Unsupervised image-to-image translation by semantics consistency and self-attention*

ZHANG Zhibin, XUE Wanli, and FU Guokai**

The Key Laboratory of Computer Vision and System of Ministry of Education, Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, Tianjin 300384, China

(Received 26 October 2020; Revised 17 September 2021)

Unsupervised image-to-image translation is a challenging task for computer vision. The goal of image translation is to learn a mapping between two domains, without corresponding image pairs. Many previous works only focused on image-level translation but ignored image features processing, which led to a certain semantics loss, such as the changes of the background of the generated image, partial transformation, and so on. In this work, we propose a method of image-to-image translation based on generative adversarial nets (GANs). We use autoencoder structure to extract image features in the generator and add semantic consistency loss on extracted features to maintain the semantic consistency of the generated image. Self-attention mechanism at the end of generator is used to obtain long-distance dependency in image. At the same time, as expanding the convolution receptive field, the quality of the generated image is enhanced. Quantitative experiment shows that our method significantly outperforms previous works. Especially on images with obvious foreground, our model shows an impressive improvement.

Document code: A **Article ID:** 1673-1905(2022)03-0175-6

DOI https://doi.org/10.1007/s11801-022-0165-3

The task of image-to-image translation is to transform an image from domain A to domain B, as shown in Fig.1. In recent years, the emergency of deep learning provided new ideas for this task. Previous works often required a large number of paired images and deliberately designed loss functions. However, in most cases, pairs of images are difficult to obtain, and specific loss functions are often unable to process images with different characteristics well. In fact, as long as we know the distribution of the two kinds of data, we can complete the transformation. The appearance of generative adversarial nets (GANs)[1] solved this problem.

GANs have shown extraordinary power in image-to-image translation. These methods are divided into supervised methods and unsupervised methods. Supervised methods usually need a large number of pairs of images. In general, it is not easy to obtain a large number of pairs of images, and cannot guarantee that the images completely correspond to each other. ISOLA et al[2] utilized conditional GANs to learn the mapping from input image to output image. Image-to-image translation was taken as a pixel-to-pixel mapping problem. They set conditions on the input image to get the corresponding output image. Unsupervised methods try to learn the distribution of different data and achieve the goal of image translation by fitting the distribution of data.

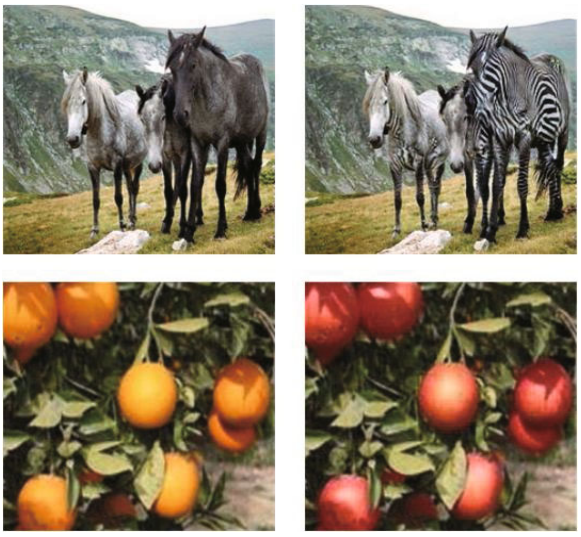

Fig.1 Examples of image translation

Based on the intuition that two kinds of distributed data are transformed into each other, CycleGAN[3] used the loss of cyclic consistency to transform images. In fact, both supervised and unsupervised methods only focused on image level conversion but ignored the processing of image features, which led to the changes of image semantics at the same time of image translation. Fig.2

* This work has been supported in part by the National Natural Science Foundation of China (Nos.61906135, 62020106004, 92048301 and 61906027), and the Tianjin Science and Technology Plan Project (No.20JCQNJC01350).
** E-mail: xuewanli@email.tjut.edu.cn

shows several fail cases of CycleGAN.

