ECycleGAN: Enhanced Cycle-Consistent Generative Adversarial Networks

Xianchao Zhang
School of Software
Dalian University of Technology
Dalian, China
xczhang@dlut.edu.cn

Changjia Zhou*
School of Software
Dalian University of Technology
Dalian, China
zcj@mail.dlut.edu.cn

ABSTRACT
Unsupervised image-to-image translation, which aims in translating two irrelevant domains of images, has increased substantially in recent years with the success of Generative Adversarial Networks (GANs) based on the cycle-consistency assumption. Especially, the Cycle-Consistent Generative Adversarial Network (CycleGAN) has shown remarkable success for two domains translation. However, the details about texture and style are often accompanied with unpleasant artifacts. To further enhance the translational quality, we thoroughly study the key components of CycleGAN - network architecture and adversarial loss, and improve each of them to derive an Enhanced CycleGAN (ECycleGAN). In particular, we propose a perceptual loss function which motivated by perceptual similarity instead of similarity in pixel space. Moreover, we introduce the Residual Dense Normalization Block (RDNB) to replace the residual blocks as the basic network building unit. Finally, we borrow the idea from WGAN-GP as the adversarial loss functions. The ECycleGAN, thanks to these changes, demonstrates appealing visual quality with more realistic and natural textures than any state-of-the-art methods.

CCS Concepts
• Information systems→Information extraction; • Computing methodologies→Image representations.

Keywords
Computer vision; image generation; image translation; generative adversarial network.

1. INTRODUCTION
Image-to-image translation, as a fundamental low-level vision problem, has been a much studied task in the research community and AI companies. Image-to-image task aims to translate an image from one domain to another, e.g., image inpainting [10], [12], super resolution [6], [17], style transfer [7], [23], colorization [18], [30] and image to semantic segmentation [29]. The task is intuitively defined when we have paired examples of an image in each domain [1], [21], [22], but unfortunately obtaining paired training data can be difficult and expensive. Sometimes these are even not available in many interesting cases like artistic stylization.

ICCAI '20, April 23–26, 2020, Tianjin, China

ACM ISBN 978-1-4503-7708-9/20/04…$15.00
DOI: https://doi.org/10.1145/3404555.3404597

Passion develops with the development of the field, using an unsupervised method to match the distribution of the two domains with the generative adversarial networks (GANs) [2], [4], [16], [20]. However, there are countless mappings between two domains, and there is no guarantee that an individual image in one domain will share any features with the representation in the other domain after mapping.

Previous methods by regularizing the generators in various ways have addressed the consistency problem, including employing U-Net [8], cross-domain weight-coupling in some layers [16] and decoding from a shared embedding space [11]. CycleGAN [14] first introduces the most common regularization named cycle-consistency property and residual blocks which is shown in Figure 1 to force the generators. They are successful for the style transfer tasks mapping local texture (e.g.,photo2monet and photo2cezanne) but unsuccessful for image translation tasks with larger shape change(e.g., apple2orange and cat2dog) [24]. For the popular task horse2zebra [14], there is the problem of ambiguity.

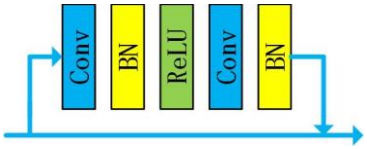

Figure 1. The construction of Residual Block, which can solve gradient vanishing and network degradation.

In this work, we propose a novel method for unsupervised image-to-image translation, which incorporates a new RDNB module and a VGG network [27] in an end-to-end manner. Our model guides the translation to force on instance details by distinguishing between source and target domains based on the high-level feature map obtained by the VGG network.

In recent years, with the development of deep learning, how to increase the depth of network effectively has gained a lot of attention from researchers in the fields of deep learning such as residual module [3] and dense module [25]. RDNB module, which is of higher capacity and easier to train, can increase generators complexity to handle the quality of generated images.

We use the features extracted from a pre-trained VGG19 network [27] instead of low-level pixel-wise error measures. Specifically we formulate a loss function based on the euclidean distance between feature maps extracted from the VGG19 network which is shown in Figure 2. As a result, our model can perform image translation tasks not only requiring style transfer but also generating high quality images.

