IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 31, 2020, accepted September 15, 2020, date of publication September 18, 2020, date of current version October 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024576

Sparse GANs for Thermal Infrared Image Generation From Optical Image

XIAOYAN QIAN, MIAO ZHANG, AND FENG ZHANG
College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 210000, China
Corresponding author: Xiaoyan Qian (qianxiaoyan@nuaa.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61803199.

ABSTRACT Thermal infrared (TIR) images are not influenced by the illumination variations and can be used in total darkness. With these advantages, TIR technology has a wide application in surveillance systems and various defense systems. However, there are not enough TIR images for wide range of application because the equipment for thermal infrared imaging is expensive and demands strict imaging conditions. To address this problem, we propose a sparse generative model based on pix2pix framework to produce synthetic TIR data from optical RGB images. Considering little texture and color information in TIR images, this model uses a U-net architecture but only selects partial low-level and high-level information for symmetric connections. Specially, we integrate intensity and gradient losses into the objective to train models, which assists generation models to learn more infrared images' characteristics. The experiments on public datasets prove that this proposed method can generate TIR data from optical images. Compared with current pix2pix networks, this method achieves increases by over 6.5% and over 1.2% separately on the metrics of SSIM and PSNR based on the public datasets. The SSIM value even gets an increase by 7% for daytime images. Meanwhile the network parameters decent by 13%.

INDEX TERMS Generative adversarial networks (GANs), image-to-image translation, style transfer, sparse U-net connections.

I. INTRODUCTION

Image-to-image translation based on generative adversarial networks (GANs) has received increasing attention. It converts an image from one representation of a given scene to another and has brought widespread applications such as image colorizing [1], [2], photo-realistic images from art paintings and reverse [3]–[5], aerial images to maps [6], daylight images to night images and reverse [7], and so on. Nevertheless, there is still little work for synthetic thermal infrared datasets from optical images. Optical images can present information similar to what the human could see but are sensitive to illumination variation, while TIR images can make up for this weakness by their thermal radiation differences. In certain fields such as military and environmental monitoring, TIR images are more useful than optical images because they are independent of the quality of the environment. Compared with capturing optical images, TIR facility is not only expensive but also demands strict testing conditions, which makes TIR images are not as available as visible images. To overcome these restrictions, this paper aims to apply GANs to construct TIR data from easily obtained optical RGB images. A supervised framework is established based on pix2pix GANs [6] for optical-to-TIR translations using labeled data.

Pix2pix framework is the first GAN-based image-to-image translation work with good performance to produce strong results in the unimodal image prediction setting when there is spatial correspondence between input and output pairs [6]. It contains an encoder-decoder architecture. During training, the learned encoder attempts to pass enough information to the generator to resolve any ambiguities regarding the output mode. Some generators select symmetric dense skip connections, which maintain both low-level and high-level information well. Our goal is to generate TIR data from optical domains. TIR images present rich geometric structure and material property by receiving thermal radiation, but have fewer details such as color and textures comparing with visible images. So dense skip connections are not needed for Optical-to-TIR images. This work therefore seeks

The associate editor coordinating the review of this manuscript and approving it for publication was Mingjun Dai.

a new generator architecture with sparse skip connections. Partial low and high layers are symmetrically connected. This model reduces both modal parameters and redundant details. To train a robust image-to-image translation system, many efforts have been made. The straightforward approach is to constrain the network training by an objective. Some networks minimize the L2 distance between predicted and ground truth pixels [8]. This tends to result in blurry problem because this distance averages all plausible outputs. To get clear structure and vivid details, some networks select L1 loss on pixel-wise space [9]. Except for GAN loss, perceptual loss has been used in image-to-image translation tasks [10]. Although we have some kinds of losses to evaluate the difference between ground truth and synthetic image, most of current networks aim to produce vivid optical results. To get realistic TIR images, this work tries to apply constraints into the object that are applicable to TIR data. Gradient and intensity losses between ground truth and generated image are added to the objective then SGD is used to optimize the network parameters. Using this proposed strategy, we can synthesize high-quality TIR images. In summary, the main contributions for the style transfer problem in this paper are:

- We propose a sparse "U-net" GAN network that strikes a better balance between style and content. The sparse skip connections have fewer modal parameters and reduce redundant details that are enough to achieve good TIR data.
- We introduce new metrics in terms of content and style. The content similarity is evaluated by gradient loss and thermal radiation difference is restricted by the L1 intensity loss. These two losses make the synthetic images more similar to the ground-truth TIR. The comparison results show that the optimization for current object brings better TIR data.
- We perform extensive and detailed experiments to verify the performance of the presented strategies. Our proposed model performs better than current state-of-the-art models on the public dataset with less network parameters.

