Unsupervised Generative Adversarial Networks with Cross-model Weight Transfer Mechanism for Image-to-image Translation

Xuguang Lai, Xiuxiu Bai*, and Yongqiang Hao
School of Software Engineering, Xi'an Jiaotong University
lxg0387@stu.xjtu.edu.cn, xiubai@xjtu.edu.cn, hyq123@stu.xjtu.edu.cn

Abstract

Image-to-image translation covers a variety of application scenarios in reality, and is one of the key research directions in computer vision. However, due to the defects of GAN, current translation frameworks may encounter model collapse and low quality of generated images. To solve the above problems, this paper proposes a new model CWT-GAN, which introduces the cross-model weight transfer mechanism. The discriminator of CWT-GAN has the same encoding module structure as the generator's. In the training process, the discriminator will transmit the weight of its encoding module to the generator in a certain proportion after each weight update. CWT-GAN can generate diverse and higher-quality images with the aid of the weight transfer mechanism, since features learned by discriminator tend to be more expressive than those learned by generator trained via maximum likelihood. Extensive experiments demonstrate that our CWT-GAN performs better than the state-of-the-art methods in a single translation direction for several datasets.

1. Introduction

Image-to-image translation transforms images from one domain to another while keeping their content unchanged, which has received extensive attention from academia. As shown in Figure 1, many tasks in computer vision can be regarded as image-to-image translation problems, such as image restoration [20], colorization [24, 3], super-resolution [14, 4, 11] and style transfer [5].

Currently, image-to-image translation is mainly implemented on the basis of GAN models. CycleGAN [26] provides a classic modeling idea of image-to-image translation by two groups of GANs. Later models [16, 22] are improved on this basis. With the development of deep learning, the results of image-to-image translation have been continuously improved.

*Corresponding author

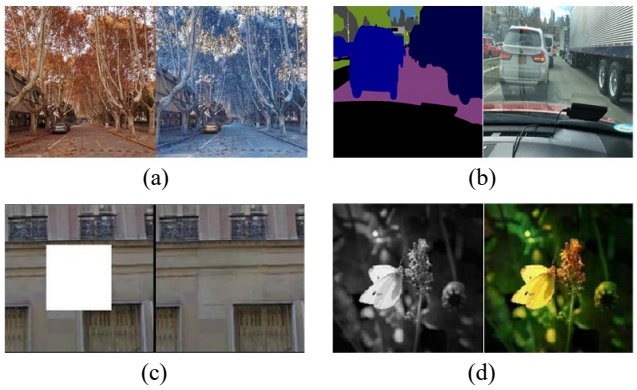

Figure 1. Examples of image-to-image translation. (a) Style transfer. (b) Semantic segmentation. (c) Restoration. (d) Colorization.

Existing GAN-based translation models [10, 15, 26] have achieved good results, but as shown in Figure 2, they still suffer two common problems: 1) Due to the defects of GAN and the imperfection of their loss function, model collapse and low diversity of results may occur; 2) The quality of the images generated by the unsupervised models using unpaired datasets still has much room for improvement.

To solve the above problems, we propose an unsupervised image-to-image translation network CWT-GAN using cross-model weight transfer mechanism. The model contains two GAN-based networks to output images from different domains and distinguish these images separately. The discriminators of CWT-GAN will transfer the weight of their encoders to those of generators in a certain proportion after each weight update phase. Since the encoders performing discriminative tasks can learn more expressive features for inference than the encoders trained via maximum likelihood, CWT-GAN can generate more diverse and higher-quality images with the help of our cross-model weight transfer mechanism.

