Unpaired Image-to-Image Translation using Adversarial Consistency Loss
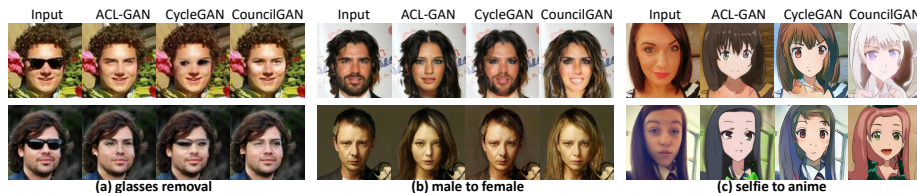
Yihao Zhao, Ruihai Wu, and Hao Dong

Hyperplane Lab, CFCS, Computer Science Dept., Peking University
{zhaoyh98, wuruihai, hao.dong}@pku.edu.cn

Abstract. Unpaired image-to-image translation is a class of vision problems whose goal is to find the mapping between different image domains using unpaired training data. Cycle-consistency loss is a widely used constraint for such problems. However, due to the strict pixel-level constraint, it cannot perform shape changes, remove large objects, or ignore irrelevant texture. In this paper, we propose a novel adversarial-consistency loss for image-to-image translation. This loss does not require the translated image to be translated back to be a specific source image but can encourage the translated images to retain important features of the source images and overcome the drawbacks of cycle-consistency loss noted above. Our method achieves state-of-the-art results on three challenging tasks: glasses removal, male-to-female translation, and selfie-to-anime translation.

Keywords: Generative Adversarial Networks, Dual Learning, Image Synthesis

1   Introduction

Fig. 1. Example results of our ACL-GAN and baselines. Our method does not require cycle consistency, so it can bypass unnecessary features. Moreover, with the proposed adversarial-consistency loss, our method can explicitly encourage the generator to maintain the commonalities between the source and target domains.

Learning to translate data from one domain to another is an important problem in computer vision for its wide range of applications, such as colourisation [34], super-resolution [7, 13], and image inpainting [25, 11]. In unsupervised

settings where only unpaired data are available, current methods [2, 6, 31, 10, 15, 20, 27, 33, 35] mainly rely on shared latent space and the assumption of cycle-consistency. In particular, by forcing the translated images to fool the discriminator using a classical adversarial loss and translating those images back to the original images, cycle-consistency ensures the translated images contain enough information from the original input images. This helps to build a reasonable mapping and generate high-quality results.

The main problem of cycle-consistency is that it assumes the translated images contain all the information of the input images in order to reconstruct the input images. This assumption leads to the preservation of source domain features, such as traces of large objects (*e.g.,* the trace of glasses in Fig. 1 (*a*)) and texture (*e.g.,* the beard remains on the face in Fig. 1 (*b*)), resulting in unrealistic results. Besides, the cycle-consistency constrains shape changes of source images (*e.g.,* the hair and the shape of the face are changed little in Fig. 1 (*b*) and (*c*)).

To avoid the drawbacks of cycle-consistency, we propose a novel method, termed ACL-GAN, where ACL stands for *adversarial-consistency loss*. Our goal is to maximise the information shared by the target domain and the source domain. The adversarial-consistency loss encourages the generated image to include more features from the source image, from the perspective of distribution, rather than maintaining the pixel-level consistency. That's to say, the image translated back is only required to be similar to the input source image but need not be identical to a specific source image. Therefore, our generator is not required to preserve all the information from the source image, which avoids leaving artefacts in the translated image. Combining the adversarial-consistency loss with classical adversarial loss and some additional losses makes our full objective for unpaired image-to-image translation.

The main contribution of our work is a novel method to support unpaired image-to-image translation which avoids the drawbacks of cycle-consistency. Our method achieves state-of-the-art results on three challenging tasks: glasses removal, male-to-female translation, and selfie-to-anime translation.

2   Related Work

Generative adversarial networks (GANs) [8] have been successfully applied to numerous image applications, such as super-resolution [17] and image colourisation [34]. These tasks can be seen as "translating" an image into another image. Those two images usually belong to two different domains, termed source domain and target domain. Traditionally, each of the image translation tasks is solved by a task-specific method. To achieve task-agnostic image translation, Pix2Pix [12] was the first work to support different image-to-image translation tasks using a single method. However, it requires paired images to supervise the training.

