.4. Other Losses

3.4.1 Adversarial loss.

An adversarial loss is employed to match the distribution of the translated images to the target image distribution. We used the Least Squares GAN [21] objective for stable training:

$$\begin{aligned}\mathcal{L}_{lsgan}(G_{s\to t}) = &\mathbb{E}_{x\sim X_s}[(D_t(x))^2] + \\ &\mathbb{E}_{x\sim X_t}[(D_t(1 - G(x,z_t)))^2].\end{aligned} \tag{9}$$

3.4.2 Identity loss.

To ensure that the color distributions of input image and output image are similar, we apply an identity consistency constraint to the generator.

$$\mathcal{L}_{identity}(G_{s\to t}) = \mathbb{E}_{x\sim X_t}[\|G(x,z_t) - x\|_1]. \tag{10}$$

3.4.3 Full objective.

Finally, we jointly train the generator, discriminators to optimize the final objective:

$$\begin{aligned}\min_{G_{s\to t}} \max_{D_t} \quad &\lambda_1\mathcal{L}_{lsgan} + \lambda_2\mathcal{L}_{cyc} + \\ &\lambda_3\mathcal{L}_{attri} + \lambda_4\mathcal{L}_{attri\_cyc} + \lambda_5\mathcal{L}_{identity}.\end{aligned} \tag{11}$$

where $\lambda_1 = 1$, $\lambda_2 = 10$, $\lambda_3 = 0.5$, $\lambda_4 = 10$, $\lambda_5 = 5$. In the second half of the training process, we remove the $\mathcal{L}_{cyc}$ to achieve better results.

4. Experiments

4.1. Experiment Setup

4.1.1 Datasets.

We evaluate our model on three tasks, including selfie-to-anime, cat-to-dog, and dog-to-cat.

Selfie-to-Anime. The selfie-to-anime dataset [13] contains 3,400/100 selfie images and 3,400/100 anime face images in the training/test set. The image size is 256 × 256.

Cat-to-Dog and Dog-to-Cat. The AFHQ dataset [7] contains 4,739/500 dog face images and 5,153/500 cat face images in the training/test set. The image size is 512 × 512.

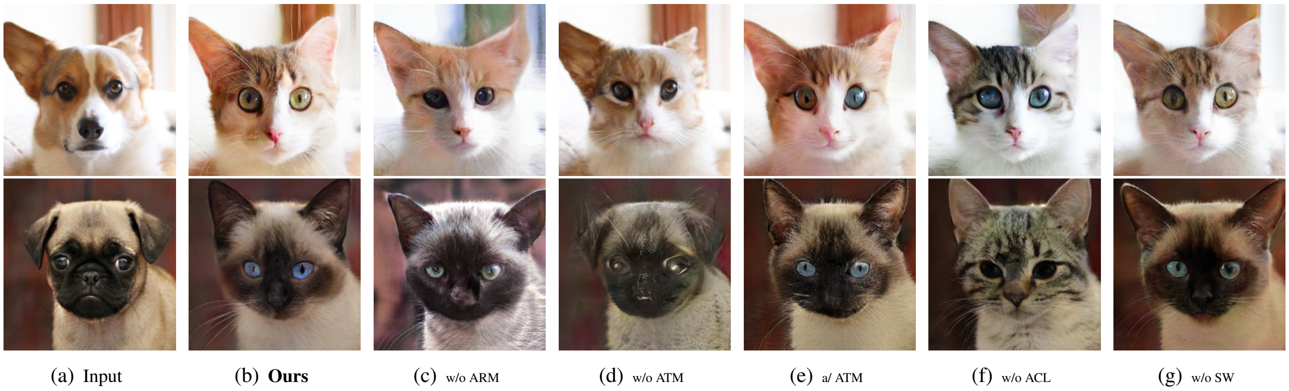


Figure 3. **Ablation studies.** The dog-to-cat translation results illustrate the effectiveness of our different settings. From left to right: input images; Ours; w/o ARM; w/o ATM; a/ ATM; w/o ACL; w/o SW.