II. RELATED WORK

Current researches on converting optical images into TIR data can be divided into two parts: one is utilizing manual ways and hand-crafted features to produce TIR style. The other uses deep learning networks to learn the mapping between different styles without manual intervention. Luo *et al.* [11] present a physical model to simulate the different thermal radiation features among different targets. They divide a gray image into small parts and then manually segment the target object from the background. After that, each area is set an infrared radiation value manually. Wu *et al.* [12] use the histogram difference between optical and infrared image pairs to convert optical images to infrared images. Li *et al.* [13] proposed a neural network-based infrared image generation method to predict the temperature of the target with different materials, but manually segmentation is needed. The above methods can consider rich physical and low-level features of different targets, but the segmentation of image has much difficulty in processing large quantities of images.

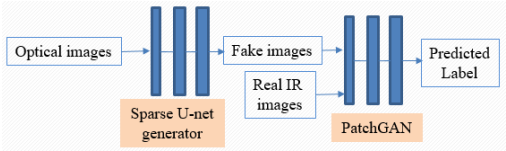


FIGURE 1. Model framework.

On the other hand, GANs have achieved promising results in image style transfer. This framework is good at capturing data distribution by learning from a big dataset. Mirza *et al.* propose Conditional GANs (CGANs) [14] to learn mapping information between inputs and outputs using a conditional convolutional neural networks. Isola *et al.* propose pix2pixGAN [6] based on CGANs. Pix2pixGAN mixes the GAN objective with L1 distance to make the output near the ground truth. Zhang *et al.* [15] use this framework to translate the labeled RGB data to TIR data. To present different radiation features for different scene, Li *et al.* [16] integrate extra scene classification network into multi-branch generators based on pix2pix network. Although this method can achieve good results, the training and testing are time-consuming. Sometimes there are wrong classification results in unreal IR images. As pix2pix needs paired training data, which are usually costly to obtain, Zhu *et al.* present CycleGAN [17], an unsupervised image-to-image translation. This framework learns the mapping between two unpaired image domains with the aid of a cycle-consistency loss. Apart from CycleGAN, many other GAN variants [18]–[20] have been proposed to tackle the cross-domain problem. Zhang *et al.* [15] also use unpaired image-to-image framework to help generate TIR data, but sometimes get worse results compared with supervised GAN network since unsupervised models are easily affected by unpredictable contend during the translation stage.

In this paper, we also use pix2pix GANs to realize optical-to-TIR style translation since it is a powerful framework for image-to-image translation. With the goal of reducing the network complexity and considering the infrared characteristics, we modify the network structure to ignore partial low-level details and add content losses into the objective.

III. THE PROPOSED METHOD

Based on pix2pix GANs, our model (seen in Fig.1) uses a "U-net" architecture [21] as the generator G. Similar to [22], we use the strided convolutional layers take the place of pooling layers because they will cause the loss of useful information during multidimensional reduction. The discriminator D is a "PatchGAN" classification by capturing local style statistics. The two deep networks compete against each other. The generator tries to generate samples that resemble the real TIR images, whereas the discriminator tries to detect whether samples generated by G are real.

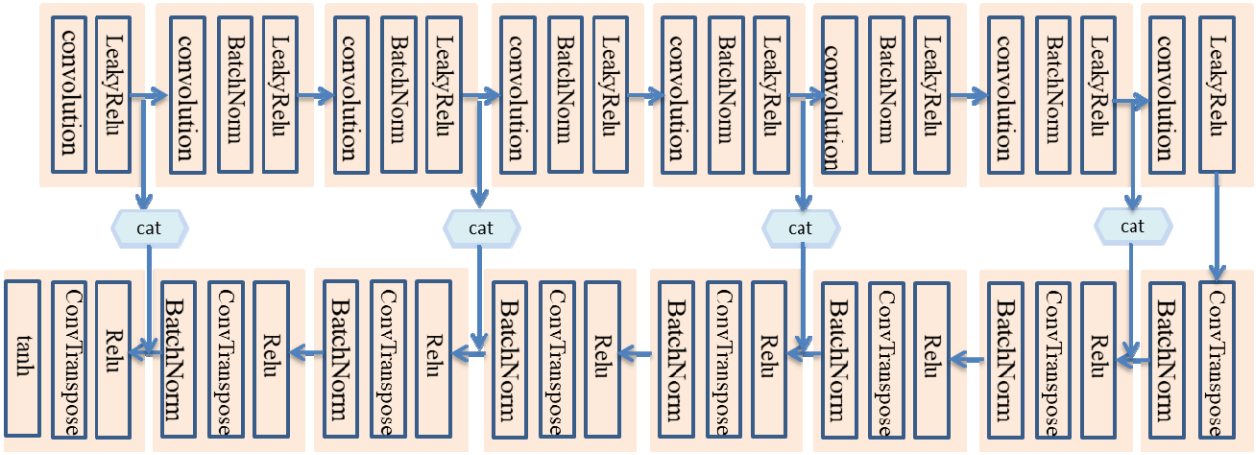


FIGURE 2. The framework for the generator.

