Generating Sea Surface Object Image Using Image-to-Image Translation

Wenbin Yin
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: 1139004179@qq.com

Zhiyi Hu
Engineering Design Institute
Army Research Laboratory
Beijing, 100042, China
E-mail: 763757335@qq.com

Jun Yu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: yujun@xatu.edu.cn

***Abstract*—Sea objects training, the conditional adversarial networks require a large number of images to solve image-to-image translation problems. In the case of insufficient samples, it leads to network overfitting and poor training results. This project proposes a conditional adversarial generative model that retains the original background features in the absence of paired samples. The goal of this project is to reduce the deviation of the corresponding output from the original input. Firstly, the object images of different categories are labeled with color masks. Second, sea objects are generated randomly in the original background using model of this project. Finally, the generated results of this approach are compared with other approaches. The experimental results show that, compared with results from other conditional adversarial generative models, the generated object images using model of this project have the characteristics of richer texture and clearer structure.**

Keywords—Image Generation; Conditional Generative Adversarial Network; Sea Surface Object; Image-To-Image Translation.

I. INTRODUCTION

Image generation is one of the popular research domain of computer vision. It is a technology for generating images based on known content (e.g. text, image). It is considered as image-to-image translation when the content is images. 3 In general, a new image can be quickly generated from several images by human beings. However, for machine learning, it can train a large number of images using Generative Adversarial Network (GAN) so as to generate new images. GAN contains a generator and a discriminator, both of which reach the maximum value of their expected benefits during the process of antagonism. 1The generator learns the data distribution of the sample set in order to generate similar data. The discriminator is used to discriminate whether the data comes from real data or generated fake data. GAN generates samples from random noise, so it exists the defect of uncontrollable information generation and free training process. Thus, CGAN2 adds additional conditional information to GAN in order to control the training process of the generator and the discriminator. The existing data-driven CGAN has far less generality than human learning ability. In the absence of data, it has practical significance that how to imitate the human learning process and design a more reasonable method to generate image.

The focus of computer vision tasks is the object itself, such as object detection18. However, due to confidentiality regulations and the high cost of acquiring images in medical, military and other fields, they are extremely short of training data sets. For example, the role of target detection in

DOI: 10.21307/ijanmc-2021-016

the search and rescue of sea ships and the recognition of sea ship types is becoming more and more significant. This often requires a sea object data set as large as possible for network training. However, in reality, the cost of acquiring such a large data set is extremely high, so it is necessary to use an adversarial generation network model to expand the existing sample set.

A significant limitation of most existing image-to-image translation methods is the lack of input. In order to solve this problem, this project proposes to generate object outputs given the same background input. This method is based on the 'pix2pix' framework, and designs a network framework for object image generation to learn the mapping from the input object label to the output real object image. Therefore, this project proposes an adversarial generation model to generate object image in the original background. This project design a conditional adversarial generation network structure for object image generation.

In this network structure, generator network structure is composed of an encoder, a converter, and a decoder. The encoder extracts feature through a convolutional neural network. The converter uses a residual module to preserve the original image features while converting. The decoder completes the work of restoring low-level features from the feature vector through the deconvolution layer. Such a generator not only retain the characteristics of the original input, such as the size and shape of the object, but also learn abundant semantic information. The discriminator uses 70×70 Patch-GANs. It predicts each individual small image block separately, and finally determines the authenticity of the whole image. The model is applied to the expansion of sea object image data set. Experiments show that the network can generate real sea objects at specified locations. Compared with other image to image translation supervision networks, the network of project can generate better quality object images.