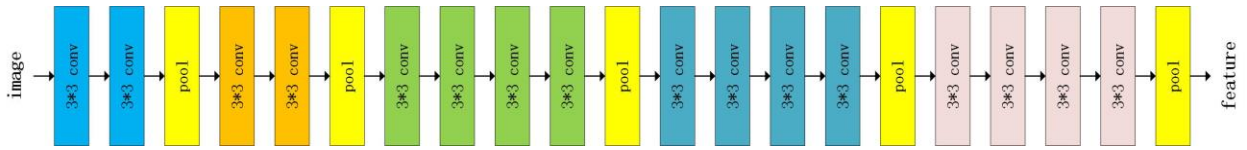

Figure 2. The overall architecture of VGG19 network. It contains sixteen convolutional layers and five max-pooling layers. High-dimensional feature are obtained after fifth max-pooling layer.

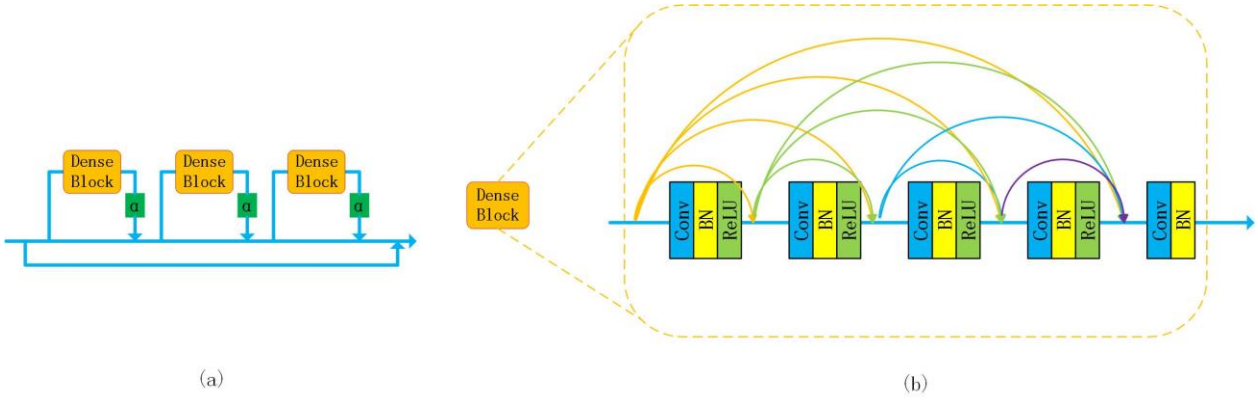

Figure 3. Overview of RDNB, consisting of two network modules, residual network and dense network. (a) Every RDNB block is made up of three dense blocks connected by residuals, α is residual scaling parameter. (b) Every dense block is made up of five convolutional layers densely connected.

The main contribution of the proposed work can be summarized as follows.

- We propose a novel method for unsupervised image-to-image translation with a new network module, RDNB, and a new regularization loss, VGG19 perceptual loss function.
- Our RDNB module, employing a deeper and more complex structure than the original residual block in CycleGAN, helps the model to improve the quality of generated images.
- We add a perceptual loss calculated on feature maps of the VGG19 network based on the L1 loss, which are more invariant to changes in pixel space.

2. RELATED WORK

2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [5] have shown impressive results in various tasks such as image translation, image generation [9], super-resolution imaging, and representation learning [13]. Conditional image generation based on GAN [15], such as generating particular images highly relevant to a given text description and image editing, has also been successfully applied. The only reason GAN has enjoyed such success thus far is that it consists of two modules: a generator and a discriminator. The generator learns to generate fake images we want, while the discriminator learns to distinguish between real and fake images. In principle, we adopt an adversarial loss to force the generated fake images to be indistinguishable from real images in the target domain.

2.2 Image-to-Image Translation

Recent works have yielded impressive results in image-to-image translation [1], [11], [14]. For example, pix2pix [1] learned the task in a supervised manner using cGAN [15]. It combines an adversarial losses [5] with L1 losses, so pairs of data samples are needed. BicycleGAN [28] extends it to multi-model translation. In order to alleviate the problem of data pair acquisition, unpaired image-to-image translation framework are proposed. UNIT [11] combines Variational Autoencoders (VAEs) [19] with CoGAN [16], a GAN framework in which two generators share weights to learn the joint distribution of images in cross domains. CycleGAN [14] and DiscoGAN [26] have exploited the loss of cycle consistency to preserve key attributes between the input and transformed images and do not rely on any task-specific, predefined similarity function between the input and output. However, the quality of the generated images sometimes is poor. Unlike the approach of CycleGAN, we use RDNB modules instead of residual modules to increase network capacity and a VGG perceptual loss to limit high-dimensional features.