Unpaired image-to-image translation has gained a great deal of attention for applications in which paired data are unavailable or difficult to collect. A key problem of unpaired image-to-image translation is determining which properties of the source domain to preserve in the translated domain, and how to
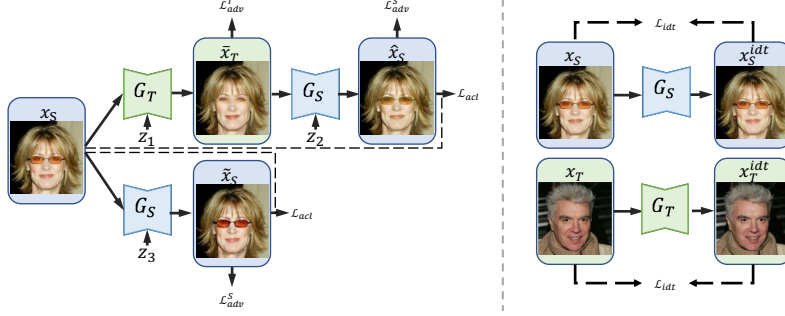
preserve them. Some methods have been proposed to preserve pixel-level properties, such as pixel gradients [5], pixel values [29], and pairwise sample distances [3]. Recently, several concurrent works, CycleGAN [35], DualGAN [33], DiscoGAN [15], and UNIT [20], achieve unpaired image-to-image translation using cycle-consistency loss as a constraint. Several other studies improve image-to-image translation in different aspects. For example, BicycleGAN [36] supports multi-modal translation using paired images as the supervision, while Augmented CycleGAN [1] achieves that with unpaired data. DRIT [19, 18] and MUNIT [10] disentangle domain-specific features and support multi-modal translations. To improve image quality, attention CycleGAN [23] proposes an attention mechanism to preserve the background from the source domain. In this paper, we adopt the network architecture of MUNIT but we do not use cycle-consistency loss.

To alleviate the problem of cycle-consistency, a recent work, CouncilGAN [24], has proposed using duplicate generators and discriminators together with the council loss as a replacement for cycle-consistency loss. By letting different generators compromise each other, the council loss encourages the translated images to contain more information from the source images. By contrast, our method does not require duplicate generators and discriminators, so it uses fewer network parameters and needs less computation. Moreover, our method can generate images of higher quality than CouncilGAN [24], because our adversarial-consistency loss explicitly makes the generated image similar to the source image, while CouncilGAN [24] only compromises among generated images. Details are described in Section 3.2.

3   Method

Our goal is to translate images from one domain to the other domain and support diverse and multi-modal outputs. Let X_S and X_T be the source and target 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, $x_S \in X_S$ and $x_T \in X_T$ be the images of different domains. We define p_X, p_S and p_T to be the distributions of X, X_S and X_T. $p_{(a,b)}$ is used for joint distribution of pair (a, b), where a and b can be images or noise vectors. Let Z be the noise vector space, $z \in Z$ be a noise vector and $z \sim \mathcal{N}(0, 1)$.

Our method has two generators: $G_S : (x, z) \to x_S$ and $G_T : (x, z) \to x_T$ which translate images to domain X_S and X_T, respectively. Similar to [10], each generator contains a noise encoder, an image encoder, and a decoder, in which the noise encoder is used only for calculating the identity loss. The generators receive input pairs (image, noise vector) where the image is from X. In detail, the image encoder receives images sampled from X. The noise vector z obtained from the noise encoder is only for identity loss, while for other losses, the noise vector z is randomly sampled from the standard normal distribution, $\mathcal{N}(0, 1)$. The noise vector z and the output of the image encoder are forwarded to the decoder and the output of the decoder is the translated image.

Fig. 2. The training schema of our model. (Left) Our model contains two generators: $G_S : (X, Z) \rightarrow X_S$ and $G_T : (X, Z) \rightarrow X_T$ and three discriminators: D_S, D_T for $\mathcal{L}^S_{adv}, \mathcal{L}^T_{adv}$ and $\hat{D}$ for $\mathcal{L}_{acl}$. D_S and D_T ensure that the translated images belong to the correct image domain, while $\hat{D}$ encourages the translated images to preserve important features of the source images. The noise vectors z_1, z_2, z_3 are randomly sampled from $\mathcal{N}(0, 1)$. (Right) $\mathcal{L}_{idt}$ encourages to maintain features, improves the image quality, stabilises the training process and prevents mode collapse, where the noise vector is from the noise encoder. The blocks with the same colour indicate shared parameters.

Moreover, there are two kinds of discriminators D_S/D_T and $\hat{D}$. $\hat{D}$ is a consistency discriminator. Its goal is to ensure the consistency between source images and translated images, and this is the core of our method. The goal of D_S and D_T is to distinguish between real and fake images in a certain domain. Specifically, the task of D_S is to distinguish between X_S and $G_S(X)$, and the task of D_T is to distinguish between X_T and $G_T(X)$.
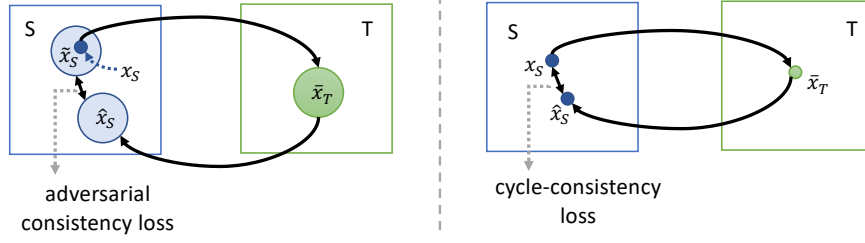
The objective of ACL-GAN has three parts. The first, *adversarial-translation loss*, matches the distributions of generated images to the data distributions in the target domain. The second, *adversarial-consistency loss*, preserves significant features of the source images in the translated images, *i.e.,* it results in reasonable mappings between domains. The third, identity loss and bounded focus mask, can further help to improve the image quality and maintain the image background. The data are forwarded as shown in Fig. 2, and the details of our method are described below.