Fig.2 Fail cases of CycleGAN with the background changed

In this paper, we propose a novel unsupervised image-to-image translation method, which adds the semantic consistency loss function to the generator in manner of end-to-end, so that the model focuses on more important not only pixel level features, but also semantic level features. Besides, a self-attention mechanism is added to the tail of the generator to obtain long-distance dependence and improve the quality of image generation. The main innovations of this paper are as follows.

(1) We propose a semantic undeformed loss function for end-to-end training of constrained unsupervised image transformation to maintain the semantic consistency between the original graph and the generated graph.

(2) Our self attention module is used to increase the receptive field of convolution network, obtain long-distance dependence and optimize the quality of image generation.

(3) In many unsupervised transformation tasks, the stupid method has higher image transformation quality than other most advanced methods.

GANs have been adopted in many images processing tasks, and have shown impressive results in image generation image inpainting, image editing, and so on[4]. The key to the success of GANs is to use the idea of zero-sum game. The task of generator is to make fake image, while the task of discriminator is to recognize the generated image and the real image. With the training process, the image produced by generator is closer to the real image, which deceived the discriminator and finally achieved Nash equilibrium.

Autoencoder can be used to extract the latent representations of data. It has been used in feature extraction, data denoising and many other fields[5]. LARSEN et al[6] combined variational autoencoders (VAEs) and GANs into an unsupervised generation model, to learn representations to better measure similarities in data space. ZHAO et al[7] proposed to deploy an energy-based model in the latent space of a pretrained autoencoder for image-to-image translation. Considering that autoencoder is often used for feature extraction, in our generator, a 256×256 image is subsampled three times and then enters nine rest blocks. The corresponding features are obtained. The decoder obtains a 256×256 image after five times of deconvolution.

Inspired by the mechanism of human attention, the neural network also uses attention mechanism to extract features from images, text and so on. Attention model has been widely used in many kinds of deep learning tasks, such as natural language processing and image recognition. Convolutional neural network (CNN) has the limitation of receptive field, so attention mechanism is used to obtain long-distance dependence of image. In recent years, there are also many works that combine GANs and attention mechanism to achieve unprecedented results. CHEN et al[8] utilized the attention network to predict spatial attention maps of images and translated images by transformation network. ZHANG et al[9] obtained the long-distance relationship on the image by self-attention.

The original GANs used the minimax game in the form of log to obtain the Nash equilibrium, which leads to difficulties in training, where the loss of generator and discriminator cannot indicate the training process, and it lacks of diversity in generating samples. Wasserstein GAN[10,11] replaced log loss with Wasserstein distance. MAO et al[12] used the least square loss to stabilize the training process of GANs. In our network, least squares GAN (LSGAN) and PatchGAN are combined. Fine-grained experiments confirm that this is effective.

The attention mechanism in neural networks is generally applied to the deeper part of neural networks, because the deeper networks can learn more detailed features, after the generator extracts image features. In the decoder part, we use self-attention mechanism, which makes our network pay more attention to the main features of the image and enables our network to capture long distance dependence. Self-attention first calculated the correlation of each pixel in the image.

$$\beta_{ji} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})}, \tag{1}$$

$$s_{ij} = f(x_i)^{\mathrm{T}} g(x_j), \tag{2}$$

$$f(x) = W_f x, \tag{3}$$

$$g(x) = W_g x, \tag{4}$$

where s_{ij} indicates the correlation between each pixel and other pixels. The image features from the previous

hidden layer $x \in \mathbb{R}^{C\times N}$, where C is the number of channels and N is the number of feature locations of features from the previous hidden layer. Both W_f and W_g represent 1×1 convolution.

$$o_j = v\left[\sum_{i=1}^{N}\beta_{ji}h(x_i)\right], \tag{5}$$

$$h(x) = W_h x, \tag{6}$$

$$v(x) = W_v x, \tag{7}$$

where the image is restored by matrix operation. The output of the attention layer is $o = (o_1, o_2, ...o_j ..., o_N) \in \mathbb{R}^{C\times N}$, and W_h and W_v represent 1×1 convolution.

$$X \to G(X), \tag{8}$$

$$Y \to F(Y), \tag{9}$$

$$F(G(x)) \approx G(F(Y)), \tag{10}$$

where for each image x from domain X, the image translation cycle should be able to bring x back to the original image. CycleGAN used the loss of cyclic consistency to translate images. Both $G(X)$ and $F(X)$ represent the generator.