II. NETWORK STRUCTURE

GAN is too free so that the result is uncontrollable. This project chooses to use additional information x to constrain the model, so as to guide the generation of samples. As shown in Figure 1, the network of project structures uses a conditional generation confrontation network model. In this model, additional information x is input to the generator and discriminator as a condition, and noise Z is input to the generator. The additional information x can be any kind of auxiliary information, such as category labels, sketches and text. This project can perform conditional processing by inputting the label as an additional input layer into the discriminator and generator. 3

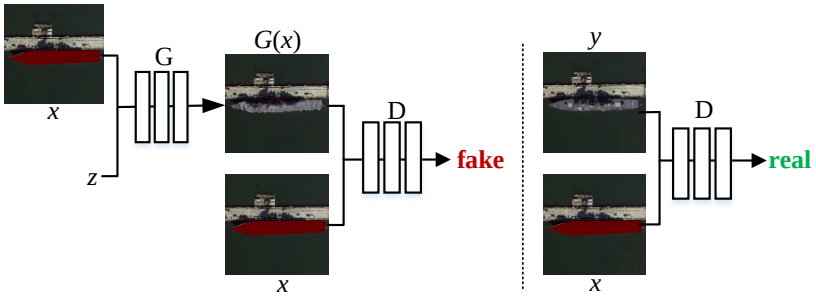

Figure 1. Conditional generative adversarial network Structure.

Conditional Gan transforms the data distribution probability in the original Gan loss function into conditional probability. This project uses the least square method as the loss function. The object L (D, G) to be optimized is defined as follows:

$$L_{cGAN}(G,D) = E_{x\sim p_{data(x)}}[\log D(x|y)] + E_{z\sim Pdata(z)}[\log(1-D(G(x|y)))] \quad (1)$$

Previous methods have found that it is beneficial to mix GAN objective with a more traditional loss such as L1 distance. The task of the generator is not only to cheat the discriminator, but also to approach the ground truth output in L1 sense. This project uses L1 distance because L1 encourages less blurring:

$$L_{L1}(G) = E_{x,y}\left(\left\|y - G(x|y)\right\|_1\right) \quad (2)$$

The final object loss function becomes as follows.

$$L(G,D) = L_{cGAN}(G,D) + \lambda L_{L1}(G) \quad (3)$$

Where $\lambda \in [0,1]$ is a variable for guaranteeing the similarity between the input and output.

The final goal is as follows.

$$G = \min_G \max_D L(G,D) \quad (4)$$

Where G tries to minimize this objective against an adversarial D that tries to maximize it.

These two networks, generator G and discriminator D, can be networks of any architecture. In the following sections, this project will provide more information about G and D. The job of project is to generate remote sensing ship images using a deep convolutional network.

A. Network structure of Generator G

A feature of image-to-image translation problems is to map a high-resolution input grid to a high-resolution output grid. The input and output of the generator are different in appearance, but their renderings have the same underlying structure.

Previous works24,25,26 used an encoder-decoder network in this area. These networks require the input flow to pass through a series of progressively lower sampling rates (that is down-sampling), up to the bottleneck level. And then progressively higher *sampling rates (that is up-sampling). In the down*-sampling stage, the network can capture some simple features of the image, such as boundary, color and so on. When more and more convolution operations are used, some useless abstract features in the image will be captured, which will lead to the inconsistency of some underlying structures between the input image and the output image, and then affect the image to image conversion. Therefore, this project adds a converter according to "Resnet"27 to build a generator network structure of "encoder-converter-decoder", as shown in Figure 2.

