Image Fusion based on Generative Adversarial Network Consistent with Perception

Yu Fu[a], Xiao-Jun Wu[a,*], Tariq Durrani[b]

[a]*Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, School of Artificial Intelligence and Computer Science, Jiangnan University, 214122, Wuxi, China*

[b]*Department of Electronic and Electrical Engineering, University of Strathclyde, G1 1XW, Glasgow, UK*

Abstract

Deep learning is a rapidly developing approach in the field of infrared and visible image fusion. In this context, the use of dense blocks in deep networks significantly improves the utilization of shallow information, and the combination of the Generative Adversarial Network (GAN) also improves the fusion performance of two source images. We propose a new method based on dense blocks and GANs , and we directly insert the input image-visible light image in each layer of the entire network. We use structural similarity and gradient loss functions that are more consistent with perception instead of mean square error loss. After the adversarial training between the generator and the discriminator, we show that a trained end-to-end fusion network- the generator network- is finally obtained. Our experiments show that the fused images obtained by our approach achieve good score based on multiple evaluation indicators. Further, our fused images have better visual effects in multiple sets of contrasts, which are more satisfying to human visual perception.

Keywords:
image fusion, generative adversarial networks, dense block, infrared image, visible image

1. Introduction

The infrared and visible image fusion task is a significant theme in the imaging field[1]. Generally, visible images contain rich reflected light information, which has a high spatial resolution, sufficient detail texture information, and great contrast. However visible images are easily affected by some environment and climatic problems such as insufficient light, fog, and vegetation occlusion resulting in a large loss of key information. Meanwhile, infrared images contain rich thermal radiation information, which can resist the negative environmental interference, however, infrared images have low spatial resolution and less texture information. Therefore, the infrared and visible image fusion offers sufficient features from the source images, and fusing them with appropriate fusion strategies delivers complementary features. The proposed algorithm can effectively extract valuable information of each source images and fuse them into high-quality stable and informative images. The fusion of infrared and visible images has many applications in practice, such as video surveillance, object recognition, tracking, remote sensing , military applications, etc.

There are currently many mature fusion algorithms[2]. According to the different approaches adopted, fusion algorithms can generally be divided into seven categories[1]:

1) Multi-scale transform fusion. Here, firstly each source image is decomposed into a set of multi-scale representation features using pyramid[3], curvelet[4], contourlet[5] , etc. Secondly, the multi-scale feature representation is fused according to the specific fusion rule, and finally the corresponding inverse multi-scale transform is used to obtain the fused image.

2) Sparse representation fusion. An over-complete dictionary is learned from high quality images, and sparse coding is used on each sliding-window patch to obtain the sparse representation coefficient. Finally, the fused images are reconstructed by the over-complete dictionary. There are representative fusion algorithms such as sparse representation (SR) and gradient histogram (HOG)[6], joint sparse representation (JSR)[7], approximate sparse representation with multi-selection strategy[8], etc.

3) LRR fusion. The low-rank representation (LRR) can be used to extract the source image features in the low-rank domain[9].Then the fusion features map is fused according to the specific fusion rule, and the fused image is reconstructed.

4) Neural network-based methods, the neural network has strong adaptability, fault tolerance and anti-noise ability. It can extract features and fuse well. The details will be explained in the next section.

5) Subspace-based methods. These methods aim to project high-dimensional images into low-dimensional space. This can reduce the interference of redundant information, and obtain the internal structure of the source image. These methods have led to successful algorithms such as PCA[10], ICA[11], NMF[12].

6) Saliency-based methods. A salient model can be used to extract the salient regions of the source image, which can obtain a weight map[13] or extract the salient object[14], and then reconstruct the fused image based on the saliency features.

7) Other methods. For example, each of the above meth-

*Corresponding author email: wu_xiaojun@jiangnan.edu.cn

ods has its advantages and short-comings, and we can combine their advantages to further improve the fusion quality, such as hybrid multi-scale transformation and saliency[14], hybrid multi-scale transformation and neural network[15], hybrid multi-scale transformation and sparse representation[16] and other methods. On the other hand, fuzzy logic theory is also a very useful tool to obtain weighted maps for infrared and visible image fusion[17].

Although traditional methods can accomplish the image fusion task well, there are still problems such as noise and artifacts. However, the emergence of deep learning has opened a new avenue for image fusion tasks.

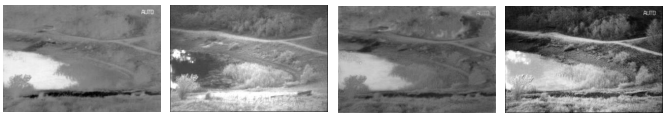
With the development of deep learning, the powerful ability of feature extraction and data representation capabilities of deep networks have become increasingly attractive. Deep learning is an excellent application in the field of computer vision, and many excellent methods in image fusion[18]. There are many effective methods in shallow neural networks such as networks less than ten layers. For example, Convolutional Neural Network (CNN)[19] or Sparse Autoencoder (SAE)[20] are used in the middle layer to determine the features as a weight calculation tool to obtain the weight map of the two source images. Further, the images can be fused according to the weight map. Besides, a set of features can be extracted by a dense block (Densefuse)[21] or PCA filters of PCANet[song2018multi], and features fused in the middle lay
through a specific decoding operatio