In addition, most of studies [26, 10, 15] have revolved around the local texture translation of images, but in the case of significant shape differences between the source do-

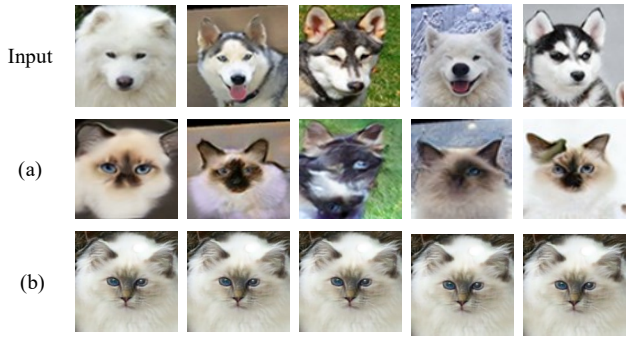

Figure 2. Problems in image-to-image translation. (a) The quality of the generated images is poor. (b) Mode collapse.

main and the target domain, the performance of these networks is unsatisfactory. To solve the above problems, CWT-GAN introduces a residual attention mechanism to further promote the spread of features in the encoders. This mechanism is implemented by using class activation maps (CAM) [25] under global pooling and average pooling, which is helpful for judging the consistency of semantic information under unsupervised conditions, so that the output images have better generation details in important areas. Meanwhile, the residual connection is introduced based on CAM, and the trade-off between the attention feature and the original feature is determined by updating the learnable parameter γ during the training process.

Our contributions can be summarized as follows.

- We propose an unsupervised image transformation network named CWT-GAN and the cross-model weight transfer mechanism, to make full use of the advantages of the encoder in the GAN discriminator.
- Our method achieves state-of-the-art results in a single translation direction for several datasets.

2. Related work

In recent years, image-to-image translation has been one of the main application scenarios of Generative Adversarial Nets (GANs) [6]. Due to the difficulty of obtaining paired data, the unsupervised method is more popular.

CycleGAN [26] consists of two sets of generators and discriminators, which can learn the two-way mapping between the source domain and the target domain. On the basis of CycleGAN, AttentionGAN [22] creates an attention mask through the built-in attention mechanism, and then fuse the generated result with the attention mask to obtain a high-quality target image. Considering that the above methods cannot distinguish the distribution of generated images and target domain images, DA-GAN [16] trains an attention encoder in a joint manner, and obtains object-level relevance through object pairs embedded with attention information, thereby constraining the whole sample and the object at the same time.

To solve significant shape changes in multiple target instances, InstaGAN [19] combines instance information to improve generation results, and proposes a context preservation loss to learn identity information outside the target instance. TransGaGa [23] introduces a novel decomposition-translation framework for the significant deformation problem. This method does not directly learn the mapping on the image space, but decomposes the image space into the Cartesian product of the appearance hidden space and the geometric hidden space.

In addition, U-GAT-IT [12] introduces an attention mechanism to better realize overall translation and large shape changes. Both the generator and discriminator of U-GAT-IT use an auxiliary classifier to obtain the attention image through the class activation mapping module, so that the generator can generate more reasonable details in the output image. NICE-GAN [2] uses the residual connection mechanism to furtherly improve the attention mechanism, through a trainable parameter to control the ratio of attention feature map and original feature map.

However, model collapse and low diversity of results may occur, since the defects of GAN. The quality of the images generated by the unsupervised models using unpaired data still has much room for improvement.

3. Method

In this section, we first introduce the general idea, and then provide the details of each component in CWT-GAN, including generators and discriminators.

3.1. Model structure and loss functions

Given unpaired samples $\{a_i\}_{i=1}^{N} \in A$, $\{b_j\}_{j=1}^{N} \in B$ from two domains, our purpose is to train an unsupervised image-to-image translation network to realize the mutual translation of samples between domain *A* and *B*. In Figure 3, CWT-GAN consists of two GAN-based networks: one is used to translate the source domain sample a into the target domain and the other is the opposite. The generator G_A is composed of an encoder E_A^G, a hidden layer module H_A and a decoder De_A, used to translate the images a from the domain *A* into $G_A(a)$. The same is true for the generator G_B. And the discriminators D_A and D_B are respectively used to identify whether the input images from their respective domains are real images.