3.1   Adversarial-Translation Loss

For image translation between domains, we utilise the classical adversarial loss, which we call *adversarial-translation loss* in our method, to both generators, G_S and G_T, and discriminators, D_S and D_T. For generator G_T and its discriminator D_T, the adversarial-translation loss is as follows:

$$\mathcal{L}^T_{adv}(G_T, D_T, X_S, X_T) = \mathbb{E}_{x_T \sim p_T}[log D_T(x_T)] \\ + \mathbb{E}_{\bar{x}_T \sim p_{\{\bar{x}_T\}}}[log(1 - D_T(\bar{x}_T))] \tag{1}$$

Fig. 3. The comparison of adversarial-consistency loss and cycle-consistency loss [35]. The blue and green rectangles represent image domains S and T, respectively. Any point inside a rectangle represents a specific image in that domain. (Right): given a source image x_S, cycle-consistency loss requires the image translated back, $\hat{x}_S$, should be identical to the source image, x_S. (Left): given a source image x_S, we synthesise multi-modal images in its neighbourhood distribution $\tilde{x}_S$, the distribution $\bar{x}_T$ in the target domain X_T, and the distribution $\hat{x}_S$ translated from $\bar{x}_T$. The distributions are indicated by the blue and green circles. Instead of requiring the image translated back, $\hat{x}_S$, to be a specific image, we minimise the distance between the distributions of $\hat{x}_S$ and $\tilde{x}_S$, so that a specific $\hat{x}_S$ can be any point around x_S. By doing so, we encourage $\bar{x}_T$ to preserve the features of the original image x_S.

where $\bar{x}_T = G_T(x_S, z_1)$ and $z_1 \sim \mathcal{N}(0, 1)$. The objective is $min_{G_T} max_{D_T}$ $\mathcal{L}_{adv}^T(G_T, D_T, X_S, X_T)$.

The discriminator D_S is expected to distinguish between real images of domain X_S and translated images generated by G_S. The generator G_S tries to generate images, $\hat{x}_S$ and $\tilde{x}_S$, that look similar to images from domain X_S. Therefore, the loss function is defined as:

$$\mathcal{L}_{adv}^S(G_S, D_S, \{\bar{x}_T\}, X_S) = \mathbb{E}_{x_S \sim p_S}[log D_S(x_S)] \\ + (\mathbb{E}_{\hat{x}_S \sim p_{\{\hat{x}_S\}}}[log(1 - D_S(\hat{x}_S))] \\ + \mathbb{E}_{\tilde{x}_S \sim p_{\{\tilde{x}_S\}}}[log(1 - D_S(\tilde{x}_S))])/2 \quad (2)$$

where $\hat{x}_S = G_S(\bar{x}_T, z_2)$, $\tilde{x}_S = G_S(x_S, z_3), z_2, z_3 \sim \mathcal{N}(0, 1)$ and the objective is $min_{G_S} max_{D_S} \mathcal{L}_{adv}^S(G_S, D_S, \{\bar{x}_T\}, X_S)$. In summary, we can define the adversarial-translation loss as follows:

$$\mathcal{L}_{adv} = \mathcal{L}_{adv}^T(G_T, D_T, X_S, X_T) + \mathcal{L}_{adv}^S(G_S, D_S, \{\bar{x}_T\}, X_S) \quad (3)$$

3.2 Adversarial-Consistency Loss

The $\mathcal{L}_{adv}$ loss described above can encourage the translated image, $\bar{x}_T$, to be in the correct domain X_T. However, this loss cannot encourage the translated image $\bar{x}_T$ to be similar to the source image x_S. For example, when translating a male to a female, the facial features of the female might not be related to those of the male. To preserve important features of the source image in the translated image, we propose the adversarial-consistency loss, which is realised by a consistency

discriminator $\hat{D}$. The consistency discriminator impels the generator to minimise the distance between images $\tilde{x}_S$ and $\hat{x}_S$ as shown in Fig. 3. The "real" and "fake" images for $\hat{D}$ can be swapped without affecting performance. However, letting $\hat{D}$ only distinguish $\hat{x}_S$ and $\tilde{x}_S$ does not satisfy our needs, because the translated images $\hat{x}_S$ and $\tilde{x}_S$ need only to belong to domain X_S; they are not required to be close to a specific source image. Therefore, the consistency discriminator $\hat{D}$ uses x_S as a reference and adopts paired images as inputs to let the generator minimise the distances between the joint distributions of $(x_S, \hat{x}_S)$ and $(x_S, \tilde{x}_S)$. In this way, the consistency discriminator $\hat{D}$ encourages the image translated back, $\hat{x}_S$, to contain the features of the source image x_S. As $\hat{x}_S$ is generated from $\bar{x}_T$, this can encourage the translated image $\bar{x}_T$ to preserve the features of the source image x_S.