A. SPARSE U-NET GENERATOR

The generator G contains an encoder-decoder architecture with symmetric skip connections. Since TIR images have less color and texture information, current dense connections in the "U-net" framework includes much redundancy and too many parameters. In this paper, we use sparse connections to simplify this model. Fig.2 shows the model framework. It includes 16 sub-models and four pairs of symmetrical connections between the first and fifteenth, third and thirteenth, fifth and eleventh, seventh and ninth layers.

As the usual practices of deep learning strategies [22], encoders have eight convolutional layers with a 4×4 kernel and each has an activation layer. The stride for all the convolutional layers is set as 2 and paddings are 1 to make input and output have the same size. In the encoders, we select LeakyRelu as the activation function to avoid dead nerve cells and its non-zero slope $a = 0.2$. The decoders have eight transposed convolutional layers with the kernel size 4×4. We choose ReLu as the activation function to enhance the non-linearity of this model and speed up convergence rate. In this way, the number of network parameters in our sparse generator decreases to 47M from 54M in the dense U-net connections (descending by about 13%).

B. ARCHITECTURE OF DISCRIMINATOR

It is well known that the L1 loss produces blurry results on image generation problems [23]. To model high-frequency crispness, we leverage PatchGANs for the discriminator network, which focuses on penalizing the structure in local image patches. Fig.3 shows the architecture of our discriminator. It includes 5 blocks with 3 channels input. The convolutional layers separately include 64-128-256-512-1filters. As the generator, the convolutional kernel is 4×4. The size of stride is set as 2 from the first to third layers and as 1 at the fourth and fifth layers, padding value is 1. Each block except for the fifth has an activation layer of Leaky Relu to ensure the parameters get enough update. In this work, the discriminator classifies whether local image patches with the size 70×70 in the synthetic TIR image are real or fake by averaging all convolutional responses to provide the ultimate 1D output of this discriminator.

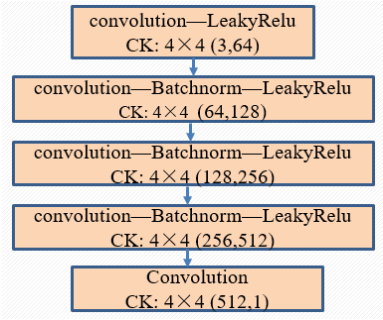


FIGURE 3. The discriminator architecture.

C. OBJECTIVE OPTIMIZATION

We use the loss function in [6], [15] to guide the training process, which consists of two parts expressed as:

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

Here, x,y and z represent the optical image, ground-truth TIR image and a random noise respectively. G tries to minimize this objective while D is to maximize it.

Additionally, pix2pix model includes a traditional construction loss, L1 distance, to ensure the quality of generated synthetic images:

$$L_{L1}(G) = E_{x,y}[||y - G(x, z)||_1] \tag{2}$$

Therefore, the full objective function is:

$$G^* = \arg \min_G \max_D L_{GAN}(G, D) + \lambda L_{L1}(G) \tag{3}$$

where λ is a weight to control the contribution of L1 loss.

Since TIR images capture the information of ground objects by their thermal radiation difference, their intensity

TABLE 1. Current datasets.

	Datasets	Number of images
Paired data	CVC-14	8,473
	KAIST	50,184
	OSU	8,544
	VAP Trimodal	5,927
	LITIV2012	6,325

can simulate the received radiation. To learn this characteristic, we add intensity loss into current objective. The intensity I can be calculated by:

$$I = 0.299R + 0.587G + 0.114B \tag{4}$$

Here, R, G and B are the color values of the image. The intensity L1 loss between the synthetic TIR and ground-truth is described as:

$$L_{L1}(I) = E_{x,y,z}[||I(y) - I(G(x,z))||_1] \tag{5}$$

In addition, in order to retain the semantic information, we propose an adversarial content loss based on the gradient information, which favors maintaining the local appearance and shape of different objects. We also formulate the gradient loss as L1 distance to minimize the difference between the generated and input TIR image:

$$L_{L1}(GRA) = E_{x,y,z}[||GRA(y) - GRA(G(x,z))||_1] \tag{6}$$

Here $GRA(y)$ and $GRA(G(x,z)$ are the gradient maps for the ground truth and the synthetic TIR image separately.

We define the final objective function as:

$$G = G^* + \lambda_{gra} L_{L1}(GRA) + \lambda_I L_{L1}(I) \tag{7}$$

where λgra and λ_I are also the weights to control the relative importance of each loss.

IV. EXPERIMENTS

A. IMPLIMENTATION DETAILS

1) DATASETS

The same as classical pix2pix model, we need paired optical and TIR images for training. TABLE 1 lists all current data sets. Among these, KAIST data set contains many multi-spectral pedestrians' data under dynamic environments and variable illumination. While other data sets present the moving objects captured by static cameras at intersections, in indoor scenes with controlled lighting or in different planes.