2.3 Design of Neural Networks

The state of the art for many computer vision problems is meanwhile set by specifically designed CNN architectures following the success of the work by Krizhevsky et al. [31]. In order to increase the depth and solve the training problem in the deep neural networks, residual networks and dense networks are brought up. In image-to-image task, U-Net [8] and residual blocks are widely used to implement the feature extraction. In the paper, we combine the residual and dense networks to propose a based block, RDNB, while the residual scaling, multiplying a constant between 0 and 1 before adding residuals to the main path, is applied.

3. APPROACH

Our main aim is to improve the visual quality for image-to-image translation between source domains X and target domain Y. Our model, which is shown in Figure 4, includes two mapping function (G_{XY}: X-Y and G_{YX}: Y-X), two discriminators (D_X and D_Y) which to distinguish between original images and translated images, and a VGG19 network for perceptual losses. In the following we first describe our proposed network architecture RDNB, a critical module to replace residual block, and then discuss the adversarial losses for matching the distribution of

generated images to the data distribution in the target domain. At last, we describe the newly introduced perceptual losses which is to minimize the error in a feature space instead of pixel space.

3.1 Network Architecture

In order to further improve the translated image quality of CycleGAN, we mainly make a huge adjustment to the structure of generator G: replace the original basic Residual block with the proposed Residual Dense Normalization Block (RDNB), which consists of multi-level residual network, dense connection and instance normalization layers.

We keep the cycle consistency architecture design of CycleGAN, and use a novel basic block namely RDNB as depicted in Figure. 3. Because of the exploration that more layers and connections could always optimize performance of neural networks, the proposed RDNB possesses a more complex and deeper construction than the original residual block in CycleGAN. Noteworthily, there are different levels of residual in RDNB that is a residual-in-residual structure. To further benefit from the dense connection, we use dense block in the primary path. For preventing the instability on training a very deep network, we introduce the residual scaling to scale down the residuals by multiplying a constant (0 to 1) before adding residuals to the primary path.

In addition to the changed basic blocks, we also use the instance normalization (IN) instead of batch normalization (BN) which normalizes the features using mean and variance to ensure consistent data distribution in a batch. But in the image-to-image translation task, the generated result depends on a single image instance. So normalizing $H \times W$ in image pixel level is to accelerate model convergence and keep each image instance independent.

3.2 Adversarial Loss

Following Goodfellow et al[5]. we apply adversarial losses to both mapping functions in order to make the generated images indistinguishable from real images. For the forward mapping function $G_{XY}: X \to Y$ and its discriminator D_Y, we express the objective as:

$$L_{GAN}(G_{XY}, D_Y) = E_{y \sim P_{data}(y)}[log D_Y(y)] + E_{x \sim P_{data}(x)}[log(1 - D_Y(G_{XY}(x)))], \quad (1)$$

where G_{XY} aims to generate a fake image $G_{XY}(x)$ conditioned on the input image x that looks similar to images from domiain Y, while D_Y tries to distinguish between real and fake images. In this paper, G tries to minimize this objective against D that tries to maximize it, i.e. $min_G max_D L_{GAN}(G_{XY}, D_Y)$. We adopt a similar adversarial loss for the mapping backward function $G_{YX}: Y \to X$ and its discriminator D_X as well: $min_G max_D L_{GAN}(G_{YX}, D_X)$.

3.3 Perceptual Loss

The pixel-wise L1 loss is calculated as:

$$L(G_{XY}) = E_{x \sim P_{data}(x)}[||G_{YX}(G_{XY}(x)) - x||_1, \quad (2)$$

This is the most widely used optimization target for image-to-image translation on which many state-of-the-art approach rely. However, solutions of L1 optimization problems often lack high-frequency content which results in perceptually unsatisfying solutions with overly smooth textures.