Figure 3 shows the structure of CWT-GAN. Take domain *A* as an example, the input images of G_A and D_A are both from *A*, so E_A^G and E_A^D both perform extracting

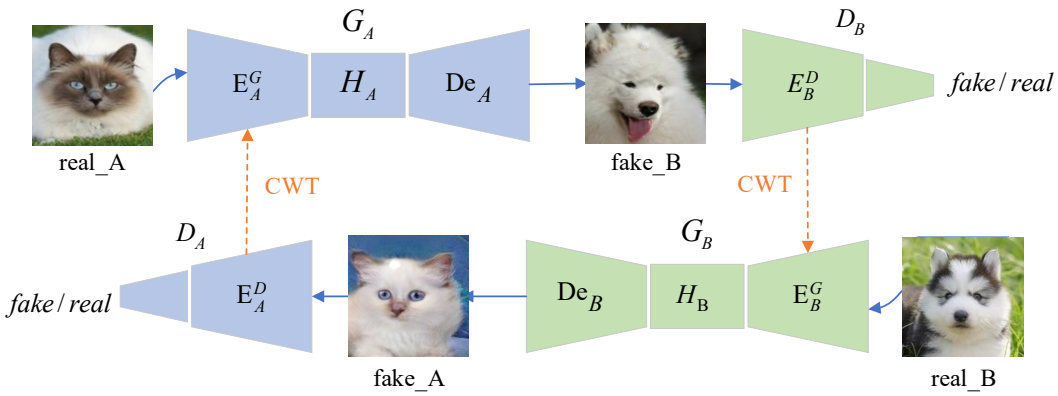

Figure 3. Illustration of the cross-model weight transfer mechanism. In CWT-GAN, the encoder of the generator and the encoder of the discriminator have exactly the same structure. During training, the discriminator will transfer the weight of its encoder to the generator in a certain proportion after each weight update phase.

features from A. This provides conditions for using cross-model weight transfer mechanism (CWT) between the generators and the discriminators. During the training process, the two discriminators will respectively transfer the weights of E_A^D and E_B^D to E_A^G and E_B^G in a certain proportion after each weight update phase. E_A^D and E_B^D, as the down-sampling modules of the discriminators, are used for the feature extraction sub-task. Therefore, while retaining the constraints of reconstruction loss and cycle consistency loss on the generators, the weight transfer mechanism of CWT-GAN uses E_A^D and E_B^D for processing discriminant tasks to improve the perception and inference capabilities of E_A^G and E_B^G.

Figure 4 shows the structure of generator in CWT-GAN. G_A can be divided into three modules: encoder E_A^G, hidden layer module H_A and decoder De_A. E_A^G is mainly composed of serial Convolution-Spectral Norm-Leaky ReLU layers, which encode the input images into high-dimensional feature maps. Among them, the Spectral Norm [18] prompts E_A^G to obey Lipschitz continuity, which limits the severity of function changes, thereby avoiding the model collapse of G_A. H_A is composed of six residual convolution blocks [7], which abstracts the features at multiple levels to accurately divide different types of data linearly. Finally, De_A contains two sub-pixel convolutional layers [21] for up-sampling.

The loss functions used by the generator include adversarial loss, cycle consistency loss and reconstruction loss. The adversarial loss can promote the matching of the distribution of the generated images with the distribution of target domain images. CWT-GAN uses the least-square adversarial loss [17] to achieve more stable training and higher

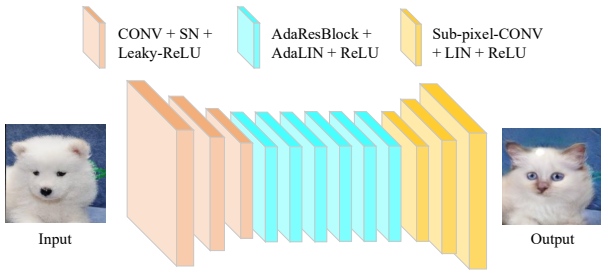

Figure 4. The structure of generator in CWT-GAN. The network can be divided into three modules: encoder, hidden layer module and decoder.

