Infrared Image Generation By Pix2pix Based on Multi-receptive Field Feature Fusion

Yangyang Ma
School of Artificial Intelligence and Automation, Huazhong University of Science & Technology
Wuhan, China
430074
myyhust@163.com

Yanling Hua
School of Artificial Intelligence and Automation, Huazhong University of Science & Technology
Wuhan, China
430074
h876477730@126.com

Zhengrong Zuo
School of Artificial Intelligence and Automation, Huazhong University of Science & Technology
Wuhan, China
430074
zhrzuo@ hust.edu.cn

***Abstract*—Infrared imaging has the advantages of strong anti-interference capability, long-range imaging, and night imaging, and has important applications in both civilian and military fields. In the development of infrared-related equipment, a large number of infrared images under a variety of conditions are required as verification test data. The field test of infrared images requires huge manpower and material resources, and it is difficult to obtain full-time infrared images. To address the problem of insufficient infrared image samples, the paper introduces generative adversarial networks into the infrared image generation task and investigates the infrared image generation method based on visible images by applying Pix2pix networks to paired visible infrared image datasets. To address the problem of missing detailed information of infrared images generated by the Pix2pix network, the paper proposes a Pix2pix network based on multi-receptive field feature fusion and constructs a multi-receptive field feature extractor based on Unet++ structure; the multi-receptive field feature fusion mechanism of nested pixel level by level is proposed. Experiments show that the Pix2pix network based on multi-receptive field feature fusion achieves finer infrared texture generation.**

Keywords—Generative Adversarial Networks, infrared image generation, multi-receptive-field feature fusion

I. INTRODUCTION

During the development of infrared-related devices and research on infrared imaging technology, a large amount of infrared image data under various conditions is needed as material for validation tests. Deep network-based algorithms also need a large number of infrared images for algorithm training, testing, and validation. However, in most of the scenarios, it is very difficult to acquire infrared images using real images, such as satellite infrared images or military target infrared images, which require a lot of human and material resources[1], and it is impractical to acquire various materials in all background infrared images at the same time. Therefore the use of infrared simulation techniques to generate infrared images has important research implications.

In order to achieve the fast and efficient generation of a large amount of infrared image data, this paper addresses the generation of heterogenous infrared images in infrared simulation. Since the technology of visible imager is mature and the data is easy and low cost to obtain, the problem of difficulty in obtaining infrared images can be alleviated if the easily available visible data can be used to generate infrared images.

There are two main ways of generating infrared images in the mainstream. One way is to analyze and model the target scene, and calculate the infrared radiation of the material according to the theory of infrared radiation and combining with atmospheric transmission parameters, and add sensor imaging effects to the infrared radiation and then grayscale it to get the infrared simulation image[2]. However, this method has problems such as low simulation degree of target temperature model and tedious processing, which is not suitable for rapid generation of a large number of IR images; another method is based on generative adversarial networks to achieve style conversion. Jiangrong Xie et al. proposed to use DCGAN networks to achieve the butrandom generation of some simple IR target images[3], Yunfei Feng for IR band expansion task To study, CycleGAN network was used to realize the conversion of mid-wave infrared images to long-wave infrared images[4]. The texture conversion from real infrared images is more consistent with the distribution characteristics of textures than modulation by visible images, and the textures generated using GAN are also smoother than the simulation effect. Therefore, this paper will investigate how to use generative adversarial networks for infrared image generation research.

In this paper, the heterogenous infrared image generation problem of generating infrared images from visible images is modeled as an image translation problem between two image domains. Among the derivative algorithms of GAN networks in recent years, the algorithms mainly address the diversity of generated images, unpaired image translation, and high-resolution image generation, while this paper focuses on the accuracy of generating infrared images from visible images, based on which the Pix2pix network proposed by Isola at CVPR 2017 is chosen[5]. the Pix2pix network is a general network architecture that uses the structure of conditional generative adversarial networks (CGAN) to solve image translation tasks[6]. The network is not only able to learn image-to-image mapping relationships but also to learn loss functions for specific translation tasks based on sample data features, making conditional generative adversarial networks applicable to image translation tasks.

The conditional generative adversarial network is the first time to introduce conditional labeling into the structure of a generative adversarial network, which somewhat improves the drawback of the original generative adversarial network that the generation pattern is too free due to any

unconstrained. Unlike discrete labeled, text-based conditional adversarial networks, Pix2pix focuses on the task of image-to-image translation with the aim of mapping high-resolution inputs to high-resolution output results, so Pix2pix uses images as conditional labels. Considering that the structures of the image pairs involved in the mapping are roughly aligned, the generator uses a multi-receptive field feature extractor with Unet++ structure which is similar in structure to an encoding-decoding network[7]. Ints input image is first passed through the encoding part of the network to reduce the resolution of the feature map, and later the decoding network is used to restore it to the original image resolution, while using cross-layer channel connections to enable the higher-level feature map to perceive the lower-level feature map The decoding network is then used to restore the original image resolution, while the cross-layer channel connections are used to enable the higher-level feature maps to perceive the detailed information of the lower-level feature maps. For the discriminator structure, the discriminator network of the original GAN is not suitable for image translation tasks requiring high resolution and high detail fidelity, and Pix2pix adopts the PatchGAN discriminator structure. The output of the conventional GAN network is a Boolean value of 'True' or 'Flase', which represents the determination of the whole image for the discriminator input, while the output of the PatchGAN discriminator is an N × N matrix, where each pixel in the matrix represents the determination of a block of M×M size corresponding to the input image perception field, which makes the discriminator focus more on the recovery of local detail information.