In deep learning networks, the p
VGGNet[22] and ResNet[23] can al
features and fuse them, and the reco
fused images. In addition to the u
as a tool for feature extraction, it c
end-to-end image fusion network.
as FusionGAN[24], here its generat
priate loss function, can directly ge
from source images. The GAN netv
age fusion tasks, including multi-f
GAN)[25], multi-exposure image fu
mote sensing image fusion (PAN-G

In recent years, with the rapid de
tive Adversarial Network (GAN)[28
with sufficient information and gooc
proaches are widely used in Super-I
mentation, and Image Enhancement.
image fusion. In FusionGan[24], M
use the GAN to fuse images. By ad
generator of the GAN network, the g
image that contains the source image
signs the appropriate loss function to control the structure of the fused image in the generator, and the discriminator improves the fusion quality. However, the fused image is unstable, and the fusion effect is not natural, as shown in Fig. 1(c). Although the features information in infrared images can be extracted, the fused image loses the edge and detail texture information of the source images. In the following year, Ma et al. proposed two improved networks based on GANs for image fusion such as DDcGAN[29] and ResNetFusion[30] which exhibited outstanding performance. However, they have shortcomings such as image blur, loss of details, and poor perception of fused images. In this case, we believe that the network loses part of the source image features during the fusion process. The loss of detail texture information needs to be supplemented.

In order to improve the end-to-end fusion quality of GAN, this paper proposes a new GAN network framework. To add detailed information, the dense block is used in the generator, and the shallow layers with richer detail features and the source image are concatenated with the deeper layers. In addition, we concatenate visible images at each layer, so that the fused image retains more visible information.

Simultaneously, to make the fused images have a similar structure to both of the source images, not just visible images, a structural similarity loss function and gradient loss are added to the generator to control the structural similarity between the generated images and the source images. On the other hand, the role of the discriminator is to compare the fused images with the visible images and obtain a loss value. Because the visible image usually has a better visual effect and is more in line with human aesthetic perception. The discriminator is intended to force the fused image towards a visible image to enhance the visual effect of the fused images. In our proposed network, the generator is an end-to-end fused network. Image fusion does not need to extract features to calculate a weight map or design



(a) Infrared Image (b) Visible Image (c) FusionGan (d) Ours

Figure 1: FusionGan fusion method and Our method proposed in this paper.

Our contributions are summarized as follows:

1) We apply dense connections as the generated backbone network, and we used the skip connection of visible images to fuse the image texture information from visible light, which is simple but very effective in enhancing the texture details of the fused image.

2) We abandoned the common mean square error loss function as the content loss function, and replaced it with structural similarity loss and gradient loss.

3) In addition, we calculate the adversarial loss between the image and only the visible light image to ensure that the generated image is real and natural enough.

Perhaps our method is simple, but it is indeed very effective, as shown in the Fig.1. The quality of the generated fusion image

is obviously improved significantly.

In Section 2, we review the Generative Adversarial Networks and DenseNet. Section 3 gives the network structure of the GAN and the design of the loss function. Section 4 describes the GAN parameters and specific training details. Section 5 compares our network with the state of the art methods. Section 6 draws the conclusions for the work.

2. Related Work

2.1. Generative Adversarial Networks

Goodfellow et al. first proposed the concept of Generative Adversarial Networks (GAN) in a paper published in NIPS 2014[28]. The algorithm for the GAN has two network models. One of them is the generator, the noise z is the input, and its task is to generate a fake image that looks real. The other one is the discriminator which has as its input a real image x or a generated fake image $G(z)$, and its task is to determine whether a given image is a real image or a fake image. Analytically, the generator aims to minimize the data distribution gap between the generated images and the real images. It makes the distribution of the fake images (P_z) as close as possible to the distribution of the real images (P_{data}) , so that the results of discriminator between the two data is similar. The purpose of the discriminator is to maximize the discrimination between real and fake images as much as possible. The two networks perform an adversarial training as follows:

$$\begin{aligned}&\min_G \max_D V(D,G) = \\ &\mathbb{E}_{x\sim P_{data}(x)}\left[log(D(x))\right]+\mathbb{E}_{z\sim P_z(z)}\left[log(1-D(G(z)))\right]\end{aligned} \tag{1}$$

The discriminator D is optimized firstly to distinguish between the real and generated pictures, and it can provide a loss value to the generator G. Then the generator is optimized. Finally $D(G(z))$ tends to 0.5, and this means that the discriminator cannot distinguish whether the fake image generated by the generator is a real image or not. It also achieves the goal of GAN that the generator G is able to produce images that look real.