Here we conduct extensive experiments on the KAIST multi-spectral pedestrian data set and other data set. 64,447-paired images are chosen randomly from these data sets to train our model and the remaining optical images are used to test the GAN network. There are no intersections between training and testing sets.

TABLE 2. The quantitative comparison of different skip connections.

	SSIM	PSNR(/dB)
1st,3rd	0.636	28.77
5th,7th	0.659	28.88
our model	0.664	29.00

2) TRAINING DETAILS

We end-to-end train the generator network and discriminator network in an adversarial manner. We alternately train the generator and the discriminator, one step on D, then one step on G. The discriminator D ensures synthetic TIR images generated by the generator can be distinguished from the ground-truth TIR image. It tries to maximize $L_{GAN(G,D)}$. During training discriminator, we also aim to maximize loss: logD(x,G(x,z)) rather than minimizing log(1−D(x,G(x,z)) brought by the generator. The generator G is trained to minimize the adversarial loss to make the synthetic image more plausible. The generator also tries to minimize the weighted loss: $\lambda L_{L1}(G) + \lambda_{\text{gra}} L_{L1}(GRA) + \lambda_1 L_{L1}(I)$ ($\lambda = 100$, $\lambda_{gra} = 100$ and $\lambda_I = 30$) by the output of forward propagation and the real labels of images. Then the model parameters are updated by back propagation with Stochastic Gradient Descent (SGD).

The model is trained with batch-size 3. A smaller learning rate is more suitable for TIR data as they have less detailed information than optical RGB images. Here, the learning rate for mini-batches is set as 0.0002. We don't use decay during training because the model does not over fit within two epochs. Momentum parameters are set as $\beta 1 = 0.5$, $\beta 2 = 0.999$.

B. EXPERIMENTAL RESULTS

1) QUALITATIVE EVALUATION

To verify the performance of our method, we first visually experiment with three base generator networks. Fig.4 presents the comparison results of image transformation from optical images to TIR data. They are produced by the generators separately with symmetrical low-level, high-level and the proposed skip connections. That shows generator with connections at the first and third layers or at the fifth and seventh layers can only produce blurred objects such as cars and trees, which sometimes cannot be distinguished from the background. The proposed method enhances the semantic features by using both low-level and high-level connections. The shapes and contours are much clearer and synthetic images are more similar to the ground truth TIR images. TABLE 2 shows using both low-level and high-level features brings highest SSIM and PSNR values.

To describe the details information in the synthetic TIR images produced by our model with sparse skip connections and the model with dense skip connections, Table 3 lists the average LBP values (local binary pattern, it's often used to describe the local texture information [24]) for synthetic TIR images in Fig.5. The TIR results of Fig.5 (a) from the left to