The loss function of CycleGAN cycle consistency is mainly applied to the image, but there is no special processing in the aspect of features. The unconstrained features will cause the deep convolution neural network cannot well represent the semantic features when updating the weights. Therefore, this paper further adds the loss function of semantic consistency in the above process, and the whole network adopts the structure of autoencoder. Deep convolution neural network can learn the advanced features of images[13]. We use the autoencoder structure to learn the latent representation of images. Fig.3 shows the general framework of our model. Encoder A is the encoder of the A domain, and Feature A is the feature extracted by the encoder for the input. We can understand that Feature A is the semantic of the input. We get the transformed image by decoding. According to the above description, we input fake images into the Encoder B of the B domain to get the corresponding feature. Then we make the consistency loss function of Feature A and Feature B. We can strengthen the relationship between A and B semantically, so as to solve the problem of semantic transformation caused by the transformation of target domain proposed at the beginning, that is, the semantics after transformation changes compared with the semantics of source domain.

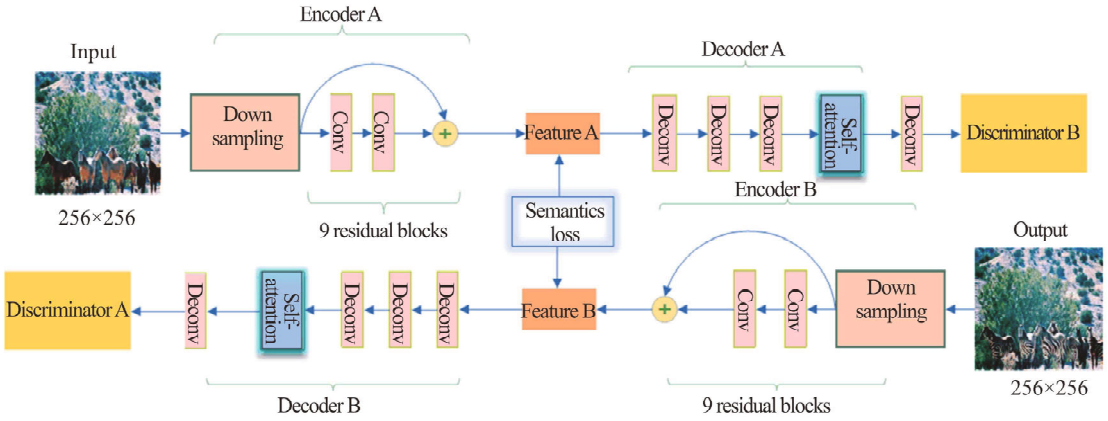

Fig.3 General framework of our network (The generator can be seen as autoencoder to extract features of images and generate the output; Semantics loss is added to maintain the semantics of images)

The training process of traditional GANs is very unstable, to a large extent, because its objective function may have gradient dispersion, especially when the objective function is minimized, which makes it difficult to update the generator. LSGAN solved this problem by samples of decision boundary of penalty principle.

$$\min_D V_{\text{LSGAN}}(D) = \frac{1}{2}E_{x\sim p_{\text{data}}(x)}[(D(x)-b)^2] + \frac{1}{2}E_{z\sim p_z}[(D(G(z))-a)^2], \tag{11}$$

$$\min_D V_{\text{LSGAN}}(G) = \frac{1}{2}E_{z\sim p_z}[(D(G(z))-c)^2], \tag{12}$$

$$Loss = Loss_{\text{GAN}} + Loss_{\text{L1}}, \tag{13}$$

where D is a generator, G is a discriminator, and D and G simultaneously perform min training as to competitors with value function of $\min V_{\text{LSGAN}}$, with p_z as the input prior distribution and $p_{\text{data}}(x)$ as the training data distribution. Finally, a, b and c are hyper-parameters, a=c=1, b=0. The LSGAN model is trained in an alternating fashion by minimizing the hinge version of the adversarial loss.

