Multi-branch Semantic GAN for Infrared Image Generation from Optical Image

Lei Li[1], Pengfei Li[1], Meng Yang[1,2(✉)], and Shibo Gao[3]

[1] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
yangm6@mail.sysu.edu.cn
[2] Key Laboratory of Machine Intelligence and Advanced Computing(SYSU), Ministry of Education, Guangzhou, China
[3] Beijing Aerospace Automatic Control Institute, Beijing 100854, China

Abstract. Infrared remote sensing images capture the information of ground objects by their thermal radiation differences. However, the facility required for infrared imaging is not only priced high but also demands strict testing conditions. Thus it becomes an important topic to seek a way to convert easily-obtained optical remote sensing images into infrared remote sensing images. The conventional approaches cannot generate satisfactory infrared images due to the challenge of this task and many unknown parameters to be determined. In this paper, we proposed a novel multi-branch semantic GAN (MBS-GAN) for infrared image generation from the optical image. In the proposed model, we draw on the idea from Ensemble Learning and propose to use more than one generator to synthesize the infrared images with different semantic information. Specially, we integrate scene classification into image transformation to train models with scene information, which assists learned generation models to capture more semantic characteristics. The generated images are evaluated by PSNR, SSIM and cosine similarity. The experimental results prove that this proposed method is able to generate images retaining the infrared radiation characteristics of ground objects and performs well in converting optical images to infrared images.

Keywords: Infrared image generation · Generative adversarial networks · Residual neural network

1 Introduction

Infrared imaging is a technique of capturing the infrared light from objects and converting them into visible images interpretable by a human eye. Near Infrared light is the portion of the electromagnetic spectrum that just past the Red light. The far infrared radiation that is also called thermal infrared radiation is heat emitted by any object that has a temperature above absolute zero. Different objects have different infrared reflectance which makes them look brighter or darker in infrared images.

Z. Cui et al. (Eds.): IScIDE 2019, LNCS 11935, pp. 484–494, 2019.
https://doi.org/10.1007/978-3-030-36189-1_40

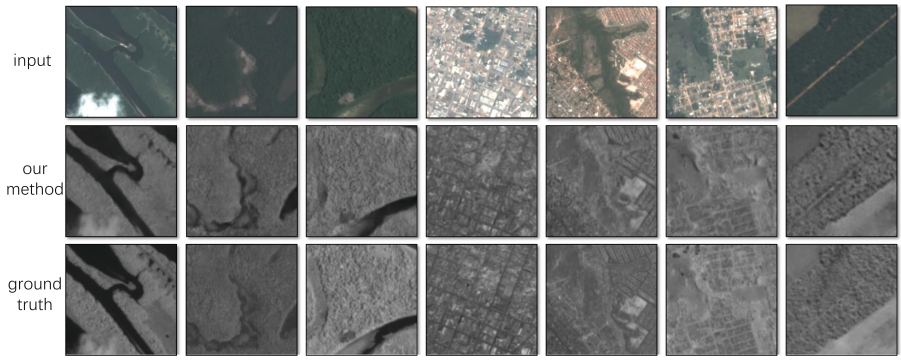

Fig. 1. The transformation results of our method. As shown, our method can generate results that are very close to groundtruth.

Compared with optical remote sensing images, infrared remote sensing images indicate more information about the essence and distribution of ground objects. However, the facility required for infrared imaging is not only priced high but also demands strict testing conditions. Converting easily-obtained optical remote sensing images into infrared remote sensing images helps overcome these restrictions.

Optical and infrared spectra provide different message. While optical images can provide information similar to what the human eye would see, optical images are incapable of providing useful information in situations where the illumination is poor or the weather is bad [1]. Infrared remote sensing images capture the information of ground objects by their thermal radiation differences. So in certain types of situations, infrared remote sensing images are useful than optical images because infrared information is independent of the quality of the environment. Moreover, infrared remote sensing images are widely applied to various fields such as military reconnaissance, climatology, and environmental monitoring. Therefore, we try to generate costly infrared images by more readily available optical images.

The challenge in our task is how to capture the infrared information from optical images. Most current researches utilize physical features, physical modeling and manual setting of environmental parameters to generate infrared images. Luo et al. [4] proposed a method that converts the infrared image into a grayscale image and then divides it into small parts. However, since the target object cannot be segmented completely, it is necessary to manually segment the target object from the background. After that, the temperature and related atmospheric parameters, which are used to calculate the amount of infrared radiation of the target object, of each segmentation area should be set manually. Finally, the infrared image is obtained by physical modeling. Wu et al. [5] use the histogram to convert optical images to infrared images by learning the characteristics of optical/infrared image pairs, but the method is mainly for the conversion of specific target objects (plants, buildings). Li et al. [6] proposed

a neural network-based infrared image generation method, which segments the visible light image into different regions, predicts the temperature of the target object of different materials, and then performs the radiation calculation, but manually segmentation is needed. And the results are directly affected by the segmentation of image. It must be mentioned that these methods have much difficulty in processing large quantities of images simultaneously.