In place of relying on pixel-wise losses we base on the ideas of 7]. and use a loss function to calculate the perceptual similarity. The perceptual loss function is euclidean distance between the high level abstract feature representation of a cycle reconstructed image $G_{YX}(G_{XY}(x))$ and the original image x. We use pre-trained 19 layer VGG network [27] which can separate the content and style abstract feature representation of an image to compute VGG loss. With $\varphi_{i,j}$ we indicate the feature map obtained by the j-th convolution (after activation) before the i-th maxpooling layer within the VGG19 network, which we consider given.

$$L_{i,j}^X = E_{i,j}[\varphi_{i,j}\left(G_{YX}\left(G_{XY}(x)\right)\right) - \varphi_{i,j}(x)]. \quad (3)$$

Here we introduce a similar perceptual loss $L_{i,j}^Y$ for the mapping reconstructed image $G_{XY}\left(G_{YX}(y)\right)$ and the original image y. So the total perceptual loss is:

$$L_{perc}(G_{XY}, G_{YX}) = \alpha(L_{i,j}^X + L_{i,j}^Y) + \beta(L(G_{XY}) + L(G_{YX})), \quad (4)$$

where α and β are the coefficients to balance different loss terms.

3.4 Full Objective

Our full objective is given by:

$$L(G, D) = L_{GAN}(G_{XY}, D_Y) + L_{GAN}(G_{YX}, D_X) + \lambda L_{perc}(G_{XY}, G_{YX}), \quad (5)$$

where λ is a hyper-parameters which controls the importance of perceptual losses. We aim to solve:

$$G^*, D^* = argmin_{G_{XY}, G_{YX}} max_{D_X, D_Y} L(G, D), \quad (6)$$

4. IMPLEMENTATION

4.1 WGAN-GP

To generate higher quality images and stabilize the training process, we use Wasserstein GAN objective with gradient penalty to replace adversarial defined as:

$$L_{GAN}(G_{XY}, D_Y) = E_{y \sim P_{data}(y)}[D_Y(y)] - E_{x \sim P_{data}(x)}\left[D_Y\left(G_{XY}(x)\right)\right] - \lambda_{gp} E_{\hat{x}}[(||\nabla_{\hat{x}} D_Y(\hat{x})||_2 - 1)^2]. \quad (7)$$

where $\hat{x}$ is generated by random interpolation sampling between a real and a generated image. In other words, $\hat{x} = \epsilon x + (1 - \epsilon) G_{XY}(x)$, ϵ is a random number between 0 and 1. We set $\lambda_{gp} = 10$ for all experiments.

4.2 Network Architecture

Adapted from Johnson et al[14]. ECycleGAN has the generator network composed of the two convolutional layers with the stride size of two for downsampling, four RDNB blocks, and two fractionally-stride convolutions with stride $\frac{1}{2}$. We use instance normalization for both generator and discriminator. For the discriminator networks we use 70×70 PatchGANs, which aim to distinguish whether image patches are real or fake.

5. EXPERIMENT

In this section, we first compare ECycleGAN against recent methods on image-to-image translation including UNIT, MUNIT and DRIT that we implement all the baseline methods using author's code. At the same time, we only change the basic blocks in CycleGAN as a contrast. In my experiments, we use Adam with β_1=0.5 and β_2=0.999 and set the batch size to one to train the model. The learning rate is 0.0001 in the first 50 epoches and linearly decayed up to 100 epoches.

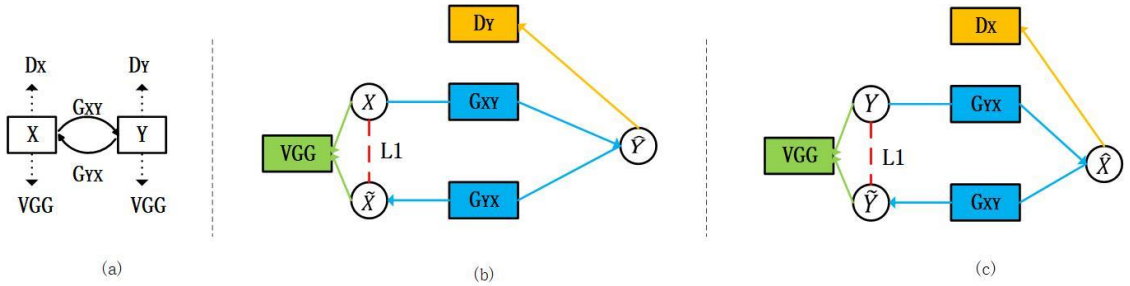

Figure 4. Our model contains two mapping function $G_{XY}: X \rightarrow Y$ and $G_{YX}: Y \rightarrow X$, and corresponding discriminators D_X and D_Y, and a VGG network with fixed parameters. D_X encourages G_{YX} to generated indistinguishable images from domain Y, and appropriate for D_Y and G_{XY}. VGG network extract high dimensional features from X and reconstructed $\tilde{X}$ to calculate the perceptual loss, and vice versa for Y and reconstructed $\tilde{Y}$ (b) forward expanded view: $X \rightarrow G_{XY}(X) \rightarrow G_{YX}(G_{XY}(X)) \rightarrow \tilde{X}$ and (c) backward expanded view: $Y \rightarrow G_{YX}(Y) \rightarrow G_{XX}(G_{YX}(Y)) \rightarrow \tilde{Y}$.