In general, the network structure of GANs is not suitable for the image field which requires high resolution and high detail preservation. Some researches have designed PatchGAN according to this situation. The main difference of this GAN is the discriminator. Generally, GANs only needs to output a true or false vector, which represents the evaluation of the whole image. However, PatchGAN outputs a matrix with size of N×N, each element of the matrix with size of N×N. We design our discriminator by combining PatchGAN and LSGAN.

In this section, we will show more training details, and

the advantages of our approach. Two time-scale update rule (TTUR)[13] used different learning rates for discriminators and generators. Generally, generators use lower update rate, and discriminators use higher update rate. With this method, we can perform discriminator and generator updates in a rating of 5:1, and only the learning rate needs to be modified. We use the learning rate of 0.001 for generator and 0.002 for discriminator and Adam optimizer, batch size of 8 for 400 000 iterations. Fig.4 shows the loss curves. The original GAN is a complex generation model. GAN learning is a game between generator and discriminator. Since the training of GAN is a game, its gradient descent may fail to converge. At that time, when using two stages, GAN converged to the stable local Nash equilibrium, but the loss value is not the best[14]. When using the semantic loss proposed in this paper, the generator and discriminator are optimized at the semantic level, so the network can quickly converge to an optimal model.

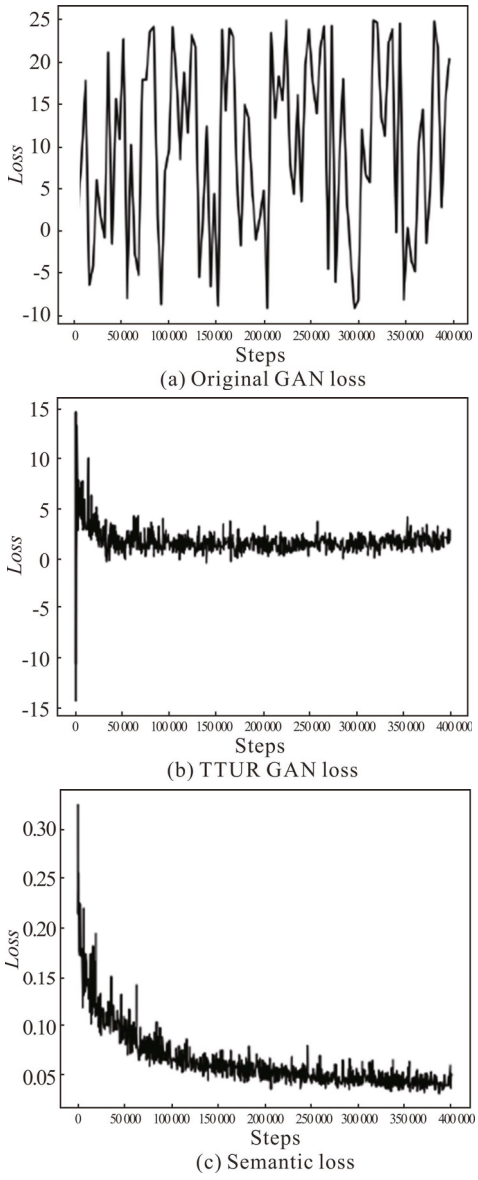

Fig.4 Loss curves in training process

To evaluate our methods, we conducted extensive experiments on the horse-to-zebra and orange-to-apple datasets. In order to evaluate the quality of the generated image, we calculated the peak signal-to-noise ratio (*PSNR*) and structural similarity (*SSIM*) of the generated image. Tab.1 shows the results. The Frechet inception distance (*FID*)[13] is a measure to calculate the distance between the real image and the feature vector of the generated image. To illustrate the advantages of our method, we calculate the *FID* of baselines. We use the same training steps to calculate the *FID* of baselines. From Fig.5, we can see that other methods lost a lot of semantic consistency, because the background has changed a lot, and our method has not lost semantic consistency while transforming the target.