generation quality, as shown in Eq.(1):

$$\begin{aligned}\min_{G_{b\to a}} L_{gan}^{b\to a} &= \mathrm{E}_{a\sim A}\left[(D_A(a))^2\right] \\ &+ \mathrm{E}_{b\sim B}\left[1 - D_A\left(G_B(b)\right)^2\right]\end{aligned} \tag{1}$$

CWT-GAN uses the cycle consistency loss function to reduce the probability of model collapse. Given an image $a \in A$, after being transformed from A into B, and then back to A, the generated image should be in the same distribution as a. The cycle consistency loss function is shown in Eq.(2):

$$L_{cycle}^{a\to b} = \mathrm{E}_{a\sim A}\left[|a - G_B\left(G_A(a)\right)|_1\right] \tag{2}$$

To ensure the output and the input maintain a similar color distribution, CWT-GAN also applies a reconstruction consistency constraint to the generator. That is, given an image $a \in A$, after using G_B to transform the image a, the output image should not be changed. The reconstruction loss function can assist the generator G_B to extract hierarchical features, and reduce the error caused by the D_A in the

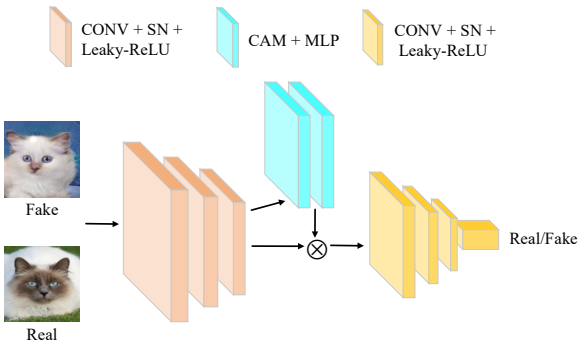

Figure 5. The structure of discriminator in CWT-GAN. The network consists of three modules: encoder, classifier and auxiliary classifier.

feature extraction process. The reconstruction loss function is shown in Eq.(3):

$$L_{identity}^{b\to a} = \mathrm{E}_{a\sim A}\left[|a - G_B(a)|_1\right] \tag{3}$$

Inspired by the work of using affine transformation parameters in the normalization layer and combining the normalization function [9], Junho *et al.* [12] combine AdaIN [9] with LN [1], and propose AdaLIN normalization method. CWT-GAN introduces this method that combines the advantages of AdaIN and LN by selectively retaining or changing content information and solves a wide range of image-to-image translation problems.

Figure 5 shows the structure of discriminator in CWT-GAN. For the discrimination task, CWT-GAN uses two discriminators to determine whether the input images are real data or generated by generators. As shown in Figure 5, D_A consists of three modules: encoder E_A^D, classifier C_A^D and auxiliary classifier η_A^D. E_A^D is composed of several Convolution-Spectral Norm-Leaky ReLU layers and used to encode the input image into high-dimensional features. C_A^D only contains a layer of convolutional neural network and outputs the judgment result of the input image. η_A^D adopts the design in NICE-GAN [2], and measures the importance of each feature map generated by E_A^D by learning an attention vector ω. Then, the feature map containing residual attention mechanism is obtained by the equation $f(a) = \gamma \times \omega \times E_a(a) + E_a(a)$, in which the trainable parameter γ determines the trade-off between the attention feature and the original feature.