In this paper, we study an infrared image generation method based on generative adversarial networks, using Pix2pix networks to automatically learn the mapping relationship from visible images to infrared images. For the problem of missing detail information of IR images generated by the Pix2pix network, we propose a Pix2pix network based on multi-receptive field feature fusion to further optimize the detail information of the generated IR images and improve the realism of the simulated images.

In the task of heterogenous IR image generation, the algorithm has to ensure the accuracy of IR images generated from visible images and generate finer IR texture information as much as possible. pix2pix can basically achieve the task of generating IR images from visible images, and the generated IR image distribution has high consistency with the real IR images, but there are still samples of conversion failure. Due to the different imaging mechanisms, there is a big difference in the distribution between visible and infrared images of the same scene. Visible images have more distinct color features and clearer texture information, while infrared images have only one channel after grayscale quantization, and most scenes have more blurred edge information than visible images. In this paper, two contributions are made to address the problem of missing detail information in infrared images generated by the Pix2pix network as follows:

(1) Multi-receptive field feature extractor. In order to enhance the feature utilization rate of smaller receptive field features, this paper builds a multi-receptive field feature extractor based on the general Unet network[8]. Unet extracts the image features of different receptive fields with the same resolution, by borrowing the model structure of Unet++ network.

(2) Multi-receptive field feature fusion mechanism. Different from the widely used feature fusion method of merging on channels, this paper achieves multi-receptive field feature fusion by nesting multiple attention mechanism modules level by level and learning the weights of different receptive field features at a pixel level.

II. Related Work

A. Infrared imaging influencing factors

Infrared radiation exists in all corners of the world, and all objects in nature with temperatures above absolute zero emit infrared radiation, and the higher the surface temperature, the stronger the infrared radiation produced. The imaging effect of infrared images is mainly related to the scene temperature, infrared imaging equipment wavelength range, atmospheric transmission medium. Next, we will analyze the effect of temperature and wavelength on the infrared imaging effect.

Figure 1 plots the spectral irradiance of the blackbody versus temperature and wavelength. From the figure, we can see that the blackbody irradiance changes with wavelength at different temperatures with roughly the same trend, rising sharply first, reaching the peak gradually decreasing, and the peak wavelength decreases gradually as the temperature increases. According to Wien's displacement law, the wavelength corresponding to the peak of blackbody irradiance is inversely proportional to the blackbody temperature[9]. To calculate the integral of wavelength 0 to infinity, the relationship between blackbody irradiance and temperature can be obtained, which is Boltzmann's law, as shown in Eq：

$$M_0(T) = \sigma T^4 \tag{1}$$

where $\sigma = 5.6694 \times 10^{-8} W/(m^2 \cdot k^4)$ .Boltzmann's law shows that the blackbody irradiance is proportional to the fourth power of the temperature.

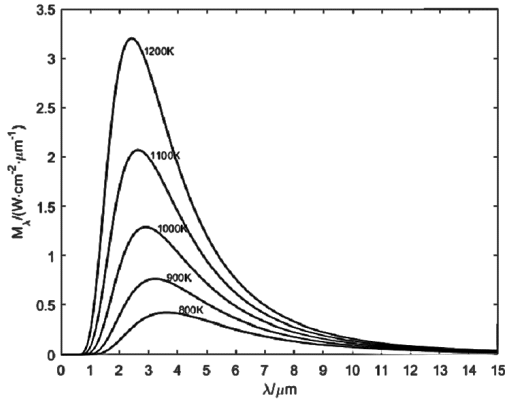

Fig.1. Blackbody irradiance as a function of temperature and wavelength

B. Generating Adversarial Network Foundation

Generative adversarial network (GAN) is a generative model based on the idea of a zero-sum game. Unlike the traditional network model, the GAN network consists of two sub-networks, the generative network, and the discriminative network, and its model structure is shown in Figure 2. The goal of the generative network is to generate

samples that match the real data distribution and apply the learned mapping function to the given noise. The goal of the discriminator is to be able to accurately identify the real samples from the real data distribution and the generated samples generated by the generator. During training, the generator gets feedback from the discriminator's decisions to further optimize the model parameters and learn how to better deceive the discriminator next time. Assuming that the real samples x conforms to the real data distribution $P_r(x)$ and the hidden vectors z conforms to the defined implicit vector distribution $P_z(z)$, such as uniform distribution or spherical Gaussian score The vector z is obtained by sampling the $P_z(z)$ distribution and then input to the generative network to generate $x'=G(z)$, and the generated data x' and the real data x are input to the discriminator respectively, and we expect the discriminator to have the accurate discriminatory ability to output the corresponding 0 (Fake) and 1 (Real)[10]. In essence, the discriminator implements the function of a binary model, which can be trained using a cross-entropy loss function. Meanwhile, the generative network, on the other hand, expects the generated data to gradually conform to the real data distribution and be able to be discriminated as 1 by the discriminator network, based on which the loss function of the generative adversarial network is defined as:

$$\min_G \max_D V(G,D) = E_{x\sim P_r}[\log D(x)] + E_{z\sim P_z}[\log(1-D(G(z)))] \quad (2)$$

Where $V(G,D)$ is the generative adversarial loss function, the generative network is trained by minimizing $V(G,D)$ and the discriminative network is trained by maximizing $V(G,D)$.