The GAN proposed for the first time still has many drawbacks. The sigmoid activation layer in GAN is easily saturated. After several training sessions, the discriminator can distinguish the authenticity of the data, however the discriminator is unable to provide an effective gradient, and the generator is unable to update effectively. Mao et al. proposed a least squares GAN (LSGAN)[31], and they changed the sigmoid activation layer in the network to the linear activation layer and calculated the loss using the least squares method. The loss function is defined as follows:

$$\begin{aligned}&\min_D V_{LSGAN}(D) = \\ &\frac{1}{2}\mathbb{E}_{x\sim P_{data}(x)}\left[(D(x)-b)^2\right]+\frac{1}{2}\mathbb{E}_{z\sim P_z(z)}\left[(D(G(z))-a)^2\right]\end{aligned} \tag{2}$$

$$\min_G V_{LSGAN}(G) = \frac{1}{2}\mathbb{E}_{z\sim P_z(z)}\left[(D(G(z))-c)^2\right] \tag{3}$$

Where a and b respectively represent the labels for fake data and real data, and c represents the labels that the generator uses to deceive the discriminator successfully. The authors propose two ways to set up the values. One way is to use $b-c=1$ and $b-a=2$, to minimize the Pearson Chi x^2 divergence between $P_{data}+P_z$ and P_z. Another way is to set $b=c$, for example, take $b=c=1$, $a=0$. In this way, the LSGAN can effectively incur a large loss for the data that is discriminated correctly by the discriminator but is quite different from the correct one, so that the network can be more effectively trained to generate higher quality images. In the method proposed in this paper, we choose the strategy training network with $b=c=1$, and $a=0$.

2.2. FusionGAN

In the FusionGAN[24] proposed by Ma et al., the GAN network generator is essentially an end-to-end image generation network. The input of the network is infrared image and visible light image, and a fusion image is output through five simple convolutional layers without designing a complicated fusion strategy. In order to control the generated results, the author uses the mean square error loss function. The discriminator is to distinguish images with clear details in order to obtain more detailed information.

FusionGAN gave an example of using GAN for image fusion, but there are still some disadvantages.

1. Although the mean square error loss function can effectively constrain the generated image, the constraint of the L2 norm will force the image to learn information from two different source images at the same time. We believe that the generated image appear to be a gray image subjectively.

2. The network structure of the generator is very simple. Although it can generate images, it is difficult to capture the detailed texture information of the source image. In the last layer of FusionGAN, only the deep semantics features of the previous layer are fused. The useful information obtained by the middle layer and the shallow layer is lost. When the network layer is deeper, the situation becomes worse, because the deep feature map is more abstract, and the shallow layer shows more edge detail information.

3. It is indeed feasible to use the discriminator to supplement the detailed texture. For example, replacing the discriminator with PatchGAN[32] may be a very good way to improve the image resolution. However, we believe that compared with judging the amount of texture information, the discriminator is easier to make the generated image biased toward a specific data distribution.

To overcome this shortcoming, we design a novel and effective generator network based on convolutional network with denseblock to enrich feature detail information. In each convolutional layer, we also concatenate the original image and the visible light image, to retain its original information. We design a reasonable loss function in which we add gradient loss and structural loss instead of the mean squared error.

2.3. Image Fusion with Denseblock

Huang and Liu et al. proposed the new network framework DenseNet[33] in the best paper of CVPR2017. In this network

framework, each layer is concatenated with all previous layers in the channel dimension as input to the next layer. That is, the input of each layer is $X_l = H_l([X_0, X_1, X_2, \ldots, X_{l-1}])$. Where H_l represents a non-linear operation including normalization, activation layer, convolution layer, etc.$[X_0, X_1, X_2, \ldots, X_{l-1}]$ represents concatenation of all layer outputs prior to layer l. Such an operation can make the reuse of features as much as possible, so only a few feature maps are needed. This not only improves the utilization of the features, but also greatly reduces the network parameters, and also reduces the problem of the disappearance of the deep network gradient.

In the paper[21] published by Li and Wu et al., the denseblock is used to concatenate features in all layers to improve the fusion quality. The autoencoder network proposed by the authors consists of a convolutional neural network and a densely connected network. Each layer of output in the encoder is concatenated to each of the other layers behind it to extract more useful information from the original image and the shallow layers during the encoding process. The pre-trained encoder model can encode the source image and then fuse the features using appropriate fusion strategies. Finally, the fused image is reconstructed by the decoder.

3. Proposed Method

In this section, the GAN network fusion method based on deep learning is introduced in detail. Our network consists of generator and discriminator. This section gives the network framework of the generator and discriminator, as well as the loss function design of the GANs.

3.1. Network framework

Our network consists of generator and discriminator. In the training phase, the network shown in Fig. 2 uses the concatenated infrared image I_{ir} and the visible image I_{vi} as inputs to the generator G. After 5 layers of convolution and dense block concatenation operations, a fused image I_f is obtained. The visible image I_{vi} and the fused image I_f are then input to the discriminator D. Both of them perform adversarial training.

Here only the visible image is input into the discriminator rather than both of source images. Since the discriminator judges whether the image generated by the generator is "real" enough or "natural" enough, the discriminator can force the generator to generate a more natural image, referring to the visible light image. The "natural" results of our method can be found in Section 4.2.1-Subjective Evaluation. In other words, the loss of the discriminator can change the style of the image generated by the generator, but it has little effect on the amount of detailed information the generator obtains from the source images. To get a natural-looking image, we decide to input the visible light image and the fused image into the discriminator rather than both source images.

In the test phase, the discriminator D is not used, and only the generator is retained as shown in the gray box in Fig. 2. The infrared image I_{ir} is concatenated with the visible image I_{vi}, and these are input to the generator to obtain the fused image I_f.