The loss function used by the discriminator is the adversarial loss. To promote the matching of the distribution of generated images with the distribution of images from target domain, CWT-GAN optimizes parameters of the discriminator when maximizing the adversarial loss, as shown in Eq.(4):

$$\begin{aligned}\max_{D_a} L_{gan}^{b\to a} = &\mathrm{E}_{a\sim A}\left[(D_A(a))^2\right] \\ &+ \mathrm{E}_{b\sim B}\left[1 - D_A\left(G_B(b)\right)^2\right]\end{aligned} \tag{4}$$

E_A^D adopts the same network structure as the generator encoder E_A^G. In the training process, CWT-GAN first updates weights of the discriminator network by maximizing the adversarial loss $L_{gan}^{b\to a}$, then passes weights of E_A^D to E_A^G according to the proportional parameter α, and finally updates weights of the generator network by minimizing the adversarial loss, cycle consistency loss and reconstruction loss. The weight transfer mechanism is shown as Eq.(5):

$$w_E^G = \alpha \cdot w_E^D + (1-\alpha)\cdot w_E^G \tag{5}$$

where w_E^G represents weights of the encoder in generator, w_E^D represents weights of the encoder in decoder, and α is used to determine the proportion of weight transfer.

Compared with E_A^G that performs the generation task, E_A^D learns features that are more expressive and more suitable for inference. Therefore, with the help of the weight transfer mechanism, while maintaining the constraints of reconstruction loss and cycle consistency loss, E_A^G can promote its feature perception and inference capabilities through E_A^D, so that the generator can output diverse and higher-quality generated images.

3.2. Total loss

The generator of CWT-GAN is trained through a joint loss function, as shown in Eq.(6):

$$\min_G \mathrm{L}_G = \lambda_1 \cdot L_{gan} + \lambda_2 \cdot L_{cycle} + \lambda_3 \cdot L_{identity} \tag{6}$$

where λ_1, λ_2 and λ_3 are all hyper-parameters that control the importance of each loss.

Under the premise of applying the weight transfer mechanism, the discriminator of CWT-GAN only needs to be configured with the adversarial loss, which reduces the coupling and training cost of the model. The formula is shown in Eq.(7):

$$\max_{\mathrm{D}} L_D = L_{gan}^{b\to a} + L_{gan}^{a\to b} \tag{7}$$

During training time, CWT-GAN will follow the process cycle of training the discriminator, encoder weight transfer and training generator to achieve the optimization of network weights. During testing time, as long as the test image is input to the corresponding generator, the generated image can be output.

4. Experiments

This section presents the detailed experiments of CWT-GAN and other advanced models. We first introduce

datasets, comparison methods and evaluation metrics, then provide comparison results, and lastly show the ablation studies.

4.1. Datasets

To verify the effectiveness of CWT-GAN, we select three popular unpaired datasets, including horse↔zebra, summer↔winter, and cat↔dog. The first two datasets are used in CycleGAN [26], and the last is studied in DRIT [15]. All images are resized to 256×256 for training and testing.

4.2. Comparison methods

To demonstrate the effectiveness of CWT-GAN, we select several representative unsupervised image-to-image translation models in the experiment, including CycleGAN[26], MUNIT[10], DRIT[15], U-GAT-IT [12] and NICE-GAN [2]. All methods being compared are performed through using public codes.

4.3. Evaluation metrics

In addition to the qualitative comparison of generated images obtained by different models through manual observation, this paper uses the Fréchet Inception Distance (FID) [8] to quantitatively evaluate the performance. FID compares the statistical data of the generated sample with the real sample. For each image set to be compared, FID obtains the features extracted from the hidden layer of the network after they are input to InceptionNet, and then calculates the Fréchet distance between the distributions through the Gaussian distribution of these features. Lower FID is better, corresponding to generated images more similar to the real.

4.4. Implementation details

The total number of iterations of CWT-GAN and other networks is 100K. We set the weight of adversarial loss, cycle consistency loss and reconstruction loss to 1, 10, and 10 respectively. The transfer ratio of our weight transfer mechanism is set to 0.9, and we use Adam [13] as the optimization algorithm with the learning rate 0.0001 and $(\beta_1, \beta_2) = (0.5, 0.999)$. The encoders of the generators and discriminators all use the Leaky-ReLU activation function with a negative slope of 0.2. The batch size is set to 1. Our code is available at https://github.com/lxg0387/CWT-GAN.