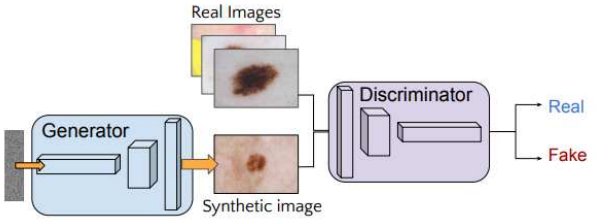

Fig.2. Generative adversarial network model structure

C. Feasibility analysis of heterogenous infrared image generation

This paper focuses on the transformation relationship between visible images corresponding to the same wavelength band and infrared images of the same time period. The proposed infrared images with a wavelength range of 7.5~13 μm and a period of 14:00 are all acquired at the same time period using uniform shooting conditions. From infrared imaging theory, it is clear that different target objects have different emissivity and different corresponding infrared radiation amounts. In the traditional infrared simulation field, the visible light images need to be material segmented and the infrared radiation of each material in the corresponding waveband and atmospheric conditions are calculated separately, and it is seen that different materials have a physical correlation with the corresponding infrared image brightness distribution. Visible images generally have good texture information and contrast. Based on the scene information obtained from visible images, if the mapping relationship from visible images to IR images between different materials can be obtained, then the mapping relationship can be used to realize the generation of IR images from visible images to corresponding scenes.

Generative adversarial networks have a wide range of applications in image generation tasks to learn the mapping relationship between two image domains. By building a generator, we can fit the complex change function from visible to infrared images of different materials, and can fully take into account the change of infrared imaging effect due to the interaction between infrared radiation of different materials. Therefore, a generative adversarial network can be built to realize heterogenous IR image generation, and the mapping function from visible to IR images can be obtained by the game learning between the generator and the discriminator. In this paper, a generator network is used to learn the mapping relationship between the visible image domain and the mid-wave infrared image domain, and experiments show that the generative adversarial network is suitable for the heterogenous infrared image generation task.

III. Introduction of Pix2pix Network Based on Multi-Receptive Field Feature Fusion

A. Overall network construction

The overall network model of the Pix2pix network based on multi-receptive field feature fusion designed in this paper is shown in Fig. 3. The generator network first extracts the multi-receptive field features by using Unet++, and then uses the multi-receptive field feature fusion module to automatically learn the pixel-level weights of different receptive field features and obtain the fused features to generate the corresponding IR images. The discriminator network has the same discriminator structure as the original Pix2pix network, and the PatchGAN discriminator structure is used.

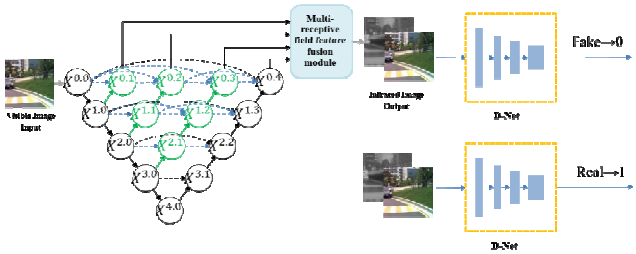

Fig.3. Structure of Pix2pix model for multi-receptive field feature fusion

The Pix2pix network consists of a generator and a discriminator, with the generator denoted as a G-network and the discriminator denoted as a D-network.

B. Analysis of Pix2pix network based on multi-receptive field feature fusion model generator construction

In this paper, we propose a Pix2pix network based on multi-receptive field feature fusion for the task of generating infrared images from visible images, firstly, we extract multi-receptive field features using Unet++ network, then we use multi-receptive field feature fusion module to obtain the fused features and generate the corresponding infrared images, and finally, we obtain the trained transform function through the game learning of generator and discriminator.

In this paper, the Unet++ network is used as a multi-receptive field feature extractor, because the original structure of Unet++ was originally proposed for image semantic segmentation, which is equivalent to pixel-level

semantic classification, but in this paper, the heterogenous infrared image generation accomplishes the task of infrared image grayscale prediction at the pixel level. The Unet++ network is applied to the image transformation task in the same structure as image segmentation, the main differences are :

- Downsampling method: In this paper, the original Unet++ downsampling layer is replaced by a convolutional layer with a step size of 2.
- Multi-receptive field feature processing: this paper adopts the Unet++ network as the multi-receptive field feature extractor, and adopts the multi- receptive field feature fusion module to fuse the extracted features, instead of the original Unet++ network in which the output of multiple up-sampled paths is averaged.
- Loss function: In this paper, we use generative adversarial loss and L1 loss function to train the network, instead of the cross-entropy loss function of the original Unet++ network.

The generator structure is mainly divided into 3 parts: the feature extractor based on the Unet++ network, the multi-receptive field feature fusion module, and the output convolution layer. The base network built based on the Unet++ network is the experimentally validated U-net network, as marked by the shaded box in Figure 4, and the unshaded part in Figure 4 is the densely connected convolutional layer on the filled cross-layer connection path.