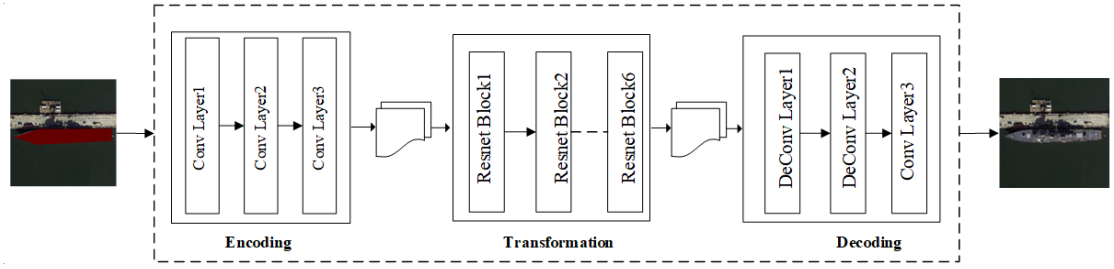

Figure 2. Network structure of Generator

In generator network structure, the encoder extracts feature from the input label x through a convolutional neural network. The converter uses 6 residual modules composed of two convolutional layers. It achieves the goal of retaining the original image features while converting. The decoder completes the work of restoring low-level features from the feature vectors through the deconvolution layer, and finally obtains a generated image G (x). The goal of project task is to preserve the characteristics of the original input, such as the size and shape of the object.

In Figure 2, the size of the original image is (256,256,3). The encoder extracts features through convolutional layers. The convolutional layer gradually extracts more advanced features. The input image changes from (256,256,3) to (256,256,64) after encoder, and continues to be transmitted to the next convolutional layer. The number of features is same with the number of filters correspondingly. The first layer has 64 filters and a batch normalization. Through the second convolutional layer of the encoder, it changes from (256,256,64) to (128,128,128). After the second convolution, the size is halved, and the

number of channels is doubled. The final output is changed from (128,128,128) to (64,64,256). Since the higher the convolutional layer is, the number of high-level features needs to be increased, so the image is compressed into 256 feature vectors with a size of 64×64.

The function of the network layer in the transformation module is to combine similar features of the image. The main goal is to retain the characteristics of the original image, such as the size and shape. Therefore, the residual network is suitable for these conversions. A 6-layer Resnet module is used here. Each sub-residual block is a network layer composed of two convolutional layers, some data of which are directly added to the output. The transformation ensure that the input information of the former layer is passed on the subsequent layer directly, so that the corresponding output has very small deviation from the input. The features of the original image can be remained in the output, and the output is similar to the object contour.

The decoder is composed of three deconvolution layers. The decoding process is completely opposite to the encoding method. Use the deconvolution to restore the low-level features from the feature vector. The difference between the decoder and the encoder is the value of window size and the moving steps. The window size of the three-layer convolution of the encoder is 7, 3, and 3.The window size of the three-layer deconvolution of the decoder is 3, 3, and 7. The moving steps of the three-layer convolution of the encoder are 1, 2, and 2. The moving steps of the three-layer deconvolution of the decoder are 2, 2, 1. Finally, a image with the size of (256,256,3) can be obtained.

B. Network structure of Discriminator D

Similar to Isola et al, the discriminator of project uses 70×70 Patch-GANs. Patch-GANs is a deep full convolu*tional network. Figure 3 describes the network struc*ture of the discriminator. The 256×256 original input image is subjected to a five-layer down sampling convolution operation to obtain a 30×30 feature map. Each point in the feature map corresponds to the size of the original image. The output is a 30 ×30 matrix, and the element in the matrix is true or false.

The discriminator uses the block algorithm to predict the true and false of each image block, and makes statistics to determine the authenticity of the final image. In this way, the image block discriminator has fewer parameters than the full discriminator to speed up the training speed. The fully convolutional manner applied to any size images can increase scalability of the framework with no restrictions on the input size. The output of the network is a prediction matrix. Specifically, each term represents an N ×N image block in the input. Finally, the discriminator gives the final prediction result by averaging the prediction results of all image blocks.

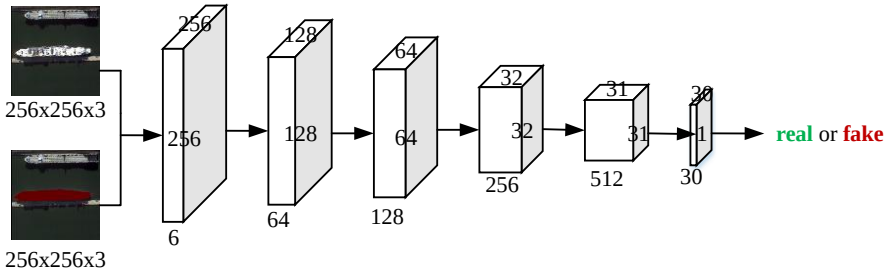

Figure 3. Network structure of Discriminator

III. EXPERIMENT AND RESULT ANALYSIS

In order to verify the method, this project uses CGAN model to generate sea object images. Next, this project introduces the experimental plan, the preparation of the experiment, the experimental results and analysis.