4.5. Comparisons with SOTA

To prove the validity of CWT-GAN, we use CWT-GAN and other advanced networks to process translation tasks on a variety of datasets, and compare the FID scores.

Table 1 shows the FID scores obtained by different models in the cat↔dog task. As we can see, CWT-GAN achieves the best FID score in the translation direction from dogs to cats. And in the opposite direction, its FID score is second only to NICE-GAN's, and the difference between the two is not significant. This shows that CWT-GAN has a good translation ability on this task. In contrast, although U-GAT-IT, MUNIT and DRIT can successfully transform the semantics of the object, the translation effect is significantly worse than our model. The FID score obtained by CycleGAN is far worse than that of all other models. The possible reason is that CycleGAN is better at transforming low-level features, such as color and texture, while the cat↔dog task involves the problem of large deformation of the object. Therefore, it is difficult for CycleGAN to generate effective output images. CWT-GAN has a similar structure to U-GAT-IT, such as attention mechanism and multi-scale discriminator, while the main difference between them is that CWT-GAN applies the cross-model weight transfer mechanism designed in this paper.

Model	FID (Dog2cat)	FID (Cat2dog)
CWT-GAN (our)	**43.77**	46.29
NICE-GAN [2]	48.79	**44.67**
U-GAT-IT-light [12]	80.75	64.36
CycleGAN [26]	119.32	125.30
MUNIT [10]	53.25	60.84
DRIT [15]	94.50	79.57

Table 1. FID scores of different models in the cat↔dog dataset. Lower is better.

Figure 6 shows the effect of CWT-GAN and other models in generating images in cat↔dog translation task. As the images generated by CWT-GAN show, after the source domain object is transformed into the target domain, the details of the object's face, ears, etc. have produced reasonable changes in color, shape and other characteristics, indicating that CWT-GAN has a good performance in transforming low-level features and significantly deformed objects. Observing the images generated by other transformation networks, we can see that results generated by NICE-GAN and MUNIT are relatively clear, but some of generated images have ambiguous expressions at the junction of target and background. And for generated samples of CWT-GAN, it can be found that the connection between the object and the background is more natural, which increases the difficulty of artificially distinguishing the authenticity of the samples. In addition, the generation results of CycleGAN fulfill the above analysis of using it to handle the translation task of significant deformation. Due to the large shape difference between cats and dogs, CycleGAN only achieves color and texture transfer during the translation process, but it performs poorly in the translation of the target shape. It can be clearly seen from the samples generated by CycleGAN that the facial features of some objects are incomplete, misplaced, or redundant.

Figure 6. Generated results of cat↔dog translation task. As the images generated by CWT-GAN shown, the details of the object's face, ears, etc. have produced reasonable changes in color, shape and other characteristics. However, some results generated by NICE-GAN and MUNIT have ambiguous expressions at the junction of target and background. Even worse, some of CycleGAN's generated images have completely collapsed.

In Table 2, CWT-GAN is better than all other models in the translation direction from winter images to summer images. CWT-GAN is second only to NICE-GAN in the opposite direction and the FID score gap is only about 0.5. Compared with the cat↔dog task, the key of seasonal translation task is the change of image color and texture, and significant deformation phenomenon rarely occurs. Therefore, by comparing Table 1 and Table 2, it can be found that CycleGAN has a significant improvement in processing seasonal transformation tasks, and is even better than U-GAT-IT-light. The above phenomenon proves again that CycleGAN is good at transforming low-level features between the two domains but is weak in dealing with significant deformation problems. Since CWT-GAN uses cross-model weight transfer mechanism, while the constraints of reconstruction loss and cycle consistency loss on the generator are retained, the discriminator encoder for processing classification tasks improves the feature perception and inference capabilities of the generators, so that the model can output effective results in a variety of image-to-image translation tasks.