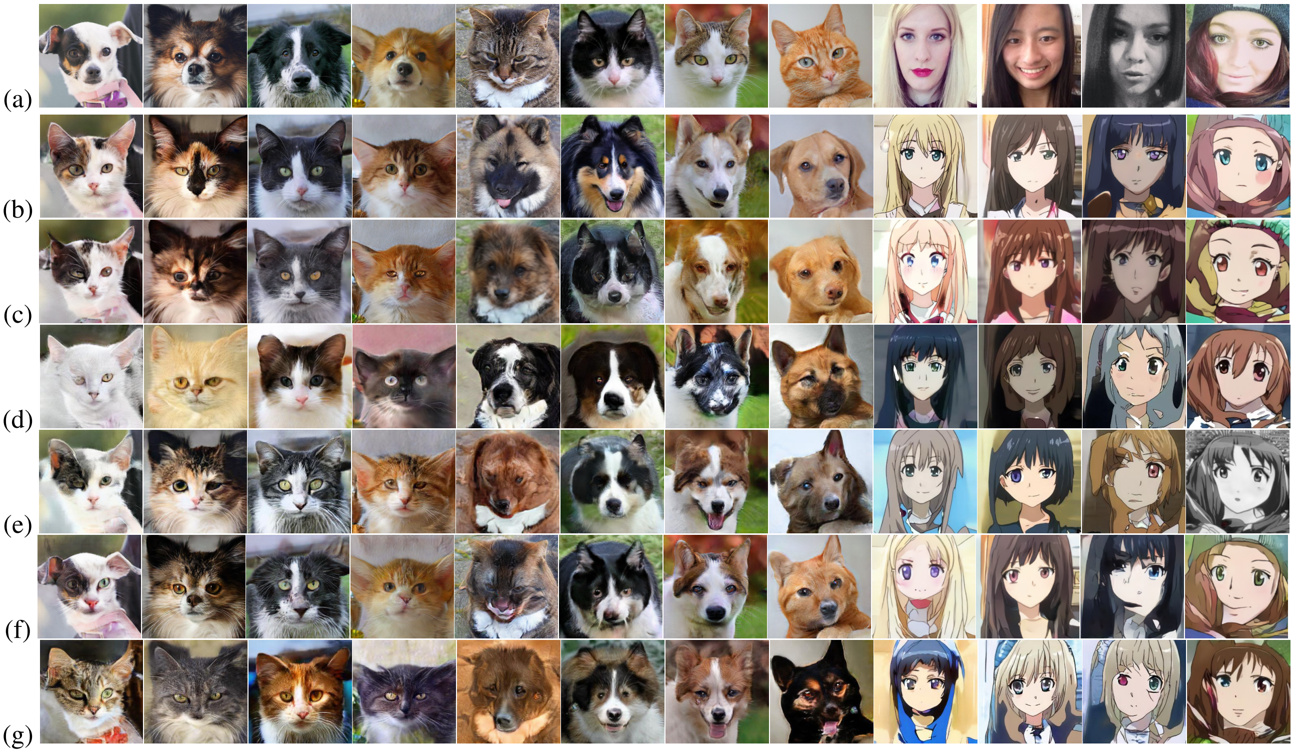


Figure 4. **Comparison against baselines on three tasks.** From left to right: dog-to-cat, cat-to-dog, and selfie2anime. (a)Inputs, (b)Ours, (c)U-GAT-IT [13], (d)ACL-GAN [28], (e)CouncilGAN [22], (f)CycleGAN [30], (g)MUNIT [10].

4.1.2 Baseline Models.

We compare our model to the state-of-the-art models for unsupervised image translation, including CycleGAN [30],U-GAT-IT [13],Council-GAN [22], MUNIT [10] and ACL-GAN [28]. We use the official pre-trained models if available, including the selfie-to-anime models of Council-GAN and U-GAT-IT. We reproduce the other results using the official open source code.

4.1.3 Evaluation Metrics.

We evaluate generated images on two metrics, the Fréch-et Inception Distance(FID) [9] and the kernel inception distance(KID) [2]. FID is a widely used metric for evaluating the image quality of generative models by comparing the distributions of the real and generated images. KID is an improvement on FID by adding unbiased estimates that more consistently match human perception.