Generative adversarial networks [13] is an effective model for image style transfer. The framework is good at capturing data distribution by learning from a big dataset. Recently, many generative adversarial models come out by extending the original GANs in different ways. For example, Pix2pixGAN [14] is a kind of conditional GAN requiring to input the real image into the generative model and discriminative model. It mixes the GAN objective with L1 distance to make the output near the ground truth output in an L1 sense. Zhu et al. [11] creatively proposed the framework that contains two generative adversarial networks. CycleGAN [11] only requires unpaired images for training rather than paired images. The training results make it possible to translate an image from each domain to another. StarGAN [12] is a novel model for multi-domain image-to-image translation. It combines domain information with image information to train just one model for multi-domain translation. With more and more research, the existing GAN methods have been able to generate higher quality images. But this is generally only for certain scenarios under the big data set. For example, learning from the natural scene image in the ImageNet [15] dataset, BigGAN [16] has been able to generate realistic and amazing results. However, for many applications in real-life scenarios, such as infrared remote sensing and visible light image dataset, corresponding adjustments are needed to generate satisfactory images. GAN [13] is a powerful framework for image style transfer. But for infrared images, different ground objects have their own infrared radiation characteristics. By simply using a framework based on GAN, e.g. pix2pixGAN, the transformation model would ignore some characteristics.

In this paper, we seek to find a method converting optical images into infrared images based on image style transfer. In order to retain the infrared radiation characteristics of different objects, we proposed a model that combines residual neural network and generative adversarial networks, which we named MBS-GAN. We draw on the idea from Ensemble Learning and propose to use more than one generator to synthesize the infrared images with different semantic information. Specially, we integrate scene classification into image transformation to train models with scene information, which assists learned generation models to capture more semantic characteristics. Our proposed model overcome this weakness and our result proves that this proposed method is able to generate images retaining the infrared radiation characteristics of ground objects and performs well in converting optical images to infrared images. Some results are demonstrated in Fig. 1.

2 Related Work

The problem to be solved in this paper is how to convert optical remote sensing images into more realistic infrared remote sensing images. As mentioned before, different ground objects have different infrared radiation characteristics. The proposed method in this paper combines scene classification and image style transfer. In this section, we review two methods of tasks above.

ResNet. In image classification, there are many classification models based on deep convolutional neural networks, e.g. AlexNet [7], GooLeNet [8] and VGGNet [9]. In most situations, the more the layers, the better the training results. Excessive number of layers may cause the problem of vanishing/exploding gradients. This problem is solved by normalized initialization and intermediate normalization layers. But there is another problem that as the depth of the network deepens, the training accuracy rate will gradually decline after getting saturated. He et al. [10] proposed a deep residual learning framework to address the degradation of training accuracy. ResNet uses building blocks to replace original convolution layers. Short connections are used in each block to pass through all information of the network input. Experiments on datasets, e.g. ImageNet and CIFAR-10, shows that ResNet are easy to optimize and can solve the degradation problem.

GANs. Goodfellow [13] proposed a novel framework that simultaneously trains two models via an adversarial process. The framework includes a generative model and a discriminative model. The generative model is like a team of counterfeiters that produces fake currency while the discriminant model is like police who detect the counterfeit currency. In the training process, the generative model tries to cheat the discriminant model and the discriminant tries to identify the fake. The competition between the two models drives them to improve their methods until the fake currency generated by the generative model can't be distinguished. Pix2pixGAN [14] is a conditional GAN requiring to input the real image into the generative model and discriminative model. It mixes the GAN objective with L1 distance to make the output near the ground truth output in an L1 sense. The generator of pix2pixGAN uses skip connections to share low-level information between input and output. Relying on an L1 term to force low-frequency correctness, the discriminator of pix2pixGAN uses PatchGAN that only penalizes small patches of an images.

3 Method

In this section, we draw on the idea from the machine learning algorithm Ensemble Learning to propose our GAN promotion method. We propose to use more than one generator to generate images to improve the performance of GAN, similar to the idea of using multiple classifiers to make decisions in the machine learning algorithm Boosting. We will start our research based on the current popular pix2pixGAN.