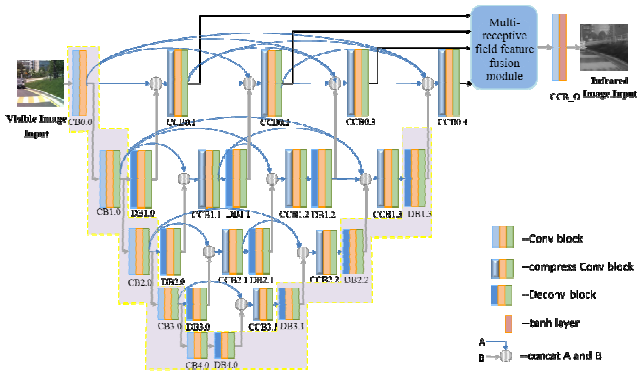

Fig.4. Pix2pix based on multi-receptive field feature fusion generator network structure

Feature extractor based on the Unet++ network. Unet ++ is a deeply supervised encoder/decoder network in which the encoder and decoder sub-networks are integrated through a series of nested dense cross-layer connectivity paths. The redesigned cross-layer connection paths aim to reduce the semantic differences between the feature maps of the encoder and decoder sub-networks. The underlying assumption behind the Unet++ architecture is to progressively enrich the high-resolution feature maps from the encoder network before fusing them with the corresponding semantic feature maps from the decoder network, thus capturing the fine-grained details of the foreground objects more efficiently. The network will handle the learning task more easily when the feature maps from the decoder and encoder networks are semantically similar. This is in contrast to the common skip connection in U-Net, which provides the corresponding resolution feature maps directly from the encoder to the decoder network, leading to feature fusion with large semantic differences.

U-net ++ consists of encoders and decoders connected by a series of nested dense convolutional blocks. The network structure is shown in Figure 5(a), where the black labeled paths represent the original U-Net structure, the green-labeled paths and blue labeled paths represent the dense connection on the cross-layer connection path, and the red labeled paths represent the added supervised learning method. As can be seen in Fig. 5, the Unet++ network is a feature extractor that "fills" the "hollow" U-net by restoring four different levels of features to the original high resolution through four different decoding paths. There are two ways to connect the features between different receptive fields: short connection at neighboring layers and long connection across layers, Figure 5(b) shows the first cross-layer connection path of Unet++. The short neighboring layer connection ensures the continuity of the solid region of the Unet++ network (the green triangularly labeled region in Fig. 5(a)) during backpropagation, and the long cross-layer connection improves the utilization of features at different levels. For the fusion of features with different receptive fields, the U-net network connects convolutional layers with a convolutional kernel size of 1 × 1 after the output paths of different decoders, after which the average value of the output of all paths is calculated.The main improvement of the Unet++ network is to make full use of the features with different receptive fields, where the features with larger receptive fields can fully recognize larger size objects in the image, while the features with smaller receptive fields can retain the features with larger receptive fields can fully recognize larger objects in the image, while the features with smaller receptive fields can preserve the edge information in the image so that the features of smaller-scale objects are not lost.

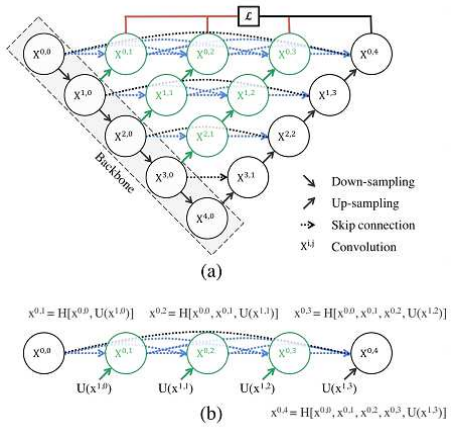

Fig.5. Unet++ model structure

There is also a contradiction between the stronger feature extraction ability of the deep network and the loss of image edge information and local information in the image translation task, so for the Pix2pix network, this paper adopts the Unet++ structure as the feature extractor of the generative network, and adds the solid part of the Unet++ network on the basis of the U-net structure. For the extracted features of different receptive fields, this paper designs a multi-receptive field feature fusion mechanism, which enables the network to automatically learn the importance of different receptive field features.

Figure 4 depicts the detailed network structure of the Unet++ network-based feature extractor. In addition to the shaded box labeled U-net network, the solid part of the

Unet++ network consists of paired upsampling modules and feature compression modules. The upsampling module ensures the feasibility of long connections in the network, while the feature compression module is added to reduce the number of feature channels in order to reduce the network size and speed up the inference process.

Multi-receptive field feature fusion module. The spatial attention mechanism and the channel attention mechanism are for feature aggregation in spatial and channel dimensions, respectively, and are not applicable to feature fusion of multi-receptive field features[11]. With the idea of attention mechanism, this paper proposes a multi-receptive field feature fusion module, whose structure is shown in Figure 6. In this paper, we adopt a nested multi-receptive field feature fusion approach, and design one attention fusion structure at a time to fuse two features with small differences in receptive fields, and the new features obtained are then fused with the larger features[12]. For the attention fusion layer, as shown in the lower right of Fig. 6, the feature maps of two features and f_i are f_{i+1} specified, and we introduce a pixel-level attention mechanism module to learn the pixel-level weights of these two features. For the learning of weights, the two features of the input fusion block are first merged on the channel $[f_i, f_{i+1}]$, after which the pixel-level attention graph is learned using a convolutional layer, while the mapping function restricts the weights in the attention graph to [0,1], and the weights are learned as shown in the following equation:

$$\alpha_i = g(H([f_i, f_{i+1}])) = \frac{1}{1+e^{-H([f_i, f_{i+1}])}} \tag{3}$$

where g denotes the mapping function,and H denotes the convolution function.