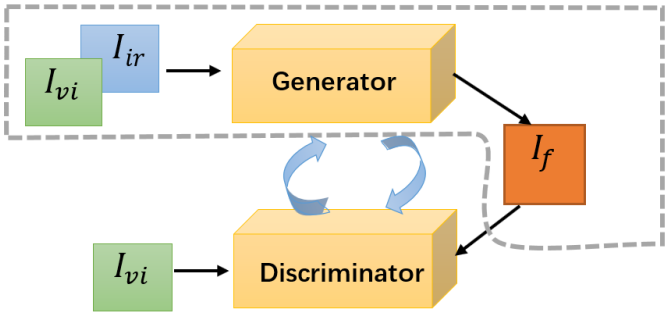


Figure 2: FusionGan fusion method and Our method proposed in this paper.

3.1.1. Generator Network

As shown in Fig. 3, the generator network G is based on a simple convolutional neural network. The visible image I_{vi} and the infrared image I_{ir} are concatenated as input images of the generator network. The first layer contains a 5×5 convolution kernel for extracting shallow features. As shown in the gray box of Fig. 3, the second to fifth layers form a denseblock convolutional layer, and the output of each layer is concatenated to all subsequent layers as inputs to following layers.

The second layer uses a 5×5 convolution kernel to expand the receptive field of the shallow network. The third to fourth layers use a 3×3 convolution kernel to reduce network parameters. The fifth layer uses a 1×1 convolution kernel to reduce the dimension of the concatenated features to a single-channel image to achieve features fusion, so that the fused image I_f is obtained in an end-to-end manner. There is a batch-normalization layer behind each layer, a Leaky ReLU[34] activation function in the first four layers, and a Tan activation function in the fifth layer. We choose Leaky ReLU as our activation function, because the ReLU activation function discards CNN negative threshold neurons. This may make the output sparse, but it is not suitable for image fusion tasks that need to retain much information. The addition, the dense blocks enables the shallow information features of the first few layers of the network to be reused as much as possible. Details such as colors and edge contours are well preserved for the deeper layers.

In addition, the reconstruction network such as DenseFuse, DeepFuse and so on cannot concatenate the source image in the middle layers, which will cause the network to secretly copy the source images to the output layer through skip connections. However, in our approach, we calculate multiple effective losses between the fused image with the two images, it must be able to generate an image that contains information from both images. Based on this, we innovatively concatenate visible images directly in the middle layer. This can retain more visible image information without losing infrared information. We use the visible light image I_{vi} as part of the input image. As shown by the green dotted line connection in Fig. 3, our network concatenates I_{vi} in each layer to supplement the detail texture information of the features.

The operation of directly inserting the input image at each layer of the network can make it easier for the network to learn

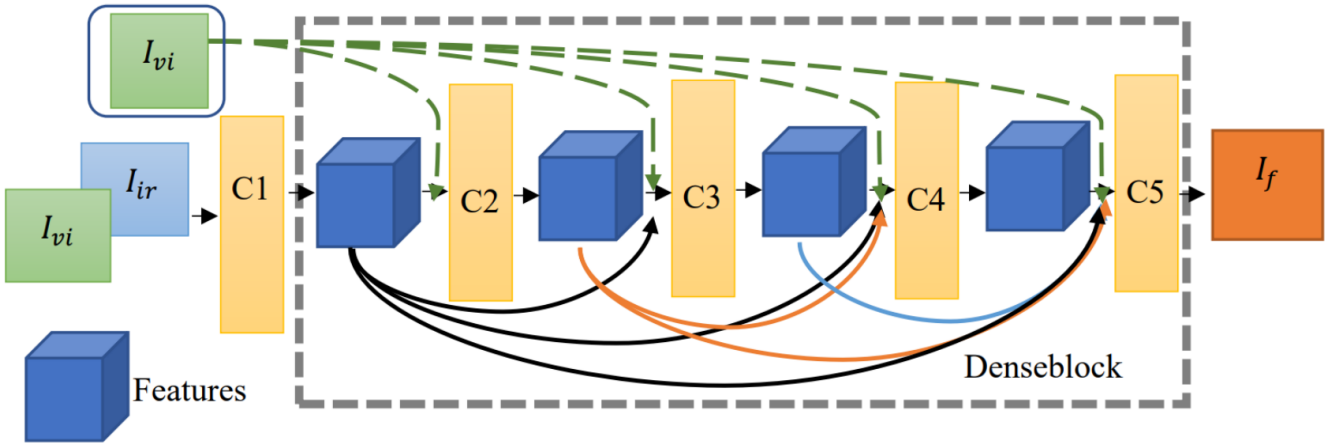


Figure 3: Generator network

visible light image information. For visible light images, we hope to retain its original details. The skip connection of visible image is equivalent to using different depth networks for feature extraction of visible light images. Therefore, the basic information of the visible light image has been extracted from networks with multiple depths, which can obtain its semantic information, and retain its texture information as much as possible. And we believe that the more important thing for infrared images is radiation information, which is a kind of local semantic information, so we did not use skip connections in the middle layers, and used the deepest network to extract the semantic information of infrared images.

3.1.2. Discriminator Network

As shown in Fig. 4, the discriminator network consists of four convolutional layers and one fully connected layer. The input image of the network is the visible image I_{vi} or the fused image I_f of the generator. There is a max-pooling layer behind each convolutional layer. The Leaky ReLU activation function is used in the first four layers, and the Tanh activation function is selected in the last layer of the full connection layer. The output of the discriminator network is a label that represents whether the input image is a real image or a fused image.