A. Experimental scheme

The purpose of this experiment is to generate new sea object images.

1) Several unprocessed original ocean background images are selected.

2) These ocean background images are pre-processed to unified into 256×256 images.

3) The object mask is randomly chosen from the existing image training set.

4) A image of sea background and a ship mask are combined to make a conditional mask image.

5) A new ship image is generated from the conditional mask image, which is the final object image.

6) Repeat the operation from step 3 to step 5 to generate multiple different images of ships on the sea.

In this experiment, different models use the same parameters when training on the sea object image data set. The batch size of training is 1. The initial learning rate is set to 0.0002. The number of training iterations epoch is set to 1000. Adam method is used as the training optimizer, in which the momentum parameter is 0.5.

C. Dataset preparation

In this experiment, 42556 original images were collected from the open sources data set of Kaggle and Google Earth. Some images are fuzzy. In order to generate more realistic high-quality images, these images are preprocessed. The pretreatment process is as follows.

1) The larger image is clipped, and the ship is clipped to the middle of the image as far as possible.

2) Clear ship pictures are selected.

3) There is reflective phenomenon in the sea surface, so the filtered image is denoised by median filtering to remove the reflection.

4) Unify the image size to 256×256.

5) Several data enhancement methods are used: *random 90° multiple rotation (P = 1); horizontal flip (P = 0.25); random brightness (P = 0.2, limit = 0.2), etc.*

6) The ship and other sea objects are masked *with different colors. The RGB value of ship tag is (100,0,0), and that of carrier condition tag is (0100,0).*

B. Environment configuration

All experiments are carried out under Linux system. In order to improve the computing speed, NVIDIA GeForce 2080Ti graphics card is used. Pytorch is chosen as the deep learning framework, which can be regarded as a powerful deep neural network with automatic derivation function. The detailed experimental configuration is shown in Table 1.

TABLE I. EXPERIMENTAL ENVIRONMENT

Hard-ware	**CPU**	**Intel(R)XEON Gold 6254**
	Display card	**Nvidia GeForce 2080Ti**
Software	Operating system	Ubuntu 18.04 LTS 64bit
	Learning framework	pytorch 0.4
	Programming language	Python
	Compiler	Pycharm2019.1

(a) Ship

(b) Carrier

Figure 4. Data annotation example

After screening and preprocessing, 315 original images of sea surface objects are obtained. And then 1334 images are obtained after data enhancement. Then the data set is divided into training set (1067 pieces) and verification set (267 pieces) by the ratio of 0.8 and 0.2. Figure 4 shows the original image and mask of ship and carrier, respectively.

D. Experimental results and analysis

In this experiment, the background this project chooses is the coastal port background. The sea

objects are mainly the ship or aircraft carrier. According to the experimental scheme described in Section 5.1, this project adds a conditional mask to the specified position in the port background, and then this project network model automatically generates a new ship image according to the conditional mask. According to this method, this project can generate object images under the specified ocean or port background conditions, which can be used as the extended data set of ship.

The project compares the method with the generated results of Pix2pixGAN3, Pix2pix-HD4 and BicycleGAN8 in terms of visual effects and evaluation indicators. Figure 5 shows the contrast results.