The input noise vector z enables multi-modal outputs, which is essential to make our method work. Without multi-modal outputs, given a specific input image x_S, the $\tilde{x}_S$ can have only one case. Therefore, mapping $(x_S, \hat{x}_S)$ and $(x_S, \tilde{x}_S)$ together is almost equivalent to requiring $\hat{x}_S$ and x_S to be identical. This strong constraint is similar to cycle-consistent loss whose drawbacks have been discussed before. With the multi-modal outputs, given a specific image x_S, the $\tilde{x}_S$ can have many possible cases. Therefore, the consistency discriminator $\hat{D}$ can focus on the feature level, rather than the pixel level. That is to say, $\hat{x}_S$ does not have to be identical to a specific image x_S. For example, when translating faces with glasses to faces without glasses, the $\tilde{x}_S$ and $\hat{x}_S$ can be faces with different glasses, *e.g.* the glasses have different colours and frames in x_S, $\tilde{x}_S$ and $\hat{x}_S$ in Fig. 2. Thus, $\bar{x}_T$ need not retain any trace of glasses at the risk of increasing $\mathcal{L}_{adv}^T$, and $\mathcal{L}_{acl}$ can still be small.

The adversarial-consistency loss (ACL) is as follows:

$$\begin{aligned} \mathcal{L}_{acl} = \; &\mathbb{E}_{(x_S, \hat{x}_S) \sim p_{(X_S, \{\hat{x}_S\})}}[log\hat{D}(x_S, \hat{x}_S)] \\ &+ \mathbb{E}_{(x_S, \tilde{x}_S) \sim p_{(X_S, \{\tilde{x}_S\})}}[log(1 - \hat{D}(x_S, \tilde{x}_S))] \end{aligned} \tag{4}$$

where $x_S \in X_S$, $\bar{x}_T = G_T(x_S, z_1)$, $\hat{x}_S = G_S(\bar{x}_T, z_2)$, $\tilde{x}_S = G_S(x_S, z_3)$.

3.3   Other Losses

Identity loss We further apply identity loss to encourage the generators to be approximate identity mappings when images of the target domain are given to the generators. Identity loss can further encourage feature preservation, improve the quality of translated images, stabilise the training process, and avoid mode collapse because the generator is required to be able to synthesise all images in the dataset [35, 26]. Moreover, the identity loss between the source image x_S and the reconstruction image x_S^{idt} can guarantee that x_S is inside the distribution of $\tilde{x}_S$ as shown in Fig. 3.

We formalise two noise encoder networks, $E_S^z : X_S \to Z$ and $E_T^z : X_T \to Z$ for G_S and G_T, respectively, which map the images to the noise vectors. Identity loss can be formalised as:

$$\mathcal{L}_{idt} = \mathbb{E}_{x_S \sim p_S}[||x_S - x_S^{idt}||_1] + \mathbb{E}_{x_T \sim p_T}[||x_T - x_T^{idt}||_1] \tag{5}$$

where $x_S^{idt} = G_S(x_S, E_S^z(x_S))$ and $x_T^{idt} = G_T(x_T, E_T^z(x_T))$.

Bounded focus mask Some applications require the generator to only modify certain areas of the source image and keep the rest unchanged. We let the generator produce four channels, where the first three are the channels of RGB images and the fourth is called bounded focus mask whose values are between 0 and 1. The translated image x_T can be obtained by the formula: $x_T = x\prime_T \odot x_m + x_S \odot (1 - x_m)$, where $\odot$ is element-wise product, x_S is the source image, $x\prime_T$ is the first three output channels of the generator and x_m is the bounded focus mask. We add the following constraints to the generator which is one of our contributions:

$$\mathcal{L}_{mask} = \delta[(\max\{\sum_k x_m[k] - \delta_{max} \times W, 0\})^2$$
$$+ (\max\{\delta_{min} \times W - \sum_k x_m[k], 0\})^2]$$
$$+ \sum_k \frac{1}{|x_m[k] - 0.5| + \epsilon} \tag{6}$$

where δ, δ_{max} and δ_{min} are hyper-parameters for controlling the size of masks, $x_m[k]$ is the k-th pixel of the mask and W is the number of pixels of an image. The ϵ is a marginal value to avoid dividing by zero. The first term of this loss limits the size of the mask to a suitable range. It encourages the generator to make enough changes and maintain the background, where δ_{max} and δ_{min} are the maximum and minimum proportions of the foreground in the mask. The minimum proportion is essential for our method because it avoids $\tilde{x}_S$ being identical to x_S under different noise vectors. The last term of this loss encourages the mask values to be either 0 or 1 to segment the image into a foreground and a background [24]. In the end, this loss is normalised by the size of the image.

3.4   Implementation Details

Full objective. Our total loss is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{adv} + \lambda_{acl}\mathcal{L}_{acl} + \lambda_{idt}\mathcal{L}_{idt} + \lambda_{mask}\mathcal{L}_{mask} \tag{7}$$

where $\lambda_{acl}, \lambda_{idt}, \lambda_{mask}$ are all scale values that control the weights of different losses. We compare our method against ablations of the full objective in Section 4.2 to show the importance of each component.

Network architecture For generator and discriminator, we follow the design of those in [10]. Specifically, a generator consists of two encoders and one decoder. Besides the image encoder and decoder that together form an auto-encoder architecture, our model employs the noise encoder, whose architecture is similar to the style encoder in [10]. Meanwhile, our discriminators use a multi-scale technique [32] to improve the visual quality of synthesised images.