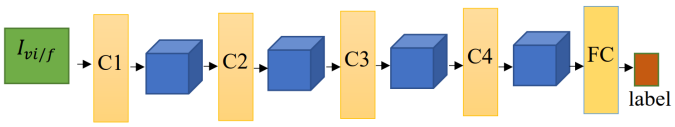


Figure 4: Discriminator network

3.2. Loss Function

The strategy here is to concatenate the infrared image I_{ir} with the visible image I_{vi} and input them into the discriminator G to obtain a fused image I_f. In order to make the fused image I_f tend to visible image I_{vi}, we input them into the discriminator to distinguish them. The discriminator D provides a loss to the generator G, enabling G to generate a fused image with a better visual effect.

3.2.1. Discriminator Loss Function

In order to enable the discriminator to better distinguish the fused image from the visible image, the loss function of the LSGAN is used here. The loss function of the discriminator is as follows:

$$L_D = \frac{1}{N}\sum_{n=1}^{N}\left(D\left(I_{vi}^n\right)-b\right)^2 + \frac{1}{N}\sum_{n=1}^{N}\left(D\left(I_f^n\right)-a\right)^2 \quad (4)$$

Where $n \in \mathbb{N}_N$, N represents the number of fused images, I_{vi}^n represents the n_{th} visible image, I_f^n represents the n_{th} fused image. The purpose is to enable the discriminator to distinguish between the I_{vi}^n and the I_f^n, so that the I_f is more biased toward the visible image, which is more real and natural, and improves the visual effect of the fused image I_f.

Here, we only calculate the loss of visible images and fused images. According to the theory of GAN, the discriminator is only used for image classification. The traditional fused images retain the infrared information, but also retain the characteristics of blurred and dim light in the infrared image. The images tend to be gray. In our network, the loss between visible image and fused image in discriminator can force the generator to generate obvious images rather than gray images. In other words, most of the fusion results of the previous fusion methods are learning an average data distribution between the different distributions of infrared images and visible light images.The fused image they generate is more like an interpolated image of an infrared image and a visible light image. The discriminator moves the data distribution of the fused images to the data distribution of the visible light image, so that the fusion result looks more natural. Our discriminator is used to increase the visual effect

of the fused images, which is in line with human aesthetics. For the fusion of infrared information, it mainly depends on the content loss in generator loss function.

3.2.2. Generator Loss Function

The generator uses LSGAN generator loss function as its loss function as follows:

$$L_{LSGAN}(G) = \frac{1}{N}\sum_{n=1}^{N}\left(D\left(G\left(I_f^n\right)\right) - c\right)^2 \tag{5}$$

where $b = c = 1$,$a = 0$ in our paper. To modify the loss function of the generator, we hope that the generator can ensure that the fused image retains the source image details. Therefore, the content loss function $L_{content}$ (see below) is added to the loss function as follows:

$$L_G = L_{LSGAN} + \gamma L_{content} \tag{6}$$

Where γ is a hyperparameter, used to balance the weight of the two, in our paper the value $\gamma = 100$.

In previous image fusion methods, authors usually use $L2$ loss function or mean square error loss. This loss function has the following shortcomings: 1. Compared with small errors, MSE is more sensitive to large errors, which will make the generated image tend to be smooth. 2. The network directly imposes pixel-level constraints on the generated image and the original image, without considering the image's overall structure, which will make the generated image tend to be an average image of infrared and visible light images—a gray image.

Based on this idea, we discarded the MSE loss function. Considering the structural similarity between the fused image and the original image, the $SSIM_f$ loss is added to the loss function. In order to enable the I_f to extract more detail from I_{ir} and I_{vi}, the $Gradient_f$ loss is added to the loss function.The $L_{content}$ loss function is as follows:

$$L_{content} = \left(1 - SSIM_f\right) + \alpha Gradient_f \tag{7}$$

where α is a hyperparameter, used to balance the weight of the two loss value, in our paper, the value is $\alpha = 5$.

Where $SSIM_f$ is calculated for structural similarity between I_{vi}, I_{ir} and I_f respectively. The two values are multiplied by 0.5. The similarity calculation algorithm $SSIM(x, y)$ is a value for calculating the brightness, contrast and structural difference between two images x and y. The calculation formula is used as follows:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{8}$$

$$SSIM_f = \frac{SSIM\left(I_f, I_{vi}\right) + SSIM\left(I_f, I_{ir}\right)}{2} \tag{9}$$

Where μ represents the mean and σ represents the standard deviation. The larger the $SSIM$, the higher the structural similarity between the two. In the $L_{content}$, we expect the $SSIM$ to be larger, so we take $(1 - SSIM_f)$ as part of the $L_{content}$.

The $Gradient(x, y)$ is calculated as the gradient difference between x and y as follows:

$$Gradient(x, y) = \frac{1}{M}\sum_{n=1}^{M}(\nabla x - \nabla y)^2 \tag{10}$$

$$Gradient_f = Gradient\left(I_f, I_{vi}\right) + \beta Gradient\left(I_f, I_{ir}\right) \tag{11}$$

Where M is the number of pixels of the image, ∇ is the gradient calculation operation, and the $Gradient(x, y)$ is calculated as the mean value with $L2$ norm. β is a hyperparameter and is used to balance the gradient weights of the I_{ir} and the I_{vi}. In our paper, the value $\beta = 2$.