4.1.4 Training.

Our model is trained using Adam [15] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We use a weight decay at a rate of 0.0001. For data augmentation, we first resize the images to 286×286, then randomly crops the images to 256×256, all images are flipped horizontally with a probability of 0.5. We train all models with a fixed learning rate of 0.0002 until 100 epochs and linearly decayed up to 200 epochs.

4.2. Ablation Studies

We analyzed our model by comparing five different settings: 1) with total module and loss(Ours), 2) without attribute-based recalibration module(Ours w/o ARM), 3) the highest scale discriminator does not introduce attention mechanism(Ours w/o ATM), 4) introduce attention mechanism into each scale discriminator(Ours a/ ATM), 5) without attribute consistency losses and cycle-consistency loss is used throughout the training(Ours w/o ACL), 6) without weight sharing of generator, using two generators instead of one(Ours w/o SW). The results of different settings are shown in Fig. 3, and the quantitative results are listed in Table 1.

We observed that the attribute-based recalibration module we introduced into the generator successfully helps preserve important attribute features of the input image in the translated image, compared with the setting of 'w/o ARM', the results are shown in Fig.3(b)(c). Besides, the ARM can also help generate more realistic images. The scores of FID and KID are significantly worse when removing the ARM. The attention mechanism introduced in the highest scale discriminator significantly promotes the generator to complete the deformation task, in 'w/o ATM' setting, as shown in Fig 3(d), it is difficult for the generator to transform the dog's eyes, nose and ears into the cat's shape. As mentioned above, the 'a/ ATM' setting will cause the model to pay too much attention to deformation and ignore the processing of texture details of generated images. Fig 3(e) shows that this setting will lead to inconsistent cat eye color and strange ear texture. Although the 'w/o ACL' setting can also generate very high-quality realistic images (its FID score and KID score are close to 'Ours'), it can not ensure the generated image retains the crucial attributes of the input image. Fig 3(f) shows the huge attribute inconsistency between the input and generated images; it illustrates the importance of attribute consistency losses. The 'w/o SW' setting gets worse scores in FID and KID, although Fig 3(g) shows that it can produce similar results with 'Ours'.

4.3. Comparison with Baselines

We compare our model to the baselines and summarize the quantitative results in Table 2. The qualitative results are shown in Fig. 4.

Model	FID	KID
Ours	**21.02**	**0.00160**
Ours w/o ARM	25.89	0.00448
Ours w/o ATM	56.37	0.00479
Ours a/ ATM	25.93	0.00178
Ours w/o ACL	21.76	0.00175
Ours w/o SW	22.21	0.00171

Table 1. Quantitative results of different ablation cases for dog-to-cat. Lower is better.

4.3.1 Selfie-to-anime.

Selfie-to-anime is a task that requires significant shape and texture change. The four columns on the right of Fig. 4 show the results of our model and baselines for selfie-to-anime. Our method can generate higher quality anime-style images while preserving the important attributes of the input images, *e.g.*, the hair color is better-preserved, and the layout of facial features are better-organized than baselines. Our method can generate images with finer hair textures, more detailed eyes and higher quality facial contour. In contrast, other methods often cause inconsistent sizes or color of the two eyes, unnatural hair texture, and important attributes such as hair color and head shape not being preserved. Our method outperforms all the baselines with a significant margin in FID and KID in Table 2.

4.3.2 Dog-to-cat and cat-to-dog.

Dog-to-cat is a task that transforms a dog face into a cat face while retaining the features of the input image; cat-to-dog is the opposite. The eight columns on the left of Fig. 4 show the results of our model and baselines for dog-to-cat and cat-to-dog. Our model does a better job than baselines in dealing with the deformation of eyes, nose, and ears. *e.g.*, ACL-GAN and CycleGAN are difficult to deal with deformation, and U-GAT-IT often cause unreal facial features and ears. Our model can better preserve attribute features than baselines. *e.g.*, the generated image inherits the input fur color and texture. The right columns of Table 2 show the quantitative results of dog-to-cat and cat-to-dog. Our method outperforms all baselines on both FID and KID.