Table 3 shows the FID scores obtained by CWT-GAN and other models in processing the transformation task between horse and zebra. According to the data in the table, CWT-GAN leads all other models in the transformation direction from zebra images to horse images, and NICE-GAN has the lowest FID score in the transformation direction from horse images to zebra images. Figure 7 shows the partial generation results of CWT-GAN in seasonal transformation and horse↔zebra transformation.

Model	FID (Winter2summer)	FID (Summer2winter)
CWT-GAN (our)	76.99	**74.92**
NICE-GAN [2]	**76.44**	76.03
U-GAT-IT-light [12]	80.33	88.41
CycleGAN [26]	79.58	78.76
MUNIT [10]	99.14	114.08
DRIT [15]	78.61	81.64

Table 2. FID scores obtained by different models for processing seasonal transformation tasks. Lower is better.

Through the above three sets of experiments, it verifies that CWT-GAN can achieve the best level in at least one direction in a variety of translation tasks, and in the other direction, it is second only to the current state-of-the-art

Model	FID (Zebra2horse)	FID (Horse2zebra)
CWT-GAN (our)	**142.27**	85.44
NICE-GAN [2]	149.48	**65.93**
U-GAT-IT-light [12]	145.47	113.44
CycleGAN [26]	156.19	95.98
MUNIT [10]	193.43	128.70
DRIT [15]	200.41	116.63

Table 3. FID scores of different models for the transformation task between horse and zebra. Lower is better.

Figure 7. Generated results of CWT-GAN in seasonal transformation and horse↔zebra transformation.

model NICE-GAN.

Why CWT-GAN can generate high-quality images to apply our cross-model weight transfer mechanism? The reason is that the discriminator will transfer its encoder weight to the generator according to the hyperparameter α after updating the weights. Since the encoder that performs the discriminative task can learn more expressive and more suitable features for inference than the encoder trained using the maximum probability method, CWT-GAN can use this mechanism to generate more diverse and higher-quality target domain images.

4.6. Ablation studies

To analyze the impact of key technologies used in CWT-GAN, we conduct ablation experiments on these technologies on the cat↔dog task. The key technologies investigated in the experiment include our Cross-model Weight Transfer (CWT) mechanism and the Residual Attention (RA) mechanism [2] introduced in the discriminator.

Table 4 shows the ablation results. It can be seen that CWT and RA both improve the transformation performance of the model. By combining all the key components, CWT-GAN's image transformation performance is better than other variants. The reason why the CWT mechanism is effective may be that the mechanism can shorten the translation distance of latent space between different domains, so that CWT-GAN based on the assumption of shared latent space can better realize domain translation in image space.

Model	FID (Dog2cat)	FID (Cat2dog)
Baseline	74.34	76.51
Baseline + CWT	52.02	63.91
Baseline + RA	48.70	50.35
Baseline + CWT + RA (our)	**43.77**	**46.29**

Table 4. CWT-GAN ablation experiment on the cat↔dog dataset. Lower is better.

Figure 8 shows qualitative results of the above ablation experiment. Due to the lack of cross-model weight transfer mechanism and residual attention mechanism, the basic model has a greater probability of ignoring the edge details of the object when performing translation processing and is not able to effectively deal with the significant deformation of the object during the translation. Consequently, it collapses in the translation of some samples. When the basic model only uses the residual attention mechanism, the generated image does not achieve a sufficiently natural transition at the intersection of different coat colors of the target, and some features of the source domain object still remain in the target's facial features. When the basic model is only equipped with the weight transfer mechanism, there is still a lot of room for improvement in the degree of completion of the content transformation of the generated image.