Training details We adopt a least-square loss [22] for $\mathcal{L}_{adv}$ (Equation 3) and $\mathcal{L}_{acl}$ (Equation 4). This loss brings more stable training process and better

results compared with [8]. For all the experiments, we used Adam optimiser [16] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The batch size was set to 3. All models were trained with a learning rate of 0.0001 and the learning rate dropped by a factor of 0.5 after every $100K$ iterations. We trained all models for $350K$ iterations. The discriminators update twice while the generators update once. For training, we set $\delta = 0.001$, $\epsilon = 0.01$ and $\lambda_{idt} = 1$. The values of λ_{acl}, λ_{mask}, δ_{min} and δ_{max} were set according to different applications: in glasses removal $\lambda_{acl} = 0.2, \lambda_{mask} = 0.025, \delta_{min} = 0.05, \delta_{max} = 0.1$; in male-to-female translation $\lambda_{acl} = 0.2, \lambda_{mask} = 0.025, \delta_{min} = 0.3, \delta_{max} = 0.5$; in selfie-to-anime translation $\lambda_{acl} = 0.5, \lambda_{mask} = \delta_{min} = \delta_{max} = 0$. For fair comparison, we follow the same data augmentation method described in CouncilGAN [24].

4   Experiments

4.1   Experimental Settings

Datasets We evaluated ACL-GAN on two different datasets, CelebA [21] and selfie2anime [14]. CelebA [21] contains 202,599 face images with 40 binary attributes. We used the attributes of gender and with/without glasses for evaluation. For both attributes, we used 162,770 images for training and 39,829 images for testing. For gender, the training set contained 68,261 images of males and 94,509 images of females and the test set contained 16,173 images of males and 23,656 images of females. For with/without glasses, the training set contained 10,521 images with glasses and 152,249 images without glasses and the test set contained 2,672 images with glasses and 37,157 images without glasses. Selfie2anime [14] contains 7,000 images. The size of the training set is 3,400 for both anime images and selfie images. The test set has 100 anime images and 100 selfie images.

Metrics Fréchet Inception Distance (FID) [9] is an improvement of Inception Score (IS) [28] for evaluating the image quality of generative models. FID calculates the Fréchet distance with mean and covariance between the real and the fake image distributions. Kernel Inception Distance (KID) [4] is an improved measure of GAN convergence and quality. It is the squared Maximum Mean Discrepancy between inception representations.

Baselines. We compared the results of our method with those of some state-of-the-art models, including CycleGAN [35], MUNIT [10], DRIT++ [19, 18], StarGAN [6], U-GAT-IT [14], Fixed-Point GAN [30], and CouncilGAN [24]. All those methods use unpaired training data. MUNIT [10], DRIT++ [19, 18], and CouncilGAN [24] can generate multiple results for a single input image and the others produce only one image for each input. Among those methods, only CouncilGAN [24] does not use cycle-consistency loss. Instead, it adopts N duplicate generators and $2N$ duplicate discriminators (N is set to be 2, 4 or 6 in their paper). This requires much more computation and memory for training. We set N to be 4 in our comparison.

Fig. 4. Ablation studies. The male-to-female translation results illustrate the importance of different losses. From left to right: input images; ACL-GAN (with total loss); ACL-A (without $\mathcal{L}_{acl}$); ACL-I (without $\mathcal{L}_{idt}$); ACL-M (without $\mathcal{L}_{mask}$).

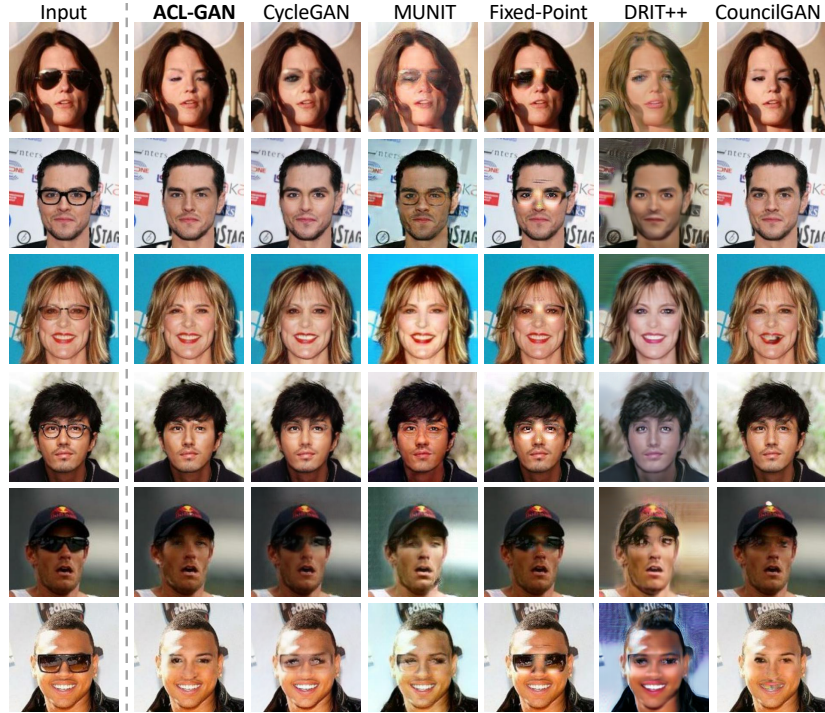| Model	$\mathcal{L}_{acl}$	$\mathcal{L}_{idt}$	$\mathcal{L}_{mask}$	FID	KID
ACL-A	-	✓	✓	23.96	0.023 ± 0.0003
ACL-I	✓	-	✓	18.39	0.018 ± 0.0004
ACL-M	✓	✓	-	17.72	0.017 ± 0.0004
ACL-GAN	✓	✓	✓	**16.63**	**0.015 ± 0.0003**

Table 1. Quantitative results of different ablation settings on male-to-female translation. For both FID and KID, lower is better. For KID, mean and standard deviation are listed. The ACL-GAN with total loss outperforms other settings.