We will explain the setting of hyperparameters in ablation studies.

We do not use the mean square error loss function, because the mean square error loss will make the data distribution generated by the generator to be a mean in the distribution of all samples, as show in Fig.5. We believe that the adversarial loss is a more reasonable loss function, which allows the generator to randomly select one of the "real" data distributions to generate it to deceive the discriminator. The fused image is not easy to become blurred and looks natural enough. In addition, we have to ensure that the generated image maintains sufficient structural similarity and certain edge information with the input image, so we add the designed content loss.

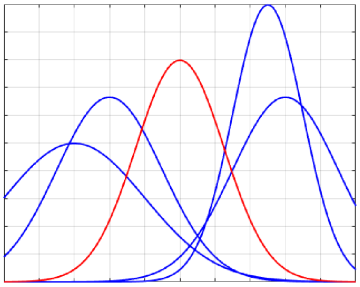


Figure 5: We set the blue distribution to be the distribution of the real sample, then the mean square error loss function learns the red distribution. The adversarial loss learns an approximate distribution from real distributions.

3.3. Network Framework Parameters

3.3.1. Generator Parameters

The generator is a convolution neural network with denseblock. The first layer and the second layer both use a 3×3 filter. The third layer and the fourth layer both use a 3×3 filter, and the fifth layer uses a 1×1 filter for channel dimensionality reduction operation. The stride is set to 1, and the padding is set to SAME. We do not perform a downsampling operation, in order to have the same shape for concatenating features. To avoid the disappearance of the network gradient, adding a batch normalization layer and a nonlinear activation function behind each layer of the convolutional layer can make the network more stable. For the activation function, the Leaky Relu activation function is selected in the first four layers, and

the Tanh activation function is selected in the last layer. The detailed network parameters of the generator network are given in Table 1.

3.3.2. Discriminator Parameters

The first four layers of the discriminator are convolutional neural networks with an active layer with a Leaky ReLU activation function and a max-pooling layer. The features obtained in the fourth layer are reshaped into one channel. The last layer is the fully connected layer with the Tanh activation function to obtain the classification result. The detailed parameters of the discriminator are shown in Table 2.

4. Experiments and Analysis

4.1. Training and Test Details

59 pairs of different infrared images and visible image pairs are selected from the TNO[35] database as the training data set of the network. In order to solve the problem that the size of the source images may be different in reality, the pixel size of each pair of source images is not the same. The largest pixel size of the training image is 678×917, and the smallest pixel size is 256×256. Note, 59 pairs of images to train the network model is not enough. Each image is cropped by setting the stride to 14 to give a size of 120×120 pixels. Then we get 54842 pairs of images. The pixel size of each cropped image is 120×120. The entire data is randomly divided into two parts: 70% of these are training data, and the rest 30% are testing data.

In the training phase,we select the batchsize=s pairs of source images as input. First,we train the discriminator M times to train the model well. Then we train the generator once. Repeat the above operation until the max iterations K. The optimizer is selected as in Adam[36]. In this experiment, the experimented settings are given as following: the discriminator training times $M = 2$, the network max iterations $K = 50$, and the batch size $s = 32$.

Generally, the discriminator converges more easily than the generator. As Traditional GAN networks are often used to transform noise into fake images, our discriminator only completes the classification task so it converges fast, and we train the discriminator M times first. Therefore, it is easy for the generator to generate a fused image, after the features are extracted and fused in the middle layers. In addition, the source images are input to the generator, and we also concatenate the source images to the middle layers as shown in Fig.3.

In the test phase, the GAN cuts off the part of the discriminator network, and the end-to-end image fusion test network is obtained. The generator network is a full-convolutional network, we input the source images without crop into the network to obtain the final fused image. We train and test using the NVIDIA GTX1080 and 64GB of memory.

4.2. Fusion Evaluation

The evaluation of image fusion quality remains an unsolved problem. The reason is that the same algorithm has different fusion effects for different kind of images. The fusion effect of the same algorithm for the same image is evaluated differently

Table 1: The architecture of Generator network.

Layer	Size (kernel)	Stride	Channel (input)	Channel (output)	Size (input)	Size (output)	Activation
Conv(C1)	5×5	1	2	256	120*120	120*120	LeakyReLU
Conv(C2)	5*5	1	258	128	120*120	120*120	LeakyReLU
Conv(C3)	3*3	1	386	64	120*120	120*120	LeakyReLU
Conv(C4)	3*3	1	450	32	120*120	120*120	LeakyReLU
Conv(C5)	1*1	1	482	1	120*120	120*120	Tanh

Table 2: The architecture of Discriminator network.