Table 1. The summary statistics of the two mainstream datasets

Datasets	Class	Train	Test
horse2zebra	horse	1067	120
	zebra	1334	140
photo2vangogh	photo	6287	751
	vangogh	1811	400

5.1 Datasets

We evaluate our models on two different datasets which are used in CycleGAN. The detailed statistics about the datasets are listed in Table 1. Each dataset is briefly described as follows.

horse2zebra[1]: It contains two domains which are horse and zebra domains selected from ImageNet. All images are resized to 256×256 .

photo2vangogh[1]: Photo domain comes from the real world and vangogh domain is an excerpt from van gogh's paintings. The size are also 256×256.

5.2 Comparison Methods

We compare our models with several state-of-the-art baseline methods:

CycleGAN: Using an adversarial loss and cyclic consistence loss to learn the mapping between two different domains X and Y.

UNIT: Assuming a shared-latent space to tackle unsupervised image translation.

MUNIT: Using adaptive instance normalization to synthesize the separated style and content to translate the source images.

DRIT: Decomposing the image into content and style, while it uses weight share and the content discriminator (auxiliary classifier) share the content spaces between the two domains.

In order to investigate the effectiveness of Perceptual loss, we also experiment with two variants of our model. One is a simplified variant with only changing the basic blocks. For another variant, it adds the VGG19 network to calculate the perceptual loss.

5.3 Experimental Results

In our experiment, we analysis the results from two sets of data in two ways, qualitative evaluation and quantitative evaluation. In qualitative evaluation, we will display the generated results compared with baseline model. And for quantitative evaluation, we use the recently proposed KID, which computes the squared Maximum Mean Discrepancy between the feature representations of real and generated images. The feature representation are extracted from the Inception network.

Qualitative Evaluation. ECycleGAN clearly generates the most natural-looking expressions in two datasets which are shown in Figure 5 and 6. In horse2zebra dataset, UNIT and MUNIT fails to pressure the personal identity in the translated images. While the CycleGAN and DRIT mostly preserve identity of the input, many generated results do not maintain the sharpness of images and become blurry. Moreover, the region around heads of two zebras are out of shape. Note that the CycleGAN with RDNB can generate relatively real-world images. ECycleGAN can translate generate undistorted image by using perceptual loss between source and target domain.

In photo2vangogh dataset, while preserving the semantic characteristics of the source domain, the generated results of ECycleGAN are visually superior to other methods. It is worth nothing that the results of MNUIT and DRIT are very different from the source images because the images they generate have random style codes to ensure diversity. In the contrast of ECycleGAN, CycleGAN and UNIT can not generate attractive results.

Table 2. Kernel inception distance$\times$ 100 $\pm$ STD.$\times$ 100 for difference image translation mode. Lower is better

Model	horse2 zebra	zebra2 horse	Photo2 vangogh	Vangog h2phot o
ECycleG AN	7.11±0.91	7.50±0.72	4.55±0.32	5.66±0. 33
CycleGA N- RDNB	7.58±0.71	8.82±0.65	6.01±0.33	5.85±0. 31
CycleGA N	8.05±0.72	8.91±0.63	6.25±0.32	5.88±0. 35
UNIT	10.44±0.67	14.93±0.75	5.26±0.29	9.72±0. 33
MUNIT	11.41±0.82	16.47±0.99	13.08±0.35	9.53±0. 35
DRIT	9.79±0.61	10.99±0.56	12.65±0.35	7.73±0. 34

Quantitative Evaluation. Compared with Frechet Inception Distance (FID), which is used to evaluate generated images quality, KID is an unbiased estimator, which makes it more

[1]https://people.eecs.berkeley.edu/~taesung_park/CycleGAN/datasets/.

reliable. The lower KID value is, the more similar real and translated images are. Therefore, the value of KID will be small if models can translate the images well. Table 2 shows that ECycleGAN achieved the lowest KID scores in horse2zebra and photo2vangogh datasets. However, we can see that there is no big difference from the lowest value. At the same time, translating between the two regions is stable. Numerically, replacing the residual blocks with RDNB blocks is useful to some extent, but the key part is the perceptual loss.