The quantitative and qualitative results of the baselines were obtained by running the official public codes, except for the CouncilGAN, where the results were from our reproduction.

4.2 Ablation Studies

We analysed ACL-GAN by comparing four different settings: 1) with total loss (ACL-GAN), 2) without adversarial-consistency loss $\mathcal{L}_{acl}$ (ACL-A), 3) without identity loss $\mathcal{L}_{idt}$ (ACL-I), and 4) without bounded focus mask and its corresponding loss $\mathcal{L}_{mask}$ (ACL-M). The results of different settings are shown in Fig. 4 and the quantitative results are listed in Table 1.

Qualitatively, we observed that our adversarial-consistency loss, $\mathcal{L}_{acl}$, successfully helps to preserve important features of the source image in the translated image, compared with the setting of ACL-A which already has identity loss
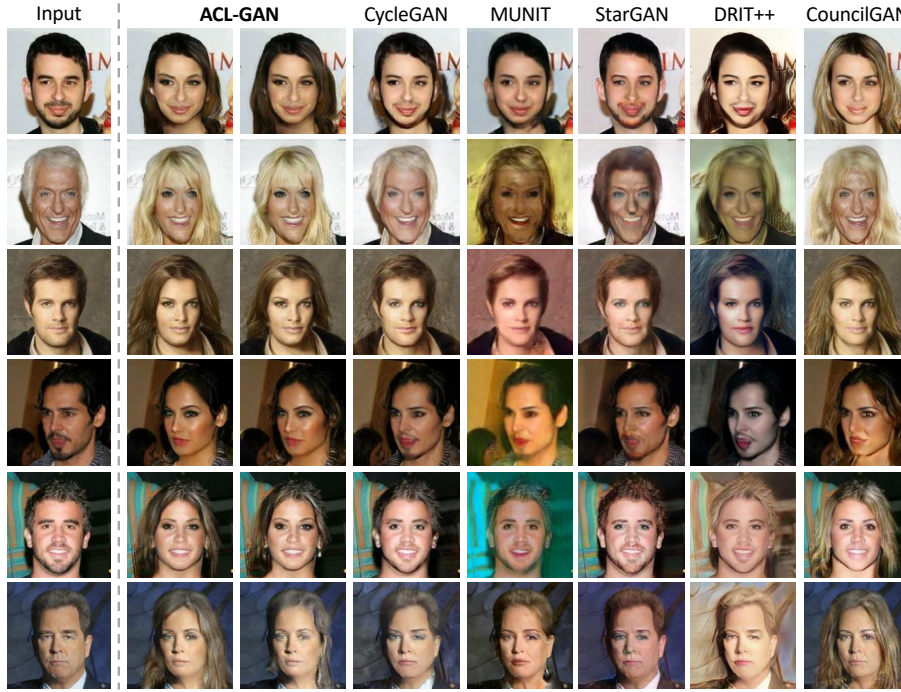
Fig. 5. Comparison against baselines on glasses removal. From left to right: input, our ACL-GAN, CycleGAN [35], MUNIT [10], Fixed-Point GAN [30], DRIT++ [19, 18], and CouncilGAN [24].

and bounded focus mask. For example, as the red arrows on results of ACL-A show, without $\mathcal{L}_{acl}$, the facial features, *e.g.* skin colour, skin wrinkles, and teeth, are difficult to maintain, and the quantitative results are the worst. We found that ACL-GAN with bounded focus mask has better perceptual quality, *e.g.,* the background can be better maintained, compared with ACL-M. The quantitative results indicate that our bounded focus mask improves the quality of images because the mask directs the generator to concentrate on essential parts for translation. Even though it is difficult to qualitatively compare ACL-I with the others, ACL-GAN achieves better quantitative results than ACL-I.

4.3   Comparison with Baselines

We compared our ACL-GAN with baselines on three challenging applications: glasses removal, male-to-female translation, and selfie-to-anime translation. These three applications have different commonalities between the source and target domains, and they require changing areas with different sizes. Glasses removal requires changing the smallest area, while the selfie-to-anime translation requires

| Input | **ACL-GAN** | CycleGAN | MUNIT | StarGAN | DRIT++ | CouncilGAN |

Fig. 6. Comparison against baselines on male-to-female translation. We show two translated imagee of our model under different noise vectors. From left to right: input, our ACL-GAN, CycleGAN [35], MUNIT [10], StarGAN [6], DRIT++ [19, 18], and CouncilGAN [24].

changing the largest area. Table 2 shows the number of parameters of our method and baselines. Table 3 shows the quantitative results for different applications and our method outperforms the baselines. Fig. 5, 6 and 7 show the results of our method and baselines.