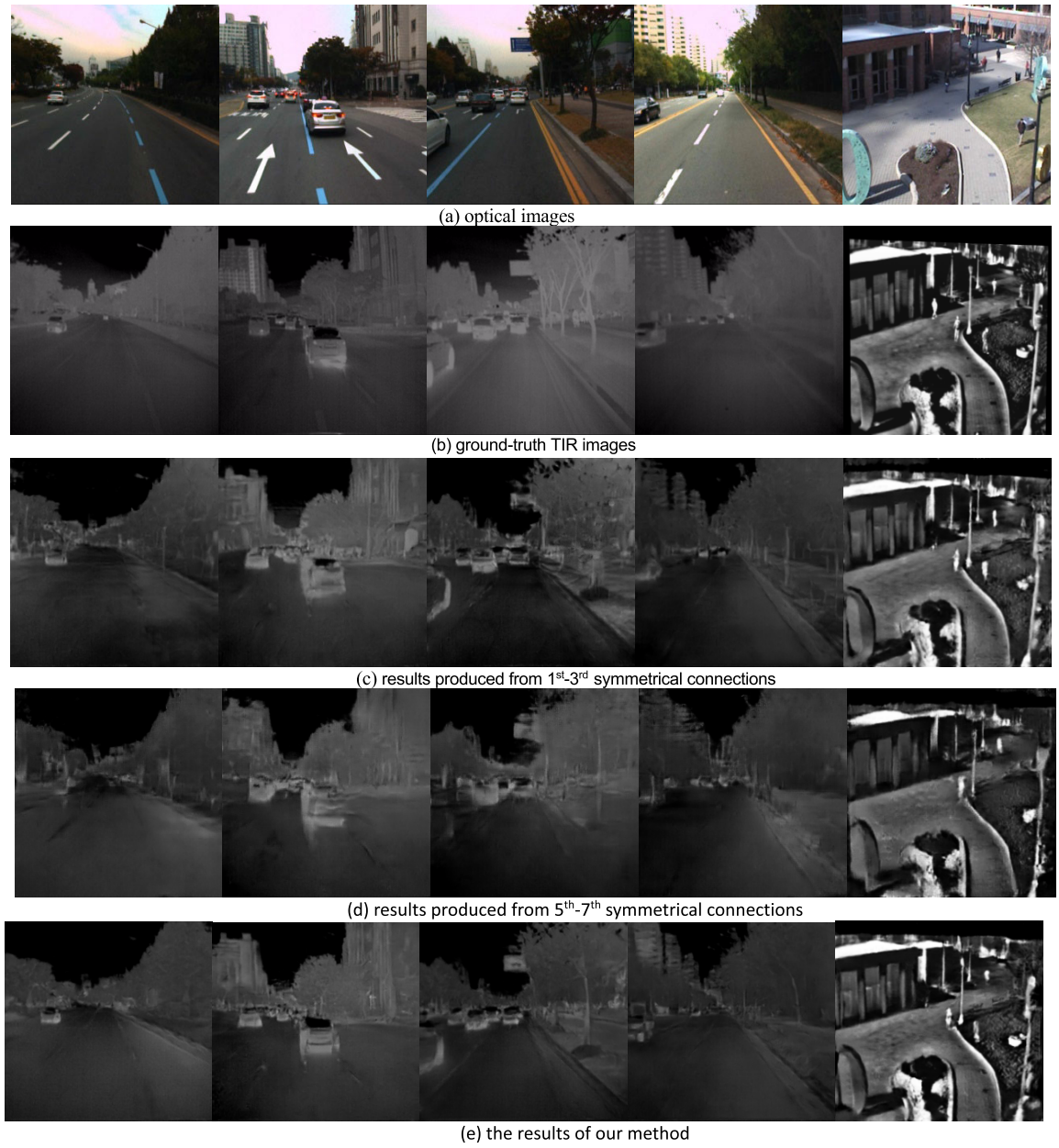


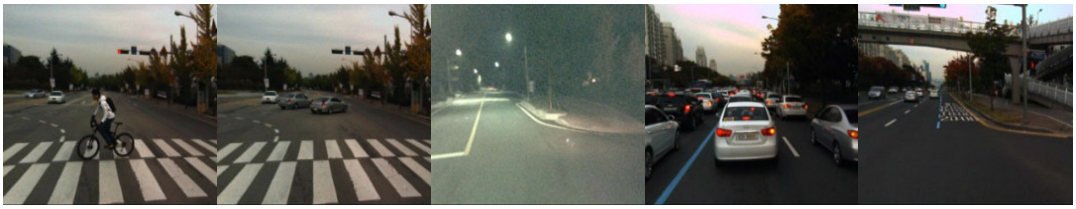
FIGURE 4. Synthetic results with different-layer skip connections.

TABLE 3. The LBP value of dense and our models.

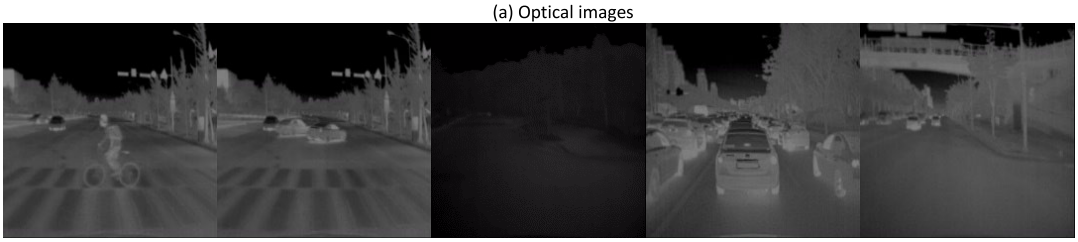
	Fig5-1	Fig.5-2	Fig5-3	Fig5-4	Fig5-5	Fig5-6
ours	155.85	155.05	152.78	168.65	148.30	149.49
dense	162.67	162.25	156.44	163.48	144.96	149.97

the right are named as Fig.5-1 to Fig.5-6. The second line is the LBP values for Fig.5 (g) from our model and the third line is those for Fig.5(c) brought by the dense model. It can be seen that our model has fewer details in Fig.5-1,2,3,6. However, Fig5-4 and Fig.5-5 have bigger LBP values. That is because our model pops out the objects better in the environment.

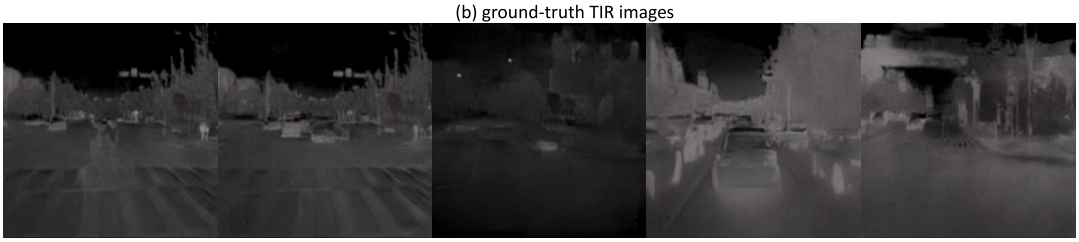
To provide a more detailed analysis of the proposed method, then we present the performance comparison of different GANs models including pix2pix network just used in [15], unsupervised CycleGAN network [17] and two baselines that ablate $L_{L1}(GRA)$, $L_{L1}(I)$ respectively. The model in [15] even fails to generate distinguishable outputs. The riding man is blurry and cars cannot be separable. Without $L_{L1}(GRA)$ or $L_{L1}(I)$, the image quality is unsatisfactory. When training with unpaired images, CycleGAN can't well transfer the characteristics of the thermal targets. Fig.5 (d) shows that riding man and cars present black color. We observed in Fig.5 (e) that the synthetic TIR images are visually realistic with a higher quality of translation results generated from out model. $L_{L1}(GRA)$ and $L_{L1}(I)$