Comparison of visual effects. Figure 5 shows the remote sensing images of sea objects generated by this project model and the other three models. After selecting four original ocean images, the Figure 5 shows that each column of images from left to right is the original image, the masked image, the image generated by this method, the image generated by Pix2pixGAN, the image generated by Bicycle-GAN, and the image generated by Pix2pix-HD. It can be seen that the images generated by Pix2pixGAN and Bicycle-GAN models have the problems of texture detail distortion and structure distortion. This is because that, part of the background information is not completely recovered in the reconstruction process, resulting in the background blur of the generated samples. The results of Pix2pix-HD are clearer than the other two methods, but the edge contours of the target are distorted. Relatively, this model can generate different ship images from the original ship according to the tag information, and the generated ship images are better than other models in terms of color and texture details.

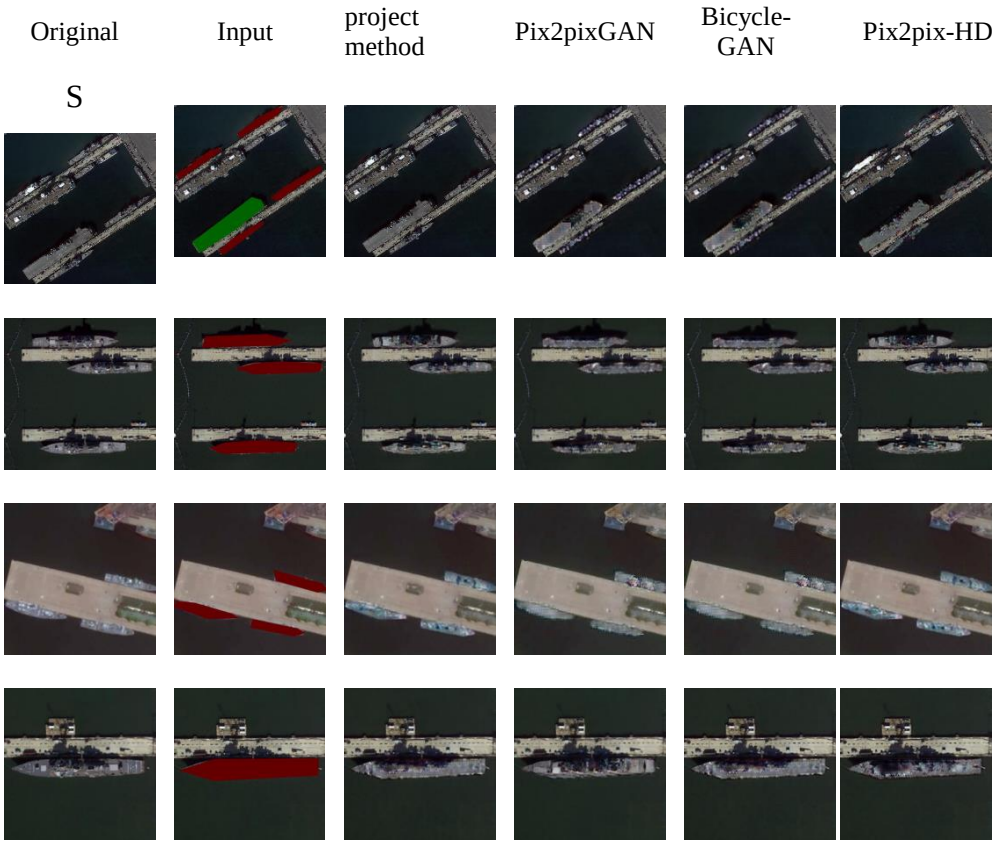

Figure 5. The Results of Comparative Experiment

1) Comparison of visual effects. Figure 5 shows the remote sensing images of sea objects generated by this model and the other three models. After selecting four original ocean images, the, Figure 5 shows that each column of images from left to right is the original image, the

masked image, the image generated by this method, the image generated by Pix2pixGAN, the image generated by BicycleGAN, and the image generated by Pix2pix-HD. It can be seen that the images generated by Pix2pixGAN and BicycleGAN models have the problems of texture detail distortion and structure distortion. This is because that, part of the background information is not completely recovered in the reconstruction process, resulting in the background blur of the generated samples. The results of Pix2pix-HD are clearer than the other two methods, but the edge contours of the target are distorted. Relatively, this model can generate different ship images from the original ship according to the tag information, and the generated ship images are better than other models in terms of color and texture details.