Layer	Size (kernel)	Stride	Channel (input)	Channel (output)	Size (input)	Size (output)	Activation
Conv(C1)	3×3	1	2	32	120×120	120×120	LeakyReLU
Pooling	2×2	2	32	32	118×118	118×118	
Conv(C2)	3×3	1	32	64	57×57	57×57	LeakyReLU
Pooling	2×2	2	64	64	57×57	28×28	
Conv(C3)	3×3	1	64	128	28×28	26×26	LeakyReLU
Pooling	2×2	2	128	128	26×26	13×13	
Conv(C4)	3×3	1	128	256	13×13	11×11	LeakyReLU
Pooling	2×2	2	256	256	11×11	5×5	
Reshape			256	1	5×5	1×6400	
FC	6400×1	1	1	1	1×6400	1	Tanh

Algorithm 1 Image Fusion Based On GANs With Denseblock

for GANs training iterations K **do**
  for D training iterations M **do**
    Select s patches I_{vi} from test data set
    Select s patches I_f from generated set
    AdamOptimizer loss fuction Eq.(4)
    Update discriminator
  end for
  Select s patches I_{ir} from test data set
  AdamOptimizer loss fuction Eq.(6)
  Update Generator
end for

because of the different interests of the observer. For applications in different fields, the requirements for images containing information are also different. Therefore, the fusion evaluation method of this paper combines two aspects: subjective evaluation and the another is objective evaluation.

4.2.1. Subjective Evaluation

Fusion images are often used in production or in practice. Therefore, subjective evaluation used here is the visual effect of fused images, such as color, lightness, fidelity and other more abstract judgments. That is, whether the fusion image gives people a satisfactory feeling. The method of subjective evaluation in this paper is to list different algorithms for the fusion images of the same groups of images and directly compare the visual effects.

4.2.2. Objective Evaluation

However, subjective evaluation is very personal. Hence, ten important objective evaluations are introduced for comparison. They are Edge Intensity (EI)[37], Cross Entropy(CE)[38], (SF)[39], Entropy (EN)[40], quality of images ($Q^{ab/f}$)[37], Spatial frequency Sum of Correlation Coefficients (SCD)[41], noise in images ($N^{ab/f}$)[38], Mutual Information between images ($MI^{ab/f}$)[42], Visual Information Fidelity (VIF)[43], Standard Deviation of Image (SD)[44], Definition (DF)[45] and a novel Visual Information Fidelity (VIFF) [46].

EI mainly represents the edge information and the contrast strength of neighbouring pixels. SF indicates the number of mutations in the image such as edges and lines. EN is defined based on information theory and can measure the amount of information contained in a fused image. The SCD calculates the difference image of the source image and the fused image, and then adds the correlation coefficients as the value of the SCD, reflecting the association between the source image and the difference image. The larger the CrossEntropy value, the greater the information difference between images. MI represents the degree of correlation between the two images. Both VIF and VIFF are used to measure the loss of image information to the distortion process. DF indicates whether the image is clear or not. The standard deviation can represent the contrast of the image to a certain extent.

Among them, the lower the value of the CrossEntropy and the $N^{ab/f}$, and the higher other values, the better the fusion quality of the image.

4.3. Comparison with State Of The Art Methods

In this section, the proposed algorithm is compared to the latest or classic fusion algorithms, including Laplacian Pyramid (LP)[47], Ratio of Low-pass Pyramid (RP)[48], Wavelet[49], Dual-Tree Complex Wavelet Transform (DTCWT)[50], Curvelet Transform (CVT)[51], Multi-resolution Singular Value Decomposition (MSVD)[52], DenseFuse[21], CNN[19], Deepfuse[53], FusionGan[24], DDcGan[29], ResNetFusion[30], Nestfuse[54] , night-vision context enhancement(NVCE)[55], FusionDN[56], HybridMSD[57], PMGI[58], infrared feature extraction and visual information preservation fusion(IFEVIP)[59], Structaware[60], U2Fusion[61], and MEF-GAN[25], respectively.

These methods can be divided into the following categories.

1. Methods based on image processing without deep learning, such as LP, RP, DTCWT, CVT, MSVD, NVCE, HybridMSD, IFEVIP. They use filters or image decomposition for image feature extraction and fusion.

2. Methods based on deep learning autoencoders, such as Densefuse, Deepfuse, Nestfuse. They use encoders to extract features and perform feature fusion at the latent layer.

3. Some methods are based on one-step fusion network such as FusionDN, PMGI, and U2fuion, using a reasonable network structure and loss function.

4. Based on the GAN network, this is the main comparison method of our method, such as FusionGAN, DDcGAN, ResNetFusion, and MEFGAN. Their generator is similar to the third one-step method, but uses a discriminator for adversarial generation to improve the quality of the fusion image.

4.3.1. Subjective Evaluation

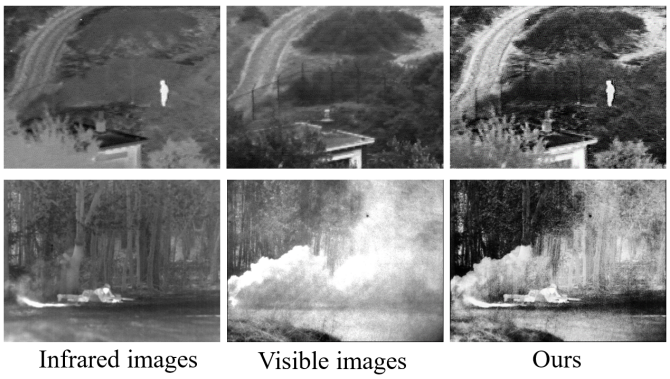


Figure 6: Fusion image of the method proposed in this paper

As shown in Fig.6, the end-to-end GAN has the advantage that the generated image can contain not only the source image information, but also rich contrast and gradient information. Our algorithm expands the distinction between objects in the image and the surrounding environment. The edge texture

Table 3: Objective evaluation of classic and latest fusion algorithms. This is the result on the TNO dataset.