Finally, our CWT-GAN, which is configured with the above two key components at the same time, can better pay attention to the edge details of the object when performing image transformation, and benefits from the strong feature perception and inference capabilities of the discriminator encoder. Its generator can use a shorter translation distance between different domains, so that it can output high-quality generated images even when solving tasks with significant deformation. In Figure 8, when CWT-GAN configures all the key modules, it generates images with the highest definition and processes the target edge most perfectly.

In our cross-model weight transfer mechanism, the hyperparameter α controls the degree of influence of the discriminator encoding module on the generator encoder. To explore the optimal ratio of the weight transfer, this paper uses CWT-GAN to perform seasonal transformation tasks under a variety of α settings.

Table 5 shows experiment results of α setting. When α is set to 0.9, the image transformation effect of model is better than other values. When α is set to 0, that is, when the model does not apply the influence of the discriminator encoding module on the generator encoder, the transformation effect of the model is reduced. When α is set to 1, that is, when the discriminator encoding module completely re-

(a) (b) (c) (d) (e)

Figure 8. Results of ablation experiment. (a) Input. (b) Baseline. (c) Baseline + CWT. (d) Baseline + RA. (e) Baseline + CWT + RA.

α setting	FID (Winter2summer)	FID (Summer2winter)
0	82.33	76.51
0.6	84.36	78.70
0.7	82.91	75.44
0.8	80.03	75.28
0.9	**76.99**	**74.92**
1.0	80.61	75.22

Table 5. FID scores of CWT-GAN performing seasonal transformation tasks under different α settings. α is used to determine the proportion of weight transfer.

places the generator encoder, the model cannot achieve the best image-to-image translation effect because the encoding modules lacks the constraints of cycle consistency loss and reconstruction loss.

It can be seen from the above phenomenon that the coding module that performs the discrimination task can promote the feature extraction and inference ability of the generator coding module. The cycle consistency loss and reconstruction loss used in the generation task are also crucial to the training of the generator coding module. Therefore, our cross-model weight transfer mechanism combines the weight transfer of the discriminator encoding module with the constraints of cycle consistency loss and reconstruction loss, so that the CWT-GAN generator achieves the relatively best image transformation performance.

5. Conclusion

This paper proposes a new image-to-image translation network CWT-GAN that performs weight transfer between the generators and the discriminators, aiming at the problems of low diversity and insufficient quality of generated images in current GAN-based image transformation tasks. Compared with the encoder trained using the maximum likelihood method, the features learned by the model performing the discriminative task are more expressive and more suitable for inference. Therefore, CWT-GAN can generate diverse and high-quality generated images with the help of our weight transfer mechanism. We compare CWT-GAN with other advanced translation methods on multiple datasets, and verify that our CWT-GAN has achieved some state-of-the-art results. In the future, CWT-GAN can also be extended to multi-modal and multi-domain image-to-image translation problems.

6. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 61802297).

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[2] Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8165–8174, 2020. 2, 4, 5, 6, 7

[3] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 415–423, 2015. 1

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 1

[5] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. 1

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 5

[9] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017. 4

[10] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Computer Vision – ECCV 2018*, pages 179–196, Cham, 2018. Springer International Publishing. 1, 5, 6, 7

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*, pages 694–711, Cham, 2016. Springer International Publishing. 1

[12] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019. 2, 4, 5, 6, 7

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[14] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017. 1

[15] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 1, 5, 6, 7

[16] Shuang Ma, Jianlong Fu, Chang Wen Chen, and Tao Mei. Da-gan: Instance-level image translation by deep attention generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2018. 1, 2

[17] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, 2017. 3

[18] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 3

[19] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*, 2018. 2

[20] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2536–2544, 2016. 1

[21] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 3

[22] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *arXiv preprint arXiv:1911.11897*, 2019. 1, 2

[23] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8004–8013, 2019. 2

[24] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision – ECCV 2016*, pages 649–666, Cham, 2016. Springer International Publishing. 1

[25] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 2

[26] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017. 1, 2, 5, 6, 7