2) Comparison of evaluation indicators. This project comprehensively evaluates the visual quality of the generated images. The evaluation indicators use namely peak signal-to-noise ratio (PSNR) and structural similarity (SSIM).

These two indexes can accurately reflect the image quality. PSNR is the ratio between information and noise, which can be used to measure the distortion or noise level of the generated image. If the PSNR value between images is larger, the images are more similar and more realistic. The definition of PSNR is shown in formulas 5 and 6.

$$\mathrm{MSE}=\frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(I_0(i,j)-I(i,j)\right)^2}{\mathrm{M}\times\mathrm{N}} \tag{5}$$

$$\mathrm{PSNR}=10\log\left(\frac{\mathrm{MAX}_I^{\,2}}{\mathrm{MSE}}\right) \tag{6}$$

Where MSE is the mean square error of the corresponding pixels of the original image and the generated image. $I_0(i,j)$ and $I(i,j)$ in the mean square error respectively represent the pixel values at (i,j) in the original image and the generated image. $M\times N$ represents the number of pixels in the calculated image. MAX_I represents the maximum value of image pixels, which is 255 by default.

SSIM is an objective index to measure the image similarity. It is more suitable for human eyes to judge the visual effect. SSIM is calculated according to the brightness and contrast of the image. The definition of SSIM is shown in formula (8).

$$SSIM(\mathrm{a},b)=\frac{(2\mu_a\mu_b+C_1)(2\sigma_{ab}+C_2)}{(\mu_a^{\,2}+\mu_b^{\,2}+C_1)(\sigma_a^{\,2}+\sigma_b^{\,2}+C_2)} \tag{7}$$

Where μ_a is the average of all pixel values of image a, and μ_{b} is the average of all pixel values of image b. $\sigma_{\mathrm{a}}^{\,2}$ is the pixel variance of image a. $\sigma_{\mathrm{b}}^{\,2}$ is the pixel variance of image b. $\sigma_{\mathrm{ab}}^{\,2}$ is the pixel covariance of image a and image b. $c_1=(k_1L)^2$ and $c_2=(k_2L)^2$ are constants used to ensure stability. L is the dynamic range of pixel values. $k_1=0.01$, $k_2=0.03$. The larger SSIM value is, the higher the similarity between the two images. The range of SSIM is [0,1]. If two images are exactly the same, the SSIM value is 1. Otherwise, if two images are completely different, the SSIM value is 0.

TABLE II. COMPARISON OF EVALUATION INDEX OF DIFFERENT METHODS

Method	PSNR	SSIM
project model	18.453	90.39%
Pix2pixGAN model	16.874	80.42%
Bicycle-GAN model	18.096	87.61%
Pix2pix-HD model	17.919	86.42%

In this experiment, this project compares the original with the generated, and calculate the PSNR and SSIM. Finally, the project gets the comparison results shown in Table 2. It can be seen that the PSNR value of this model is significantly higher than that of other models, which indicates that the image generated by this model is more similar to the original image. The SSIM index of this model is also higher than other models, which indicates that the image generated by this model is closer to the original object image. Therefore, this model is superior to pix2pixGAN, Bicycle-GAN and pix2pix-HD.

IV. CONCLUSIONS

In this paper, the sea object image is generated by adding conditional mask in the specified posi-

tion under the condition of less input of original images. In order to verify the effectiveness of this model, the visual effect and evaluation indexes are compared with the existing Pix2pixGAN model, Bicycle-GAN model and Pix2pix-HD model. The results show that the object image generated by this model is obviously better than the other three models. This model is a promising method for many images to image translation tasks.

REFERENCES

[1] Goodfellow, J. Pouget-Abadie, M. Mirza, et al, Generative adversarial nets, Advances in neural information processing systems (2014), pp. 2672-2680.

[2] M. Mirza, S. Osindero, Conditional Generative Adversarial Nets, Computer Science (2014)pp.2672-2680.