Tab.1 Comparison with baselines

Data set	Method	*PSNR*	*SSIM*	*FID*
Horse2zebra	CycleGAN[3]	18.43	0.75	41.5
	MUNIT[15]	19.44	0.79	39.5
	AgGAN[16]	22.63	0.81	35.6
	Ours	**25.59**	**0.93**	**33.6**
Orange2apple	CycleGAN[3]	15.51	0.54	59.7
	MUNIT[15]	17.26	0.59	54.1
	AgGAN[16]	19.24	0.61	44.7
	Ours	**21.52**	**0.68**	**38.4**

In order to strengthen the connection between the generated image and the original image, L1 loss is used at the end of the network. In the process of back propagation, L1 loss promotes the semantic consistency between the generated image and the original image. The attention mechanism in the model is used to increase the receptive field of convolution network, so as to improve the quality of image generation. Fine-grained experiments show that our method is very effective. Tab.2 shows the results. After removing a part of the model, we retrained it. With learning rate of 0.001, we trained 400 000 steps and recalculated our evaluation metrics. As a result, in Tab.2, we can see that when only L1 loss is added, the performance of the model still exceeds that of baselines. After the attention mechanism is added, the performance of the model is further improved. When the two are combined, the best result is obtained, which shows that under the semantic constraints of attention, GAN model will optimize the target more accurately according to the target semantics, so the optimization result of the network model will be better.

Our model has a total of 6.5M parameters and is implemented on TensorFlow v1.14, CUDNN v7.0, CUDA v9.0. We performed all training runs on central processing unit (CPU) Intel(R) Xeon(R) CPU E5-2697 v3 (2.60 GHz) and graphics processing unit (GPU) GTX 1080. Our full model ran at 0.35 second per frame on

GPU and 1.2 seconds per frame on CPU for images with resolution of 256×256.

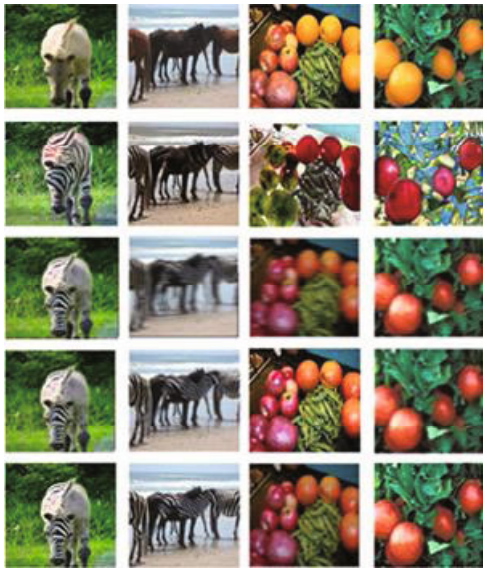

Fig.5 Translation results of CycleGAN[3], MUNIT[15], AgGAN[16], and our proposed method

Tab.2 Albation study

Data set	Method	*PSNR*	*SSIM*	*FID*
Horse2zebra	GAN+Semantic loss	22.34	0.83	39.6
	GAN+Attention	22.46	0.88	36.4
	GAN+Semantic loss+Attention	**25.59**	**0.93**	**33.6**
Orange2apple	GAN+Semantic loss	14.52	0.55	54.2
	GAN+Attention	17.76	0.59	43.6
	GAN+Semantic loss+Attention	**21.52**	**0.68**	**38.4**

Fig.6 shows several examples of failure. We can see that the apple to orange does not perform as well as horse to zebra. This may be due to the obvious difference between horse and zebra. How to learn less obvious features is still our future work. Due to the lack of certain types of data, our model cannot learn certain distributions. How to learn with a small number of samples is still a prominent problem.

In this paper, we proposed an unsupervised image translation method based on encoder-decoder structure. L1 loss on semantics features of the model ensures semantic consistency, and the use of attention mechanism enhances the quality of the generated image.