According to the obtained attention diagram α_i , the specific fusion of features f_i and f_{i+1} is shown in Eq:

$$FeatureFuse(f_i, f_{i+1}) = f_i \odot \alpha_i + f_{i+1} \odot (1-\alpha_i) \tag{4}$$

where the attention weight of the two features f_i and f_{i+1} are negatively correlated.

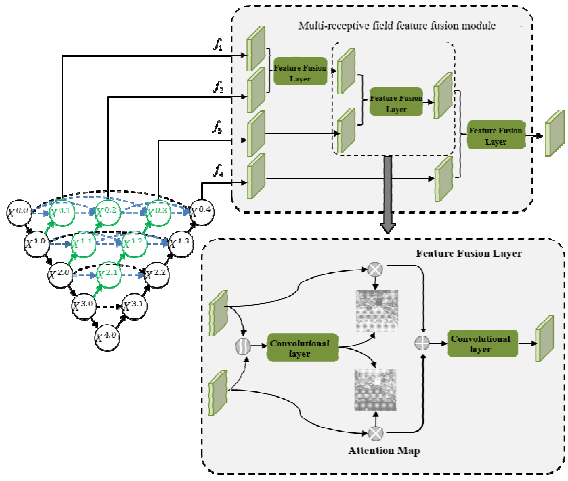

Fig.6. Multi-receptive field attention fusion module.

The multi-receptive field feature fusion module fuses the features of different receptive fields by using a pixel-level attention mechanism, in which one fusion block is used for every two features, nested layer by layer, and its structure is shown in Figure 6, and the parameters of each fusion block are not shared. Given four features of the same size f_1、f_2、f_3 and f_4, we first sample four convolutional layers to downscale the features separately and then use two-by-two nesting to achieve the fusion of features with different perceptual fields.

Output convolution layer. The output convolutional layer is a convolutional layer with a convolutional kernel size of 3 and outputs a 3-channel image.

Generator network parameters. The parameters of the G-network are shown in Table 1. The convolutional layer parameters (f,k,s,p) in the table indicate the number of convolutional kernels, size, step size, and fill size, respectively, and the input and output parameters (b,c,h,w) indicate the training batch size, number of channels, feature map height, and feature map width, respectively, and the convolutional layer parameters of the multi-receptive field feature fusion module have been labeled in the figure, so they are not described in detail in the table.

TABLE I. TABLE OF PARAMETERS OF THE PIX2PIX BASED ON MULTI-RECEPTIVE FIELD FEATURE FUSION GENERATOR NETWORK

Modules	Convoluti-on block	Convolutional layer parameters (f,k,s,p)	Input Features (b,c,h,w)	Output (b,c,h,w)
Input	-	-	(1, 3,256,256)	-
Unet++ Feature Extractor	CB0.0	(64,4,1,2)	(1, 3,256,256)	(1,64,128,128)
	CCB0.1	(64,4,1,1)	(1,128,128,128)	(1,64,128,128)
	CCB0.2	(64,4,1,1)	(1,192,128,12,8)	(1,64,128,128)
	CCB0.3	(64,4,1,1)	(1,256,128,128)	(1,64,128,128)
	CCB0.4	(64,4,1,1)	(1,320,128,128)	(1,64,128,128)
	CB1.0	(128,4,1,2)	(1,64,128,128)	(1,128,64,64)
	CCB1.1	(128,4,1,1)	(1,256,64,64)	(1,128,64,64)
	CCB1.2	(128,4,1,1)	(1,384,64,64)	(1,128,64,64)
	CCB1.3	(128,4,1,1)	(1,512,64,64)	(1,128,64,64)
	DB1.0-1.3	(64,4,1,2)	(1,128,64,64)	(1,64,128,128)
	CB2.0	(256,4,1,2)	(1,128,64,64)	(1,256,32,32)
	CCB2.1	(256,4,1,1)	(1,512,32,32)	(1,256,32,32)
	CCB2.2	(256,4,1,1)	(1,768,32,32)	(1,256,32,32)
	DB2.0-2.2	(128,4,1,1)	(1,256,32,32)	(1,128,64,64)
	CB3.0	(512,4,1,2)	(1,256,32,32)	(1,512,16,16)
	CCB3.1	(512,4,1,1)	(1,1024,16,16)	(1,512,16,16)
	DB3.0,3.1	(512,4,1,2)	(1,512,16,16)	(1,256,32,32)
	CB4.0	(1024,4,1,2)	(1,512,16,16)	(1,1024,8,8)
	DB4.0	(512,4,1,2)	(1,1024,8,8)	(1,512,16,16)
Output	CCB_O	(3,4,1,1)	(1,32,32,32)	(1,3,256,256)

C. Pix2pix network based on multi-receptive field feature fusion model discriminator construction analysis

The discriminator network is used to discriminate the "true or false" input, which is essentially a trainable loss function. Unlike the original GAN network, the D-network built in this paper adopts the idea of PatchGAN, and the output of the network is a 30×30 matrix so that each pixel value of the output corresponds to a 70×70 size image block

of the input, which is able to reproduce more detailed information for image translation tasks. The structure of the D-network built in this paper is shown in Fig. 7, where the convolution module is the same as the G-network.