3.1 Multi-generator Training

Ensemble Learning [17] is a machine learning method that uses a series of classifiers to learn and uses a certain rule to integrate individual learning results to achieve better learning outcomes than a single classifier. Based on this idea, We propose to use multiple generators to improve the image generation quality of GAN. The basis for dividing multiple generators and how to use multiple generators are two issues that need to be addressed. For the first question, the first solution we think of is the semantic category. Obviously, we need to generate images in multiple scenes, and each scene has its own unique characteristics, so it is reasonable and simple to use different generators for different scenes according to semantics. For the second problem, one method that is very easy to solve is to use a classifier to classify the input and then pass the image to the appropriate generator. So as stated above, we propose our model as shown in Fig. 2. As the figure shows, the input image is first passed through a resnet-50 neural network for classification, and then the image is input to the corresponding generator for image generation and finally get the output.

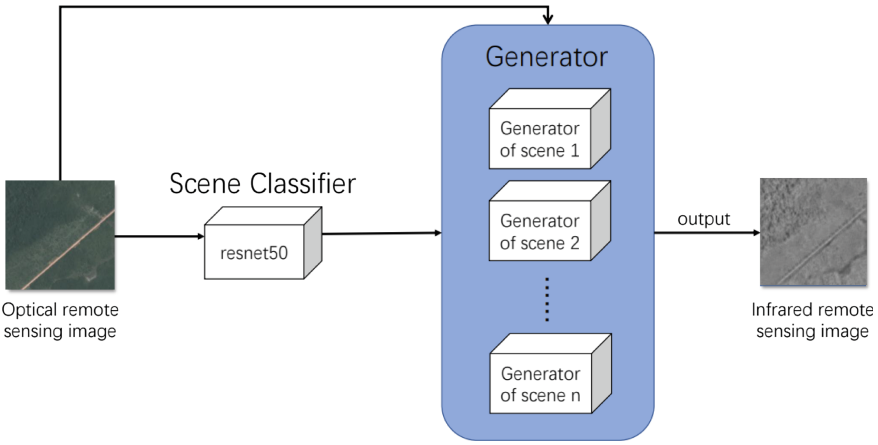

Fig. 2. The architecture of our model. The input image will be first inputted to the scene classifier and then go through the corresponding generator.

However, when a dataset has many different semantic classes and using a generator for each semantic category, the model structure will become very large. So in order to optimize our model, we have to mask some constraint on the semantics. One way to reduce semantic categories is cluster semantics because similar semantics have some common features which can be learned by a common generator. Another simpler approach is to train a generator to find the semantic categories that are difficult to train, ie, the generated categories with lower metrics, and then use separate generator training for difficult categories.

We recommend that the cluster method can be used when there are many semantic categories. Otherwise, the latter can be used when there are few semantic categories, which directly solves the problem of poor semantic generation. With reference to the idea of Ensemble Learning and solving difficulties separately, the quality of image generation of GAN can be improved by our multi-generator training method.

3.2 Objective

We use the loss function of conditional GAN to guide the training process. And the objective of a Conditional GAN in pix2pix [14] can be expressed as

$$L_{cGAN}(G, D) = E_{x,y}[\log D(x, y)] + E_{x,z}[\log(1 - D(x, G(x, z)))] \tag{1}$$

where z means a random noise, x and y represent conditional image and target image, respectively. Meanwhile, generator G tries to minimize this objective against an adversarial discriminator D that tries to maximize it.

For our method, let X and Y be two image domains (e.g., the optical and infrared image domains). And the training samples from domain X and Y are denote by $\{x^i\}_{i=1}^{M} \subset X$ and $\{y^i\}_{i=1}^{M} \subset Y$ respectively, where M is the number of semantics and i means the i^{th} semantic category; this shows that we will have M generators but only one discriminator for all inputs. Here $x^i \subset X$ will be inputted to corresponding generator G_i. G is expected to well learn the scheme of image style transfer and output the images from target domain Y by the feedback from D. So we propose our L_{GAN} objective as:

$$L_{GAN}(G, D) = E_Y[\log D(y^i)] + E_X[\log(1 - D(G_i(x^i)))] \tag{2}$$

According to [14], previous studies have found it beneficial to mix the GAN objective with a more traditional loss, such as L2 distance [18]. And [14] propose that using L1 distance rather than L2 as L1 encourages less blurring. Therefore, we introduce L1 loss as:

$$L_{l1}(G) = E_{x,y}[\left\|y^i - G_i(x^i)\right\|_1] \tag{3}$$

The full objective of our proposed method is:

$$L = L_{GAN}(G, D) + \lambda L_{l1}(G) \tag{4}$$

where λ is used to control the weights of L1 loss.