4.4. User Study

We used user studies for qualitative evaluation. In the user studies, 138 people participated in the selfie-to-anime evaluation, 120 participated in the cat-to-dog evaluation, and 112 participated in the dog-to-cat evaluation. We inform only the name of the target domain, i.e., anime, dog, and cat to the participants. We randomly selected 100 images for user studies, including 34 in selfie-to-anime, 33 in dog-to-cat, and 33 in dog-to-cat. For each input image, we show participants the image results generated by

Model	Sefile-to-Anime		Dog-to-Cat		Cat-to-Dog	
	FID	KID	FID	KID	FID	KID
Ours	**77.69**	**0.01614**	**21.02**	**0.00160**	**44.59**	**0.00765**
U-GAT-IT	87.15	0.01627	26.07	0.00396	57.55	0.01804
ACL-GAN	97.94	0.03371	39.80	0.02021	94.75	0.05169
CouncilGAN	91.81	0.02541	40.19	0.01916	98.75	0.06633
CycleGAN	89.37	0.02269	70.81	0.03242	123.77	0.07093
MUNIT	94.80	0.02964	42.29	0.02215	85.58	0.04319

Table 2. Quantitative results of Ours and the baselines. Lower is better.

Model	Sefile-to-Anime	Dog-to-Cat	Cat-to-Dog
Ours	**72.42**	**77.03**	**80.41**
U-GAT-IT	12.70	15.69	13.38
ACL-GAN	7.21	1.84	1.79
CouncilGAN	4.39	4.25	1.77
MUNIT	3.28	1.19	2.65

Table 3. Preference score on translated images by user studies. Higher is better.

our method and other baseline methods (but do not tell participants which is the output of our method). We instructed participants to select images with higher image quality and better preservation of input image attributes. We compare our model to the baselines and summarize the qualitative results in Table 3. The results in Table 3 show that the score of our method is higher than that of all other baseline methods. User studies show that the images generated by our method generally have higher quality and can better retain more attribute information of the input image. At the same time, we also received feedback from some participants. Like other methods, our method is hard to retain detailed attribute information such as eye color. We will discuss it in the limitations section.

5. Limitations

The experiments have demonstrated that our method outperforms state-of-the-art methods both quantitatively and qualitatively on three tasks. Nevertheless, our method also has some limitations. The typical failure cases are shown in Fig. 5. We only use one generator to translate real images from domain X_s to X_t synchronously, however, the performance may be decreased in tasks with huge differences in objects or backgrounds between X_s and X_t. In addition, for the attributes of eye color and expression in the input image, our method can not ensure that the output image is consistent with the original image in these details. Thus, supporting the preservation of eye color and expression is an interesting direction for future studies.

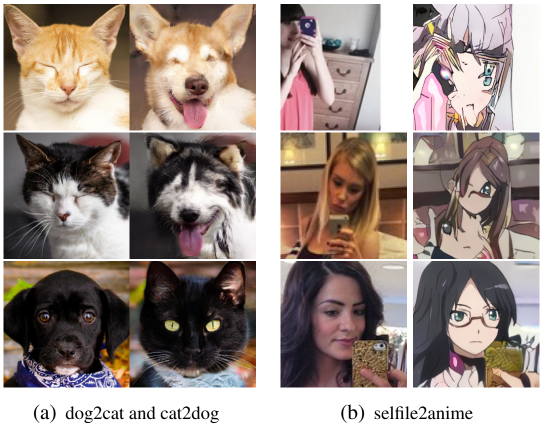

(a) dog2cat and cat2dog (b) selfile2anime

Figure 5. **Typical failure cases of our method.** (a) When the input cat closes its eyes, it may produce a dog without eyes; The eye color of the input image and the output image may be inconsistent. (b) The generated anime image may be distorted when the face in the input image is too small; Sometimes the generated image may appear glasses that are not in the input image.

6. Conclusions

This paper has proposed a novel generator with ARM and attribute consistency constraint and has proposed a new multi-scale discriminator for unsupervised image translation. Our generator better ensures the preservation of the attribute features of the input images. Our discriminator facilitates the shape deformation and improves generated image quality. We have shown in the experiments that our method improves the quality of the generated images, and generated images better preserve the attribute features of inputs. Our model outperforms the existing state-of-the-art methods on both FID and KID metrics.