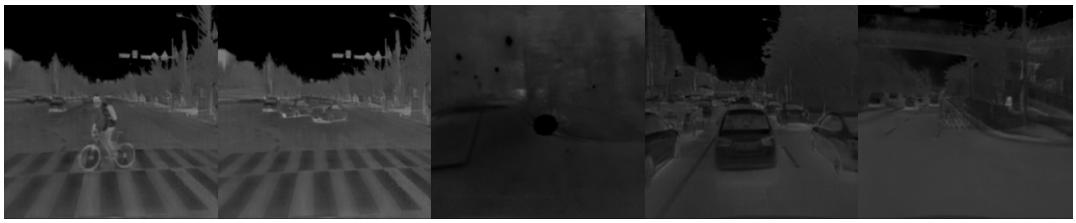
(a) Optical images



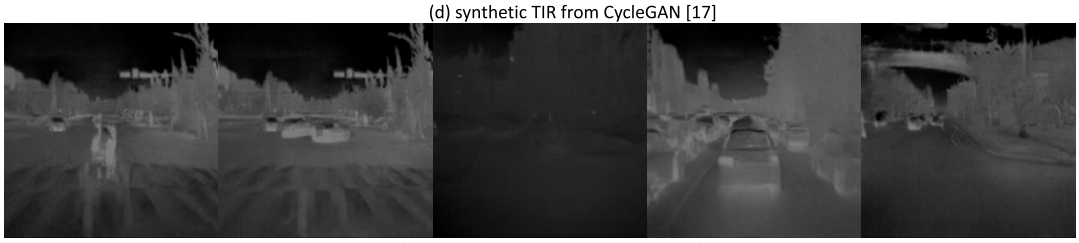
(b) ground-truth TIR images



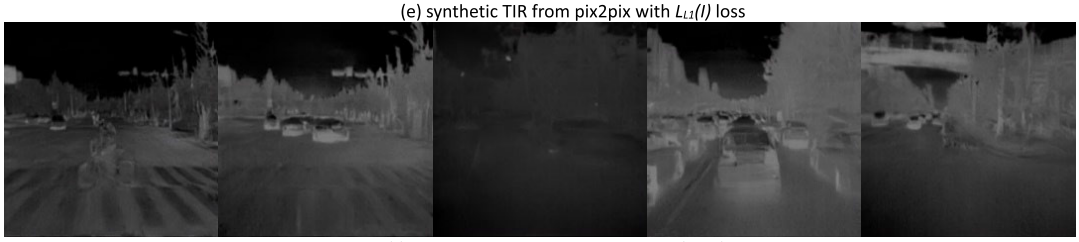
(c) synthetic TIR from pix2pix network [15]



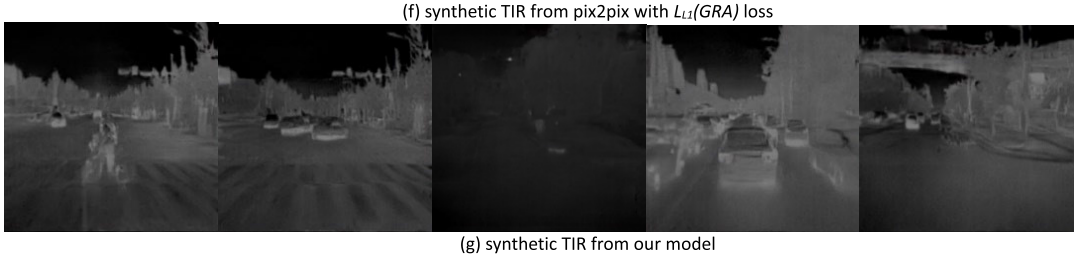
(d) synthetic TIR from CycleGAN [17]