4 Implementation

In this work, we focus on converting optical remote sensing images to near infrared remote sensing images. we consider to classify the optical images into three categories, such as water, habitation and other. We simply divide our work into two parts for scene classification and image transformation.

We train a scene classification model by ResNet50 and image transformation model by pix2pixGAN.

For the scene classification model, we train it based on transfer learning. A ResNet50 model pretrained on ImageNet dataset is preferred. Pre-trained model can reduce training time and greatly improve the accuracy rate of training. There are two points that we have to pay attention to in the training process. The first one is that we have to preprocess images in our dataset. The input image size required for ResNet50 is 224×224 while the size of our images is 256×256. Thus, we have to crop, flip and normalize images before training. A 224×224 crop is centrally sampled from an image or its horizonal flip. The second one is that we have to change fully connected layers in the ResNet50 model to fit our classification model. Cause the number of categories in ImageNet dataset is 1000 and the number of categories in our dataset is 3. We use cross entropy loss function and SGD. The learning rate starts from 0.001 and is divided by 10 after several epoch. Each epoch contains training and validation. The model is set training mode in training and compute loss by the output of forward propagation and the real labels of images. Then the parameters in the model are updated by back propagation. The parameters of the model which gets the highest accuracy rate in validation are returned after the whole training process.

For training of the image transformation model, we have to concatenate the paired images into an 256×512 image. We alternately train the generator and the discriminator in training. The weight of L1 loss in the objective is 100. The models are trained using minibatch SGD and Adam [19]. The learning rate starts from 0.0002 and is linearly decayed after the first 100 epochs. Momentum parameters are set as $\beta_1 = 0.5$ and $\beta_2 = 0.999$.

5 Experiment

5.1 Dataset

A satellite imaging company called Planet recently released a dataset about remote sensing images of Amazon basin. The dataset is used in a Kaggle competition for labeling the ground feature [20]. The chips were derived from Planet's full-frame analytic scene products using 4-band satellites in sun-synchronous orbit and International Space Station orbit. Each image is 256×256 pixels and contains four bands of data: red, green, blue, and near infrared. In our experiment, we use the data which for training in the Kaggle competition to train our model. There are 40479 tiff image files which contains both optical information and near infrared information.

5.2 Experimental Results of Scene Classification

In this section, we simply divide the images into three categories: water, habitation and other based on their own labels. For each category, there are 2000 images for training. Here we use residual neural networks of 18, 34 and 50 layers

to train our scene classification models. Table 1 shows the accuracy rates of scene classification using different models and we can observe that ResNet50 produces the highest accuracy rate.

Table 1. The accuracy rates of different ResNet models.

Model	The accuracy rate
ResNet18	0.87
ResNet34	0.92
ResNet50	0.95

5.3 Experimental Results of Image Transformation

Qualitative Evaluation. Figure 1 shows the results of image transformation from optical remote sensing images to near infrared remote sensing images. Compared with the original pix2pixGAN model, the method that training models for each scene produces better results. This method allows models to learn more information of the scene feature. For example, the color of water in infrared images is black. As the Fig. 3 shows, when we train our models with scene information, the color of water is close to ground truth, and results are not well when we simply use pix2pixGAN to train our model.

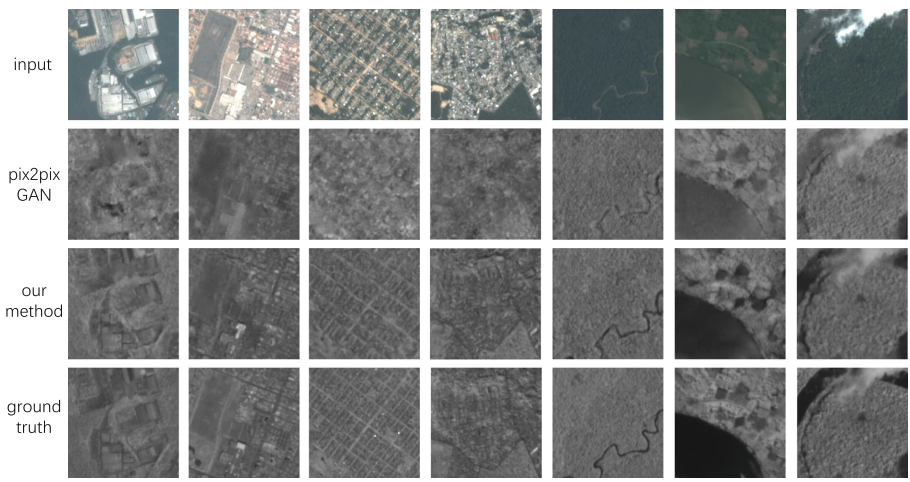

Fig. 3. Comparison of transformation between our method and pix2pixGAN.