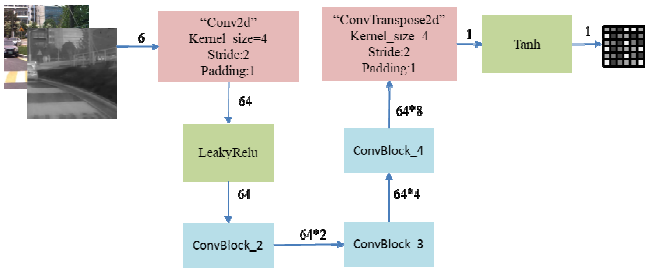

Fig.7. Pix2pix discriminator network structure

The input of the discriminator D network of Pix2pix is an image pair, and the visible image is stitched with the real IR image and the generated IR image in the third channel dimension, respectively, as the true sample pair and the false sample pair of the discriminator. the parameters of the D network are shown in Table 2, and the convolution layer parameters (f,k,s,p) in the table denote the number of convolution kernels, size, step size, and fill size, respectively, and the input and output parameters (b,c,h,w) denote the training batch size, number of channels, feature map height, and width, respectively.

TABLE 2 NETWORK DISCRIMINATOR NETWORK PARAMETERS

Modules	Convolution block	Convolutional layer parameters (f,k,s,p)	Input Features (b,c,h,w)
Input	-	(1, 6,256,256)	-
Down-sampling	(64,4,1,2)	(1, 6,256,256)	(1,64,128,128)
	(128,4,1,2)	(1,64,128,128)	(1,128,64,64)
	(256,4,1,2)	(1,128,64,64)	(1,256,32,32)
	(512,4,1,1)	(1,256,32,32)	(1,512,31,31)
	(3,4,1,1)	(1,512,31,31)	(1,3,30,30)

As can be seen from the table, the D network input is (batchsize, 6 , 256, 256) sample pairs, after three downsampling, according to the data in the table we can calculate the perceptual field of each value of the D network output matrix more than corresponding to the original image. In the convolutional neural network, the perceptual field is used to represent the size of the mapping range of the pixel points in the feature layer of each convolutional layer output relative to the input image. The perceptual field is calculated by the formula:

$$RF^N = \begin{cases} 1, N = 0 \\ k_{1,} N = 1 \\ RF^{N-1} \times k_n - (k_n - 1) \times (RF^{N-1} - \prod_{i=1}^{n-1} s_i), N \geq 2 \end{cases} \quad (5)$$

where RF^N denotes the perceptual field of the Nth convolutional layer, and similarly RF^{N-1} denotes the perceptual field of the N-1th feature layer, k_n denotes the size of the nth convolutional kernel, and s_i denotes the step size of the i-th convolutional layer. According to the formula, the perceptual field of each convolutional layer of the D-network is:

Conv1： k_1 =4， s_1 =2， RF^1 =4；

Conv2： k_2 =4， s_2 =2， RF^2 =10

Conv3： k_3 =4， s_3 =2， RF^3 =22

Conv4： k_4 =4， s_4 =2， RF^4 =46

Conv5： k_5 =4， s_5 =2， RF^5 =70

From the above analysis, it is clear that each pixel in the discriminator output 30×30 matrix corresponds to an image block of 70×70 pixels of the original image, reflecting the judgment of the local true and false information of the image.

D. Network loss function

The loss function of the Pix2pix network based on multi-receptive field feature fusion consists of the G network loss function and the D network loss function, and the training process of the network G network and D network game process, the goal of the G network is to minimize L_{pixGAN} and the goal of the D network is to maximize $L_{pixGAN}(G,D)$ as shown in the following equation:

$$L_{pixGAN}(G,D) = E_{V,I}[\log D(V,I)] + E_V[\log(1 - D(V,G(V)))] \quad (6)$$

At the same time, considering that the task of G-network is not only to "cheat" the discriminator but also to implement the image translation task, so that the generated image is infinitely close to the real IR image in the sample pair. Therefore, the loss function of the G network is added to the L1 loss function, and the final loss functions of the G and D networks are shown in Eqs. 7 and 8.

$$L_{G\_net}(G) = L_{pixGAN}(G) + \lambda L_{l1}(G) = E_V[\log(1 - D(V,G(V)))] + \lambda L_{l1}(I,G(V)) \quad (7)$$

$$L_{D\_net}(D) = -L_{pixGAN}(D) = -(E_{V,I}[\log D(V,I)] + E_V[\log(1 - D(V,G(V)))]) \quad (8)$$

where λ is the weight of the loss term $L_{l1}(I,G(V))$.

The loss function of the G-network is constrained by a loss so that the generated IR image distribution matches the target image distribution, and by b loss so that the generated IR image is consistent with the target image at the pixel level in terms of grayscale.

IV. MODEL TRAINING AND EXPERIMENTAL RESULTS ANALYSIS

The dataset used in the experiments is the Kaist lab release in-vehicle road scene dataset. The infrared images in the dataset are taken by FLIR infrared camera in the range of 7.5~13um, and the selected time period is 14:00, while the visible images have the same structure and one-to-one matching with the corresponding infrared images. The scenes of the whole dataset have some similarities, including scenery such as roads, trees, grass, vehicles, and a small amount of pedestrian interference. The resolution of the images is 256×256. The training dataset contains 2500 pairs of visible-IR image samples and the test dataset has 500 pairs.

For the training parameters, the batchsize is set to 1 in this paper, and the Adam optimizer is used in this paper, with β_1 、 β_2 set to 0.5 and 0.999, respectively.