[3] P. Isola, J.-Y. Zhu, T. Zhou, et al. Image-to-image translation with conditional adversarial networks, in the IEEE conference on computer vision and pattern recognition (CVPR) (2017), pp. 1125-1134.

[4] T. C. Wang, M. Y. Liu, J. Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional GANs, In the IEEE conference on computer vision and pattern recognition (CVPR) (2018), pp. 8798-8807.

[5] M. Zhai, L. Chen, F. Tung, J. He, M. Nawhal, G. Mori, Lifelong gan: Continual learning for conditional image generation. In the IEEE International Conference on Computer Vision(ICCV) (2019), pp. 2759-2768.

[6] D.Bau, H.Strobelt, W.Peebles, et al. Semantic photo manipulation with a generative image prior[J]. arXiv preprint arXiv:2005.07727, 2020.

[7] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, Infogan: Interpretable representation learning by information maximizing generative adversarial nets, Advances in neural information processing systems (NIPS) (2016), pp. 2172-2180.

[8] J.Y.Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, E. Shechtman, Toward multimodal image-to-image translation, Advances in neural information processing systems (NIPS) (2017), pp. 465-476.

[9] J. Y. Zhu, T. Park, P. Isola, et al, Unpaired image-to-image translation using cycle-consistent adversarial networks, In the IEEE international conference on computer vision (ICCV) (2017), pp. 2223-2232.

[10] W. Xian, P. Sangkloy, V. Agrawal, et al. Texturegan: Controlling deep image synthesis with texture patches, In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018), pp. 8456-8465.

[11] Y. Lu, S. Wu, Y. W. Tai, et al, Image generation from sketch constraint using contextual gan, in the European Conference on Computer Vision (ECCV) (2018), pp. 205-220.

[12] A. Gonzalez-Garcia, J. Van De Weijer, Y. Bengio, Image-to-image translation for cross-domain disentanglement, Advances in neural information processing systems (NIPS) (2018), pp. 1287-1298.

[13] H. Tang, D. Xu, G. Liu, W. Wang, N. Sebe, Y. Yan, Cycle in cycle generative adversarial networks for keypoint-guided image generation. In the 27th ACM International Conference on Multimedia (2019, October), pp. 2052-2060.

[14] Z. Gan, L. Chen, W. Wang, Y. Pu, Y. Zhang, H. Liu, C. Li, L. Carin, Triangle generative adversarial networks. In NIPS. (2017) pp. 5253–5262.

[15] 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 CVPR. (2018)

[16] Huang X, Liu M Y, Belongie S, et al. Multimodal unsupervised image-to-image translation. In ECCV. (2018)

[17] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In NIPS, 2017.

[18] Taigman, Y., Polyak, A., Wolf, L. Unsupervised cross-domain image generation. In ICLR. (2017)

[19] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, D. Krishnan, Unsupervised pixel-level domain adaptation with generative adversarial networks. In CVPR. (2017)

[20] 20. E. Hosseini-Asl, Y. Zhou, C. Xiong, R. Socher, Augmented cyclic adversarial learning for low resource domain adaptation (2018). arXiv preprint arXiv:1807.00374.

[21] M. Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, J. Kautz, Few-shot unsupervised image-to-image translation, In the IEEE International Conference on Computer Vision (ICCV) (2019), pp. 10551-10560.

[22] T. C. Wang, M. Y. Liu, A. Tao, G. Liu, J. Kautz, B. Catanzaro, Few-shot video-to-video synthesis, (2019) arXiv preprint arXiv:1910.12713.

[23] A. Torralba, Contextual priming for object detection, International journal of computer vision, 2003, 53(2): 169-191.

[24] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In ECCV. (2016)

[25] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In CVPR. (2016)

[26] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon. Pixel- level domain transfer. In ECCV. (2016)

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. In the IEEE con ference on computer vision and pattern recognition (CVPR) (2016), pp. 770-778.