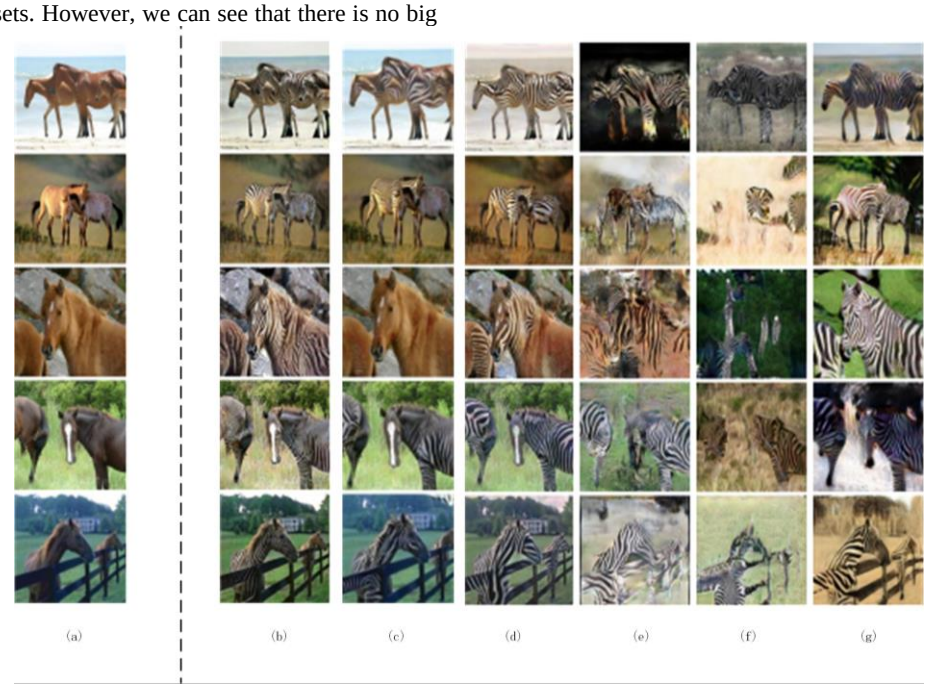

Figure 5. Different methods for mapping horse to zebra. (a) Source images, (b) ECycleGAN, (c) CycleGAN with RDNB, (d) CycleGAN, (e) UNIT, (f) MUNIT, (g) DRIT.

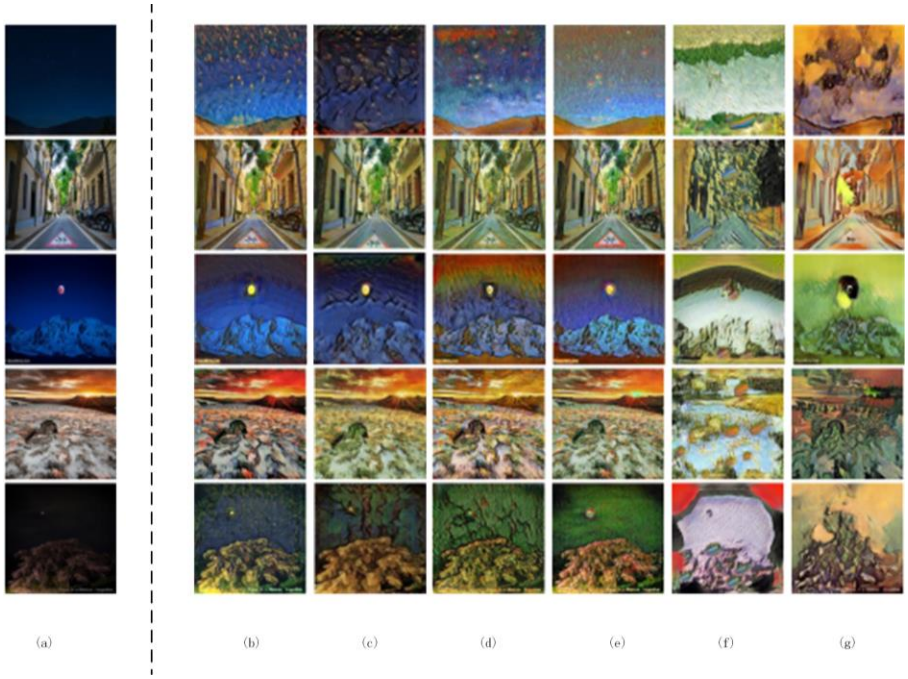

Figure 6. Different methods for mapping photo to vangogh. (a) Source images, (b) ECycleGAN, (c) CycleGAN with RDNB, (d) CycleGAN, (e) UNIT, (f) MUNIT, (g) DRIT.