Glasses removal The goal of glasses removal is to remove the glasses of a person in a given image. There are two difficulties in this application. First, the area outside the glasses should not be changed, which requires the generator to identify the glasses area. Second, the glasses hide some information of the face, such as the eyes and eyebrows. The sunglasses in some images make it even more difficult because the eyes are totally occluded and the generator is expected to generate realistic and suitable eyes.

Fig. 5 shows the results of ACL-GAN and baselines for glasses removal. Our results leave fewer traces of glasses than [35, 10, 30, 19, 18] because cycle-consistency is not required. The left columns of Table 3 show the quantitative results of glasses removal. Our method outperforms all baselines on both FID and KID. It is interesting that our method can outperform CouncilGAN and MU-

Fig. 7. Comparison against baselines on selfie-to-anime translation. From left to right: input, our ACL-GAN, CycleGAN [35], MUNIT [10], U-GAT-IT [14], DRIT++ [19, 18], and CouncilGAN [24].

NIT, which use the same network architecture as ACL-GAN. Compared with MUNIT, our method does not require the image translated back to be a specific image so that no artefacts remain. Compared with CouncilGAN, it might be that, rather than requiring multiple duplicate generators to compromise each other by minimising the distances between their synthesised images, our method explicitly encourages the generated image to be similar to the source image.

Male-to-female translation The goal of male-to-female translation is to generate a female face when given a male face, and these two faces should be similar except for gender. Comparing with glasses removal, this task does not require to remove objects, but there are three difficulties of this translation task. First, this task is a typical multi-modal problem because there are many possible ways to translate a male face to a female face *e.g.,* different lengths of hair. Second, translating male to female realistically requires not only changing the colour and texture but also shape, *e.g.,* the hair and beard. Third, the paired data is impossible to acquire, *i.e.,* this task can only be solved by using unpaired training data.

Model	CycleGAN	MUNIT	DRIT++	Fixed-Point GAN
Params	28.3M	46.6M	65.0M	53.2M

Model	StarGAN	U-GAT-IT	CouncilGAN	**ACL-GAN**
Params	53.3M	134.0M	126.3M	**54.9M**

Table 2. The number of parameters of our method and baselines. We set the image size to be 256×256 for all methods. The parameters of U-GAT-IT [14] is counted in light mode. The number of councils is set to be four in CouncilGAN [24].
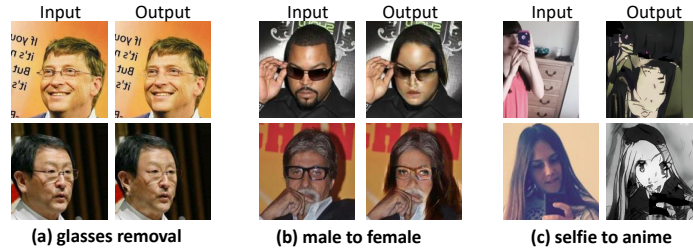
Model	glasses removal		Model	male to female		Model	selfie to anime	
	FID	KID		FID	KID		FID	KID
CycleGAN	48.71	0.043 ±0.0011	CycleGAN	21.30	0.021 ±0.0003	CycleGAN	102.92	0.042 ±0.0019
MUNIT	28.58	0.026 ±0.0009	MUNIT	19.02	0.019 ±0.0004	MUNIT	101.30	0.043 ±0.0041
DRIT++	33.06	0.026 ±0.0006	DRIT++	24.61	0.023 ±0.0002	DRIT++	104.40	0.050 ±0.0028
Fixed-Point GAN	44.22	0.038 ±0.0009	StarGAN	36.17	0.034 ±0.0005	U-GAT-IT	99.15	0.039 ±0.0030
CouncilGAN	27.77	0.025 ±0.0011	CouncilGAN	18.10	0.017 ±0.0004	CouncilGAN	98.87	0.042 ±0.0047
ACL-GAN	**23.72**	**0.020** ± **0.0010**	ACL-GAN	**16.63**	**0.015** ± **0.0003**	ACL-GAN	**93.58**	**0.037** ± **0.0036**

Table 3. Quantitative results of glasses removal, male-to-female translation, and selfie-to-anime translation. For KID, mean and standard deviation are listed. A lower score means better performance. U-GAT-IT [14] is in light mode. Our method outperforms all other baselines in all applications.

Fig. 6 compares our ACL-GAN with baselines for male-to-female translation. For each row, two images of ACL-GAN are generated with two random noise vectors. This shows the diversity of the results of our method, which is the same in Fig. 7. Our results are more feminine than baselines with cycle-consistency loss [35, 6, 19, 18] as our faces have no beard, longer hair and more feminine lips and eyes. We found that MUNIT [10] can generate images with long hair and no beard, which may contribute to the style latent code. However, the important features, *e.g.* the hue of the image, cannot be well-preserved. Quantitatively, the middle columns of Table 3 show the effectiveness of our method and it can be also explained by the lack of cycle-consistency loss [35] and compromise between duplicated generators [24].

Selfie-to-anime translation Different from the previous two tasks, generating an animated image conditioned on a selfie requires large modifications of shape. The structure and style of the selfie are changed greatly in the target domain, *e.g.,* the eyes become larger and the mouth becomes smaller. This may lead to the contortion and dislocation of the facial features.