(e) synthetic TIR from pix2pix with $L_{L1}(I)$ loss



(f) synthetic TIR from pix2pix with $L_{L1}(GRA)$ loss



(g) synthetic TIR from our model

FIGURE 5. **Synthetic results from different models.**

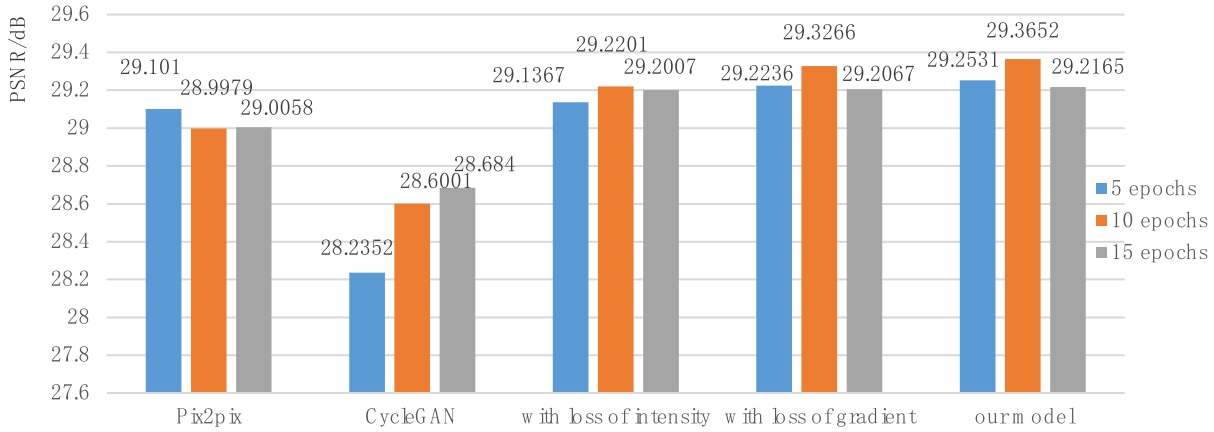


FIGURE 6. PSNR values brought by 5, 10 and 15 training epochs for different models.

TABLE 4. The synthetic results under daytime and night.

	environment	PSNR/dB	SSIM
Pix2pix [15]	daytime	29.2238	0.7480
	night	28.9935	0.6918
CycleGAN [17]	daytime	28.9340	0.6807
	night	28.7234	0.6744
Pix2pix+ $L_{L1}(I)$ loss	daytime	29.5015	0.7633
	night	29.0131	0.6981
Pix2pix+ $L_{L1}(GRA)$ loss	daytime	29.5069	0.7968
	night	29.1939	0.6991
Our model	daytime	29.5331	0.8011
	night	29.1821	0.7012

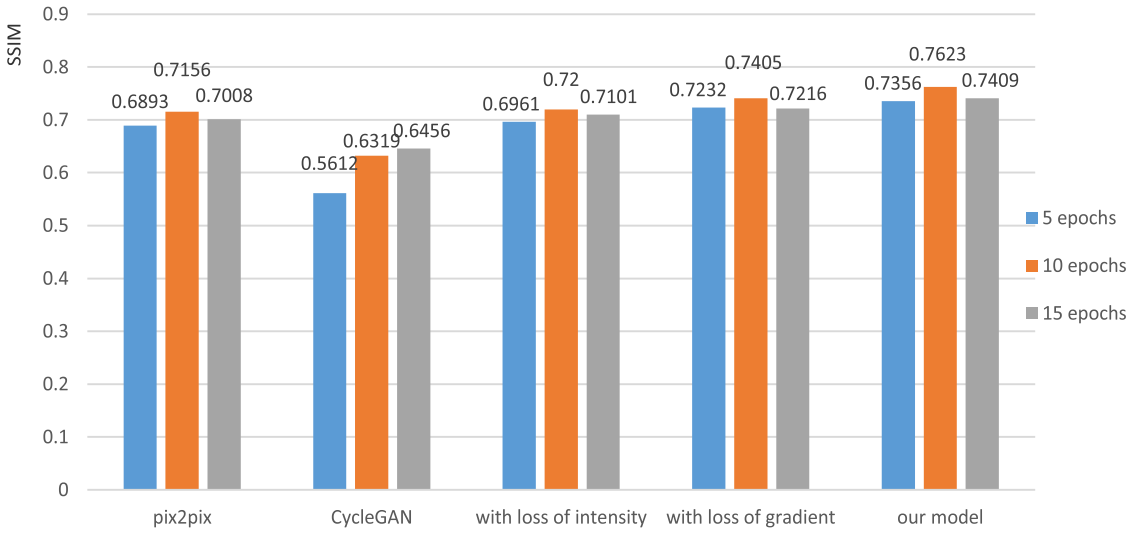


FIGURE 7. SSIM values brought by 5, 10 and 15 training epochs for different models.

losses allow our model to learn reliable features with different attribute values universally applicable to produce real TIR images.