6. CONCLUSION

In this paper, we propose a novel enhance cycle-consistent generative adversarial network (ECycleGAN) for image-to-image translation. With an adversarial loss constraint, the generator are focus on translating an image from source to target domain. The perceptual loss encourages the translated images more realistic in detail, while the introduction of RDNB block promotes the model to generate high-quality images. Experimental results demonstrate that our model achieves superior and comparable results.

7. REFERENCES

[1] Isola, P., Zhu, J. Y., and Zhou, T.2017. Image-to-image translation with conditional adversarial network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125--1134.

[2] Russo, P., Carlucci, F. M., and Tommasi. T. 2018. From source to target and back: symmetric bi-directional adaptive gan. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8099--8108.

[3] He, K., Zhang, X., and Ren, S. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770--778.

[4] Taigman, Y., Polyak, A., and Wolf, L. 2016. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200.*

[5] Goodfellow, I., Pouget-Abadie, J., and Mirza, M. 2014. Generative adversarial nets[C]//Advances in neural information processing systems, 2672--2680.

[6] Kim, J., Kwon, L. J., and Mu, L. K. 2016. Accurate image super-resolution using very deep convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1646--1654.

[7] Gatys, L. A., Ecker, A. S., and Bethge, M. 2016. Image style transfer using convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414--2423.

[8] Ronneberger, O., Fischer, P., and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer, Cham, 234--241.

[9] Denton, E. L., Chintala, S., and Fergus R. 2015. Deep generative image models using a laplacian pyramid of adversarial networks[C]//Advances in neural information processing systems, 1486--1494.

[10] Pathak, D., Krahenbuhl, P., and Donahue, J. 2016. Context encoders: Feature learning by inpainting[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2536--2544.

[11] Liu, M. Y., Breuel, T., and Kautz, J. 2017. Unsupervised image-to-image translation. Advances in neural information processing systems, 700--708.

[12] Iizuka, S., Simo-Serra, E., and Ishikawa, H. 2017. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4): 1--14.

[13] Radford, A., Metz, L., and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[14] Zhu, J. Y., Park, T., and Isola, P. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE international conference on computer vision*, 2223--2232.

[15] Goodfellow, I., Pouget-Abadie, and J., Mirza, M. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 2672--2680.

[16] Liu, M. Y., and Tuzel, O. 2016. Coupled generative adversarial networks. *Advances in neural information processing systems*, 469--477.

[17] Dong, C., Loy, C. C., and He, K. 2015. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2): 295--307.

[18] Zhang, R., Zhu, J. Y., and Isola P. 2017. Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999.*

[19] Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[20] Dumoulin, V., Belghazi, I., and Poole, B. 2016. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*.

[21] Wang, T. C., Liu, M. Y., and Zhu, J. Y. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798--8807.

[22] Li, C., Liu, H., and Chen, C. 2017. Alice: Towards understanding adversarial learning for joint distribution matching. *Advances in Neural Information Processing Systems*, 5495--5503.

[23] Huang, X., and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. *Proceedings of the IEEE International Conference on Computer Vision*, 1501--1510.

[24] Lee, H. Y., Tseng, H. Y., and Huang, J. B. 2018. Diverse image-to-image translation via disentangled representations. *Proceedings of the European conference on computer vision (ECCV)*, 35--51.

[25] Huang, G., Liu, Z., and Van Der Maaten, L. 2017. Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700--4708.

[26] Kim, T., Cha, M., and Kim, H. 2017. Learning to discover cross-domain relations with generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org*, 1857--1865.

[27] Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[28] Zhu, J. Y., Zhang, R., and Pathak, D. 2017. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 465--476.

[29] Cordts, M., Omran, M., and Ramos, S. 2016. The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213--3223.

[30] Zhang, R., Isola, P., and Efros, A. A. 2016. Colorful image colorization. *European conference on computer vision. Springer, Cham*, 649--666.

[31] Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097--1105.