According to our built Pix2pix network based on multi-receptive field feature fusion, the loss functions of generator and discriminator are shown in Fig. 8 and Fig. 9, respectively, where the generator loss function and L1 loss function in been kept decreasing and the training process is relatively stable. The discriminator loss starts as the generator structure of the Pix2pix network based on multi-receptive field feature fusion is more complex compared to the discriminator network, and converges more slowly compared to the discriminator network, and the discriminator network loss is very small at the beginning, and with the gradual enhancement of the fitting ability of the generator network, the discriminator network loss slowly becomes larger and finally converges to the theoretical equilibrium point log(1-0.5) $\approx$ 0.7 or so.

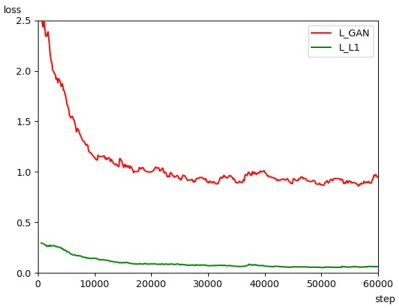

Fig.8. Pix2pix network based on multi-receptive field feature fusion generator loss curve.

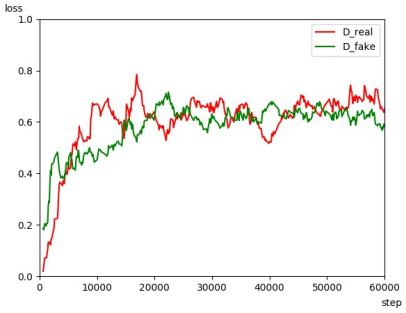

Fig.9. Pix2pix network based on multi-receptive field feature fusion discriminator loss curve

In order to compare the effect of Unet++ as a generator, we use the pix2pix model with Unet as a network generator to experimentally compare with the Unet++ model under the same data set. the trend of the generators and discriminators of the pix2pix model with Unet as a generator during the training process is shown in Figure 10 and Figure 11.

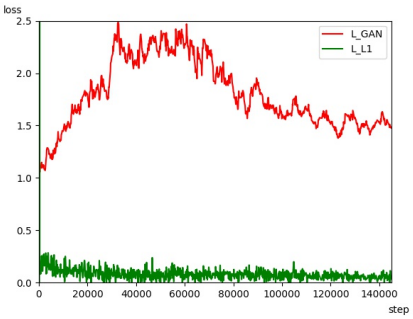

Fig.10. Pix2pix generator loss curve

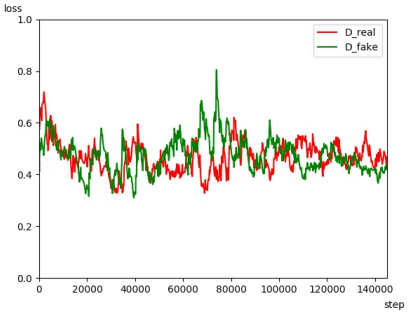

Fig.11. Pix2pix discriminator loss curve.

By comparison, it was found that the Pix2pix network based on multi-receptive field feature fusion converged after 60,000 iterations, and the training speed of the Pix2pix network based on multi-receptive field feature fusion was significantly faster compared to the convergence of the Pix2pix network in 140,000 steps.

The trained network was used to test the data in the test set, and the results obtained are shown in Figure 12, where the first column is the visible image, the second column is the real infrared image corresponding to the visible image, the third column is the infrared image generated by the Pix2pix network, and the fourth column is the infrared image generated by the Pix2pix network based on multi-receptive field feature fusion. In the first row, the Pix2pix model does not generate the contour information of the car owner, and the edge information of the rear car is blurred; in the second row, the roof edge of the car is not generated smoothly; in the third row, there are two pedestrians in the scene but we can only observe one pedestrian information after generation. In the third row, there are two pedestrians in the scene but we can only observe one pedestrian after generation. The comparison shows that the multi-receptive feature fusion network achieves more accurate detail recovery.

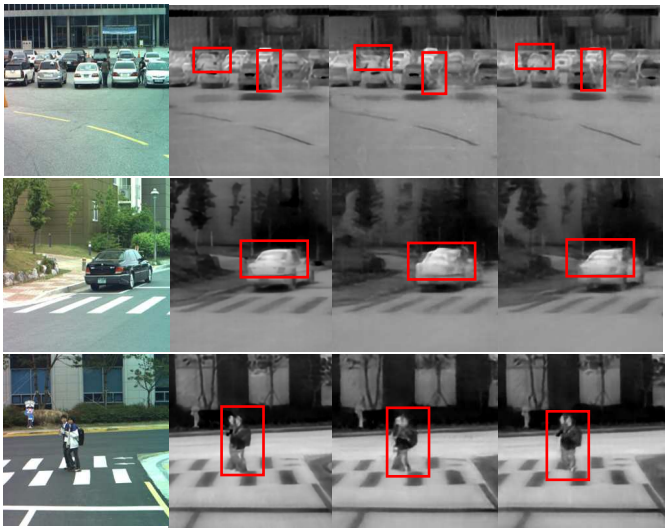

Fig.12. Experimental results of Pix2pix and Pix2pix based on multi-receptive field feature fusion

TABLE 3 QUANTITATIVE ANALYSIS OF THE PERFORMANCE OF PIX2PIX AND PIX2PIX BASED ON MULTI-RECEPTIVE FIELD FEATURE FUSION.