Fig. 7 shows the results of ACL-GAN and baselines for selfie-to-anime translation. The results of methods with cycle-consistency are less in accordance with

Input   Output   Input   Output   Input   Output

(a) glasses removal   (b) male to female   (c) selfie to anime

Fig. 8. Typical failure cases of our method. (a) Some glasses are too inconspicuous to be identified and they can not be completely removed by the generator. (b) We also found that the glasses might be partly erased when translating a male to a female and this is because of the imbalance inherent to the datasets. (c) The generated anime image may be distorted and blurred when the face in the input image is obscured or too small, because it is far from the main distribution of the selfie domain.

the style of anime. In contrast, without cycle-consistency loss, our method can generate images more like an anime, *e.g.* the size and the layout of facial features are better-organized than baselines. The adversarial-consistency loss helps to preserve features, *e.g.,* the haircut and face rotation of the source images. U-GAT-IT [14] utilises an attention module and AdaIN to produce more visually pleasing results. The full mode of U-GAT-IT has 670.8M parameters, which is more than twelve times the number of parameters our model has. For fairness, we evaluated U-GAT-IT in light mode for comparison. Our method can still outperform U-GAT-IT when it is in light mode, which still uses more than twice as many parameters as ours as shown in Table 2.

5   Limitations and Discussion

In this paper, we present a novel framework, ACL-GAN, for unpaired image-to-image translation. The core of ACL-GAN, adversarial-consistency loss, helps to maintain commonalities between the source and target domains. Furthermore, ACL-GAN can perform shape modifications and remove large objects without unrealistic traces.

Although our method outperforms state-of-the-art methods both quantitatively and qualitatively on three challenging tasks, typical failure cases are shown in Fig. 8. We also tried learning to translate real images from domain X_S to X_T and from domain X_T to X_S synchronously, but the performance was decreased, which may be caused by the increase of network workload. In addition, our method is not yet able to deal with datasets that have complex backgrounds (*e.g.* horse-to-zebra) well. Thus, supporting images with a complex background is an interesting direction for future studies. Nevertheless, the proposed method is simple, effective, and we believe our method can be applied to different data modalities.

References

1. Almahairi, A., Rajeswar, S., Sordoni, A., Bachman, P., Courville, A.: Augmented cyclegan: Learning many-to-many mappings from unpaired data. In: International Conference on Machine Learning (2018)
2. Anoosheh, A., Agustsson, E., Timofte, R., Gool, L.V.: Combogan: Unrestrained scalability for image domain translation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (2018)
3. Benaim, S., Wolf, L.: One-sided unsupervised domain mapping. In: Conference on Neural Information Processing Systems (2017)
4. Bikowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: International Conference on Learning Representations (2018)
5. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
6. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
7. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2016)
8. Goodfellow, I., Pougetabadie, J., Mirza, M., Xu, B., Wardefarley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Conference on Neural Information Processing Systems (2014)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Conference on Neural Information Processing Systems (2017)
10. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: European Conference on Computer Vision (2018)
11. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. In: ACM Transactions on Graphics (2017)
12. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
13. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
14. Kim, J., Kim, M., Kang, H., Lee, K.H.: U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In: International Conference on Learning Representations (2020)
15. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: International Conference on Machine Learning (2017)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (2014)
17. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)

18. Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M., Yang, M.H.: Drit++: Diverse image-to-image translation via disentangled representations. International Journal of Computer Vision (2020)
19. Lee, H., Tseng, H., Huang, J., Singh, M., Yang, M.: Diverse image-to-image translation via disentangled representations. In: European Conference on Computer Vision (2018)
20. Liu, M., Breuel, T.M., Kautz, J.: Unsupervised image-to-image translation networks. In: Conference on Neural Information Processing Systems (2017)
21. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: IEEE International Conference on Computer Vision (2015)
22. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
23. Mejjati, Y.A., Richardt, C., Tompkin, J., Cosker, D., Kim, K.I.: Unsupervised attention-guided image-to-image translation. In: Conference on Neural Information Processing Systems (2018)
24. Nizan, O., Tal, A.: Breaking the cycle – colleagues are all you need. In: arXiv preprint arXiv 1911.10538 (2019)
25. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
26. Rosca, M., Lakshminarayanan, B., Warde-Farley, D., Mohamed, S.: Variational approaches for auto-encoding generative adversarial networks. In: arXiv preprint arXiv:1706.04987 (2017)
27. Royer, A., Bousmalis, K., Gouws, S., Bertsch, F., Mosseri, I., Cole, F., Murphy, K.: Xgan: Unsupervised image-to-image translation for many-to-many mappings. In: International Conference on Machine Learning (2018)
28. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Conference on Neural Information Processing Systems (2016)
29. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
30. Siddiquee, M.M.R., Zhou, Z., Tajbakhsh, N., Feng, R., Gotway, M.B., Bengio, Y., Liang, J.: Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In: IEEE International Conference on Computer Vision (2019)
31. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. In: International Conference on Learning Representations (2017)
32. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
33. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: IEEE International Conference on Computer Vision (2017)
34. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European Conference on Computer Vision (2016)
35. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (2017)

36. Zhu, J., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Conference on Neural Information Processing Systems (2017)