2) QUANTITATIVE EVALUATION

It is necessary to measure quantitatively the difference between the translated images and true ones. All quantitative

results based on the test dataset, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [25], are shown in Figure 6-7 and Table 4. Figure 6 and 7 show the detailed PSNR and SSIM scores of different models under 5, 10 and 15 training epochs. The model using source pix2pix network [15] gets about 29dB PSNR. After adding $L_{L1}(GRA)$ and $L_{L1}(I)$ losses, PSNR has slight improvement. In contrast, the CycleGAN model with unpaired images [16] has lowest PSNR. While the bar graphs in Figure 7 show each model except CycleGAN has the highest SSIM values with the 10-epoch-trained networks. Our model achieves an increase by 6.5%∼10.5% on the metric SSIM compared with the source pix2pix model [15] and by 3%∼6% after adding $L_{L1}(GRA)$ or $L_{L1}(I)$ losses. Though CycleGAN [17] can get higher and higher SSIM values with the increasing epochs, it is time-consuming and has no better results than the others have. Our proposed model enhances the semantic consistency using intensity and gradient constraints that encourage appearance and structural similarity. The sparse connections not only maintain low-level and high-level features but also reduce network's parameters. This simplification is more applicable to TIR images.

In addition, we can find in TABLE 4 that the synthetic daytime TIR data have better qualities than those at night resulted from any GANs model. The SSIM increase for daytime synthetic TIR images reaches 7% while only 1.4% for the night images when comparing with the dense model [15]. That is because lower temperature at night cause weak thermal radiation difference. Indistinguishable objects tend to produce low-quality synthetic results.

V. CONCLUSION

In this work, we propose an image-to-image translation model to learn the mapping from the optical images to TIR domains. We design a sparse U-net architecture to learn the low-level and high-level features with less networks' parameters. To optimize this deep network, we introduce intensity and gradient losses into the objective to improve the semantic and appearance consistency during translation. The experimental results on the current dataset show our method works well. Although our method can bring better results, there are cases where the synthetic TIR data at night are different from the real infrared images and that needs further study.

REFERENCES

[1] R. Zhang, C. Mu, M. Xu, L. Xu, Q. Shi, and J. Wang, "Synthetic IR image refinement using adversarial learning with bidirectional mappings," *IEEE Access*, vol. 7, pp. 153734–153750, 2019.

[2] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708.

[3] W. Cho, S. Choi, D. K. Park, I. Shin, and J. Choo, "Image-to-image translation via group-wise deep whitening-and-coloring transformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10639–10647.

[4] M. Amodio and S. Krishnaswamy, "TraVeLGAN: Image-to-image translation by transformation vector learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8983–8992.

[5] M. Tomei, M. Cornia, L. Baraldi, and R. Cucchiara, "Art2Real: Unfolding the reality of artworks via semantically-aware image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5849–5859.

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[7] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. IEEE Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 35–51.

[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[10] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," in *Proc. IJCAI*, 2017.

[11] X. Luo, J. Sun, J. Liu, and J. Xia, "Realization of infrared image acquisition by inversion of visible light image," *J. Infrared Laser Eng.*, 2008.

[12] G. Wu, T. Bai, and F. Bai, "Infrared image inversion based on visible light image," *J. Infr. Technol*, vol. 33, no. 10, p. 574, 2011.

[13] M. Li, Z. Xu, H. Xie, and Y. Xing, "Infrared image generation method based on visible light image and its detail modulation," *J. Infr. Technol*, vol. 40, no. 1, pp. 34–38 ,2018.

[14] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[15] L. Zhang, A. Gonzalez-Garcia, J. van de Weijer, M. Danelljan, and F. S. Khan, "Synthetic data generation for end-to-end thermal infrared tracking," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1837–1850, Apr. 2019.

[16] L. Li, P. F. Li, M. Yang, and S. Gao, "Multi-branch semantic GAN for infrared image generation from optical image," in *Proc. IScIDE*, 2019, pp. 484–494.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[18] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[19] A. Anoosheh, E. Agustsson, R. Timofte, and L. Van Gool, "ComboGAN: Unrestrained scalability for image domain translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 896–897.

[20] H. Tang, D. Xu, W. Wang, Y. Yan, and N. Sebe, "Dual generator generative adversarial networks for multi-domain image-to-image translation," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2018, pp. 3–21.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.

[22] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *Comput. Sci.*, 2015.

[23] A. B. L. Larsen, S. K. Sonderby, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1558–1566.

[24] T. Ojala, M. Pietikinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognit.*, vol. 29, pp. 51–59, Jan. 1996, doi: 10.1016/0031-3203(95)00067-4.

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

XIAOYAN QIAN received the B.S. and Ph.D. degrees in computer science from the Nanjing University of Science and Technology, China. She is currently an Associate Professor with the Nanjing University of Aeronautics and Astronautics. She has published more than 40 journal articles and one book. Her current research interests include deep learning and intelligent surveillance.

MIAO ZHANG received the B.S. degree in traffic information engineer from the Nanjing University of Aeronautics and Astronautics, China, in 2018, where she is currently pursuing the M.S. degree. Her current research interests include deep learning and image processing.

FENG ZHANG received the B.S. degree in traffic information engineer from the Nanjing University of Aeronautics and Astronautics, China, in 2018, where he is currently pursuing the M.S. degree. His current research interests include deep learning and image processing.