Models	SSIM	PSNR
Pix2pix	0.793	26.646
Pix2pix based on multi-receptive field feature fusion	**0.861**	**28.709**

In this paper, the attention maps of different receptive field features learned by the multi-receptive field feature fusion module are visualized as shown in Fig. 13 and Fig. 14. (a) and (b) are the visible image and the corresponding infrared image, respectively, and (c), (d), (e) and (f) are the attention maps of features with small to large receptive fields, respectively.

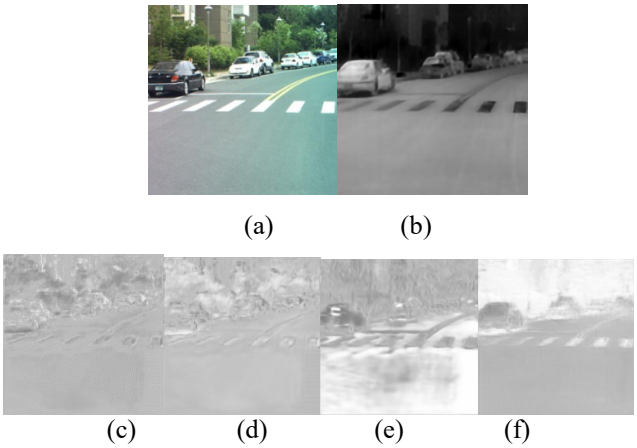

Fig.13. The fused attention of different receptive field features1

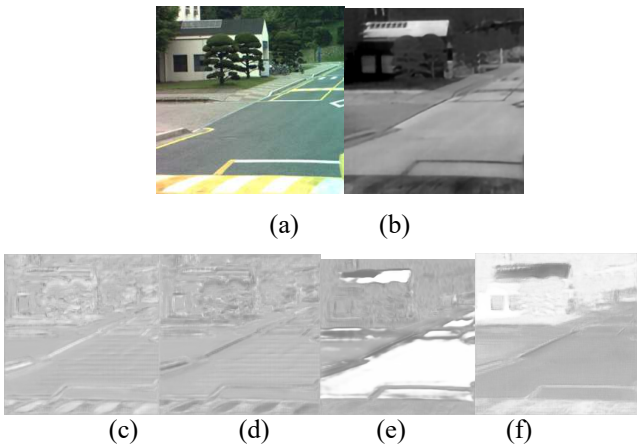

Fig.14. The fused attention of different receptive field features2

From the figures, we can see that the attention maps corresponding to features with smaller receptive fields focus more on the texture, local edges of the images, while the attention maps corresponding to features with larger receptive fields focus on the larger contour information of the images, which have greater structural similarity with the infrared images.

V. Conclusion

In this paper, we propose a Pix2pix network based on multi-receptive field feature fusion for heterogenous infrared image generation task, build a multi-receptive field feature extraction based on Unet++ network, design a multi-receptive field feature fusion mechanism, and experimentally show that compared to existing pix2pix networks, the pix2pix based on multi-receptive field feature fusion has higher performance and has greater structural similarity with the infrared images.

Acknowledgment

We thank the National Natural Science Foundation of China for supporting this study and all the authors for their hard work on this paper.

References

[1] Xiaochun Luo, Jiyin Sun, and Jing Liu, “Infrared image acquisition using inversion of visible images,” Infrared and Laser Engineering, 2008(05): 773-776.(in Chinese)

[2] Wang, R. F. , W. D. Wu , and J. X. Huo . "The design of real time infrared image generation software based on Creator and Vega." ISPDI 2013 - Fifth International Symposium on Photoelectronic Detection and Imaging International Society for Optics and Photonics, 2008.

[3] Xie J R, Li F M, and Wei H, “Infrared target simulation method based on generative adversarial neural network,” Journal of Optics, 2019, 39(03): 150-156.

[4] Yunfei. Feng, “Research on infrared image band expansion method based on real images,” Xi'an University of Electronic Science and Technology, 2019. (in Chinese)

[5] Isola P, Zhu J Y, Zhou T, et al. “Image-to-image translation with conditional adversarial networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1125-1134.

[6] Liu Y , Qin Z , Wan T , et al. “Auto-painter: Cartoon Image Generation from Sketch by Using Conditional Generative Adversarial Networks,” Neurocomputing, 2017:S0925231218306209.

[7] Zhou, Z. , et al. "UNet++: A Nested U-Net Architecture for Medical Image Segmentation." 4th Deep Learning in Medical Image Analysis (DLMIA) Workshop 2018.

[8] Ronneberger O , Fischer P , Brox T . “U-Net: Convolutional Networks for Biomedical Image Segmentation,” Springer, Cham, 2015.

[9] Baird, James K. , and T. R. King . "A wien displacement law for impact radiation." International Journal of Impact Engineering 23.1(1999):39-49.

[10] Goodfellow, I. J. , et al. "Generative Adversarial Networks." Advances in Neural Information Processing Systems 3(2014):2672-2680.

[11] [1] Zhao, W. , J. J. Huang , and B. Tian . "An Image Fusion Algorithm Based on Receptive Field Model." Acta Electronica Sinica 36.9(2008):1665-1669.

[12] Zhang H, Goodfellow I, Metaxas D, et al. “Self-Attention Generative Adversarial Networks,” in International Conference on Machine Learning. 2019: 7354-7363.