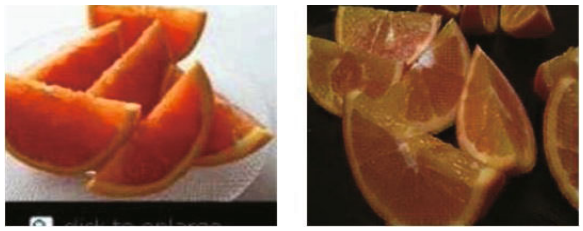

Fig.6 Fail cases of our model

Although our method can generate more realistic images, our approach has good generalization ability in many cases and can transform high-quality images. We propose a method to improve the quality of image generated by translation. Our model is superior to the existing unsupervised models in image quality and the original semantics can be reserved.

Statements and Declarations

The authors declare that there are no conflicts of interest related to this article.

References

[1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems (NeurIPS), December 8-13, 2014, Montreal, Quebec, Canada. Cambridge：MIT Press, 2014, 27：2672-2680.

[2] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, Hawaii, USA. New York：IEEE, 2017：1125-1134.

[3] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York：IEEE, 2017：2223-2232.

[4] CHEN X, JIA C Y. An overview of image-to-image translation using generative adversarial networks[C]// International Conference on Pattern Recognition (ICPR), January 10-15, 2021, Milan, Italy. Berlin, Heidelberg：Springer Cham, 2021：366-380.

[5] ZHU J Y, PHILIPP K, SHECHTMAN E, et al. Generative visual manipulation on the natural image manifold[C]//European Conference on Computer Vision (ECCV), October 8-16, 2016, Amsterdam, Netherlands. Berlin, Heidelberg：Springer-Verlag, 2016：597-613.

[6] LARSEN A, SØNDERBY S, LAROCHELLE H, et al. Autoencoding beyond pixels using a learned similarity metric[C]//International Conference on Machine Learning (ICML), June 19-24, 2016, New York City, NY, USA. New York：ACM, 2016：1558-1566.

[7] ZHAO Y, CHEN C Y. Unpaired image-to-image translation via latent energy transport[C]//Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021. New York：IEEE, 2021：16418-16427.

[8] CHEN X Y, XU C, YANG X K, et al. Attention-GAN for object transfiguration in wild images[C]//European Conference on Computer Vision (ECCV), September 8-14, 2018, Munich, Germany. Heidelberg：Springer-Verlag, 2018：167-184.

[9] ZHANG H, GOODFELLOW I, METAXAS D, et al. Self-attention generative adversarial networks[C]//International Conference on Machine Learning (ICML), June 10-15, 2019, Long Beach, CA, USA. New York：ACM, 2019：7354-7363.

[10] CAO J Z, MO L Y, ZHANG Y F, et al. Multi-marginal wasserstein GAN[C]//Advances in Neural Information Processing Systems (NeurIPS), December 8-14, 2019, Vancouver Convention Center, Vancouver, Canada. Cambridge：MIT Press, 2019, 32：1776-1786.

[11] LIU H D, GU X F, SAMARAS D. Wasserstein GAN with quadratic transport cost[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 27- November 2, 2019, Seoul, Korea. New York：IEEE, 2019：4832-4841.

[12] MAO X D, LI Q, XIE H R, et al. Least squares generative adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York：IEEE, 2017：2794-2802.

[13] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[C]//Advances in Neural Information Processing Systems (NeurIPS), December 8-14, 2017, Long Beach, CA, USA. Cambridge：MIT Press, 2017, 30：6626-6637.

[14] ZEILER M, FERGUS R. Visualizing and understanding convolutional networks[C]//European Conference on Computer Vision (ECCV), September 6-12, 2014, Switzerland, Zurich. Berlin, Heidelberg：Springer-Verlag, 2014：818-833.

[15] HUANG X, LIU M Y, BELONGIE S, et al. Multimodal unsupervised image-to-image translation[C]//European Conference on Computer Vision (ECCV), September 8-14, 2018, Munich, Germany. Berlin, Heidelberg：Springer-Verlag, 2018：185-208.

[16] MEJJATI Y, RICHARDT C, TOMPKIN J, et al. Unsupervised attention-guided image-to-image translation[C]//Advances in Neural Information Processing Systems (NeurIPS), December 3-8, 2018, Montreal Convention Centre, Montreal, Canada. Cambridge：MIT Press, 2018, 31：3693-3703.