Table 2. Evaluation on different image transformation models

Model	Category	PSNR	SSIM	Cosine similarity
pix2pix	water	24.5083	0.7448	0.9863
	habitation	23.5000	0.6996	0.9903
	other	24.8334	0.7258	0.9923
Our method	water	25.4383	**0.7580**	0.9866
	habitation	**26.6665**	0.7502	0.9906
	other	25.7735	0.7547	**0.9947**
pix2pix	overall	24.2805	0.7234	0.9897
Our method	**overall**	**25.9596**	**0.7543**	**0.9906**

Quantitative Evaluation. For quantitative evaluations, we perform to evaluate our experiments by using standard image quality assessment, e.g. PSNR, SSIM [21] and cosine similarity. The peak-signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) are widely used objective metrics due to their low complexity and clear physical meaning. The PSNR value approaches infinity as the MSE approaches zero. This shows that a higher PSNR value provides a higher image quality. Also means, a small value of the PSNR implies high numerical differences between images. The SSIM is a well-known quality metric used to measure the similarity between two images. And it is designed by modeling any image distortion as a combination of three factors that are loss of correlation, luminance distortion and contrast distortion [22]. The SSIM is considered to be correlated with the quality perception of the human visual system (HVS). Cosine similarity is a commonly used approach to match similar vector, with the advantages of simplicity and effectiveness.

Table 2 shows the results of our method and pix2pixGAN generating images of each category. Obviously, the results of our method have higher PSNR, SSIM and cosine similarity scores and our method performs better than the original pix2pixGAN model.

6 Conclusion

For the task that converts optical remote sensing images into infrared remote sensing images, our method is able to retain features of different scenes. By combining scene classification with image transformation, the models capture infrared characteristics of each scene perfectly and the generated infrared images are close to ground truth. The experimental results shows our method works well. Although our method can achieve better results, there are cases where the classification is wrong during experiments. In this case, the results are different from the real infrared images and that needs further study.

Acknowledgement. This work is partially supported by the National Natural Science Foundation of China (Grant no.61772568, Grant no.61603364), the Guangzhou Science and Technology Program (Grant no. 201804010288), and the Fundamental Research Funds for the Central Universities (Grant no.18lgzd15).

References

1. Rodhouse, K.N.: A comparison of near-infrared and visible imaging for surveillance applications (2012)
2. Siesler, H.W.: Near-Infrared Spectroscopy: Principles, Instruments, Applications. Wiley, Hoboken (2008)
3. Reich, G.: Near-infrared spectroscopy and imaging: basic principles and pharmaceutical applications. Adv. Drug Delivery Rev. **57**(8), 1109–1143 (2005)
4. Luo, X., Sun, J., Liu, J., Xia, J.: Realization of infrared image acquisition by inversion of visible light image. J. Infrared Laser Eng. (2008). (in Chinese)
5. Wu, G., Bai, T., Bai, F.: Infrared image inversion based on visible light image. J. Infrared Technol. **33**(10), 574 (2011). (in Chinese)
6. Li, M., Xu, Z., Xie, H., Xing, Y.: Infrared image generation method based on visible light image and its detail modulation. J. Infrared Technol. **40**(1), 34–38 (2018). (in Chinese)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
8. Szegedy C, Liu W, Jia Y.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9 (2015)
9. Simonyan, K, Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
10. He, K., Zhang, X., Ren, S.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778 (2016)
11. Zhu, J.Y., Park, T., Isola, P.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223-2232 (2017)
12. Choi, Y., Choi, M., Kim, M.: Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789-8797 (2018)
13. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
14. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134 (2017)
15. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: CVPR (2009)
16. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
17. Dietterich, T.G., et al.: Ensemble learning. In: The handbook of brain theory and neural networks(2), pp. 110–125 (2002)
18. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2536–2544 (2016)

19. Kingma, D.P., Ba, J..: Adam: a method for stochastic optimization. In: ICLR (2015)
20. Kaggle competition. planet: Understanding the amazon from space. https://www.kaggle.com/c/planet-understanding-the-amazon-from-space. Accessed 29 Apr 2019
21. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans. Image Process. **15**(11), 3440–3451 (2006)
22. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)