7. Acknowledgments.

This work is supported by Industrial Prospective Project of Jiangsu Technology Department under Grant No.BE2018119.

References

[1] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. C. Courville. Augmented cyclegan:learning many-to-many mappings from unpaired data. In *Proceedings of ICML*, pages 195–204, 2018. 1

[2] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans. In *Proceedings of ICLR*, 2018. 6

[3] R. Chen, W. Huang, B. Huang, F. Sun, and B. Fang. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of CVPR*, pages 8165–8174, 2020. 3

[4] S.-Y. Chen, F.-L. Liu, Y.-K. Lai, P. L. Rosin, C. Li, H. Fu, and L. Gao. DeepFaceEditing: Deep face generation and editing with disentangled geometry and appearance control. *ACM Trans. Graph.*, 40(4):90:1–90:15, 2021. 3

[5] X. Chen, C. Xu, X. Yang, and D. Tao. Attention-gan for object transfiguration in wild images. In *Proceedings of ECCV*, pages 167–184, 2018. 5

[6] 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 *Proceedings of CVPR*, pages 8789–8797, 2018. 3

[7] Y. Choi, Y. Uh, J. Yoo, and J. Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of CVPR*, pages 8185–8194, 2020. 1, 3, 5

[8] I. Durugkar, I. Gemp, and S. Mahadevan. Generative multi-adversarial networks. In *Proceedings of ICLR*, 2017. 3

[9] 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 *Proceedings of NIPS*, pages 6626–6637, 2017. 6

[10] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of ECCV*, pages 179–196, 2018. 1, 3, 6

[11] G. Ian, P.-A. Jean, M. Mehdi, X. Bing, W.-F. David, O. Sherjil, C. Aaron, and B. Yoshua. Generative adversarial nets. In *Proceedings of NIPS*, pages 2672–2680, 2014. 1, 2, 3

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of CVPR*, pages 5967–5976, 2017. 2, 3

[13] 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 *Proceedings of ICLR*, 2020. 1, 3, 4, 5, 6

[14] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of ICML*, pages 1857–1865, 2017. 1, 3

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, 2015. 7

[16] C. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of CVPR*, pages 5548–5557. 3

[17] H. Lee, H.-E. Kim, and H. Nam. Srm: A style-based recalibration module for convolutional neural networks. In *Proceedings of ICCV*, pages 1854–1862, 2019. 3

[18] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. K. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of ECCV*, pages 36–52, 2018. 1, 3

[19] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Proceedings of NIPS*, pages 700–708, 2017. 1

[20] G. Lu, Z. Zhou, Y. Song, K. Ren, and Y. Yu. Guiding the one-to-one mapping in cyclegan via optimal transport. In *Proceedings of AAAI*, pages 4432–4439, 2019. 3

[21] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *Proceedings of ICCV*, pages 2813–2821, 2017. 5

[22] O. Nizan and A. Tal. Breaking the cycle - colleagues are all you need. In *Proceedings of CVPR*, pages 7857–7866, 2020. 1, 3, 6

[23] R. Or-El, S. Sengupta, O. Fried, E. Shechtman, and I. Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *Proceedings of ECCV*, pages 739–755, 2020. 4

[24] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. In *Proceedings of ICLR*, 2017. 1

[25] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. In *Proceedings of NIPS*, pages 1144–1156, 2018. 2

[26] 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 *Proceedings of CVPR*, pages 8798–8807, 2018. 2, 3

[27] Z. Yi, H. Zhang, P. Tan, and M. Gong. Unsupervised dual learning for image-to-image translation. In *Proceedings of ICCV*, pages 2849–2857, 2017. 1, 3

[28] Y. Zhao, R. Wu, and H. Dong. Unpaired image-to-image translation using adversarial consistency loss. In *Proceedings of ECCV*, pages 800–815, 2020. 1, 3, 6

[29] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of CVPR*, pages 2921–2929, 2016. 3, 4

[30] J.-Y. Zhu, T. Park, P. Isola, and A. AEfros. Unpaired image-to-image translation using cycle consistent adversarial networks. In *Proceedings of ICCV*, pages 2223–2232, 2017. 1, 3, 5, 6