Methods	EI	CE	SF	EN	$Q^{ab/f}$	SCD	$N^{ab/f}$	$MI^{ab/f}$	VIF	SD	DF	VIFF
CBF	52.3045	1.3163	13.0921	6.8275	0.4539	1.3193	0.0621	13.6550	0.6521	34.6491	6.4624	0.2834
CVT	42.9631	1.5894	11.1129	6.4989	0.4946	1.5812	0.0275	12.9979	0.5930	27.4613	5.4530	0.3316
DTCWT	42.4889	1.6235	11.1296	6.4791	0.5276	1.5829	0.0232	12.9583	0.5986	27.3099	5.4229	0.3231
GTF	35.0073	1.1440	9.5022	6.5781	0.4412	1.0516	0.0126	13.1562	0.4071	30.1806	4.6583	0.2072
LP	44.7055	1.4291	11.5391	6.6322	***0.5923***	1.5920	0.0237	13.2644	0.7721	30.5623	5.7422	0.4120
MSVD	27.6098	1.4202	8.5538	6.2807	0.3331	1.5857	***0.0022***	12.5613	0.3031	24.0288	4.2283	0.2768
RP	44.9054	1.3420	12.7249	6.5397	0.4351	1.5769	0.0583	13.0794	0.6420	28.8385	6.1799	0.2919
Wavelet	24.6654	1.4278	6.2567	6.2454	0.3219	1.5877	**0.0000**	12.4907	0.2921	23.6867	3.1353	0.2742
CNN	44.8334	**0.8939**	11.6483	7.0629	0.5833	1.6006	0.0234	14.1259	0.7872	44.9159	5.7266	0.4504
DeepFuse	33.8768	1.7779	8.3500	6.6102	0.3850	1.5523	0.0202	13.2205	0.5752	31.5605	4.3660	0.4898
DenseFuse	36.4838	1.4015	9.3238	6.8526	0.4744	1.5329	0.0352	13.7053	0.6875	38.0412	4.6176	0.3876
FusionGan	32.5997	1.9353	8.0476	6.5409	0.2685	0.6876	0.0352	13.0817	0.4928	29.1495	4.2727	0.1473
IFCNN	44.9725	1.4413	11.8590	6.6454	0.4968	1.6126	0.0346	13.2909	0.6090	33.0086	5.9808	0.3599
MDLatLRR	***86.8855***	1.0159	***21.6245***	7.0897	0.3238	1.5641	0.3021	14.1793	***2.2258***	43.0009	***11.0665***	***0.8776***
DDcGAN	54.8233	1.0979	12.8351	***7.4215***	0.3613	1.3831	0.1323	***14.8429***	1.0206	46.7319	6.9395	0.4634
ResNetFusion	26.2360	1.6188	5.9182	6.6546	0.0958	0.1937	0.0550	13.3092	0.4526	46.9280	2.5853	0.0421
NestFuse	38.4401	1.4458	9.7098	6.8856	0.5002	1.5839	0.0491	13.7713	0.7236	38.3311	4.9099	0.3744
NVCE	60.1502	1.8396	15.2304	7.1440	0.4576	1.5155	0.1338	14.2879	1.2602	39.4821	7.9169	0.6490
FusionDN	61.3491	1.2903	14.2256	7.4073	0.3826	***1.6148***	0.1540	14.8147	1.3721	***48.5659***	7.4565	0.7994
HybridMSD	46.3200	1.3157	12.3459	6.7103	0.5014	1.5773	0.0435	13.4206	0.7872	31.2742	6.0954	0.4448
PMGI	37.2133	1.5656	8.7194	6.8688	0.3790	1.5738	0.0340	13.7376	0.6904	33.0167	4.4328	0.4237
IFEVIP	37.2133	1.5656	8.7194	6.8688	0.3790	1.5738	0.0340	13.7376	0.6904	33.0167	4.4328	0.4237
StructAware	44.9083	***0.9165***	11.1984	7.0549	**0.6228**	1.2611	0.0147	14.1098	0.8350	40.3137	5.4062	0.3524
U2Fusion	48.4915	1.3255	11.0368	6.7227	0.3937	1.5946	0.0800	13.4453	0.7680	31.3794	5.8343	0.5331
MEF-GAN	36.9050	1.4980	7.8481	6.9727	0.2076	1.3083	0.0750	13.9454	0.7330	43.7332	3.8742	0.3240
Ours	**88.3545**	1.3337	**21.8903**	**7.7660**	0.2288	**1.6564**	0.3113	**15.5319**	**2.3350**	**68.8495**	**11.6163**	**1.1901**

information in the visible images, as well as the desired radiation information in the infrared image, is better reflected in the fused images.

In the experiment, the subjective evaluation of the fusion image of the algorithm is compared with the 21 pairs of typical images selected from the TNO. For each group of images, the first two pictures are the visible light image and the infrared image. The following images are the results of other approaches, except the last image is ours.

As a subjective evaluation, it can be clearly seen that the fusion images of most methods are well fused with infrared and visible information. Obviously, the image quality of our proposed method is higher.

In Fig.11, for the details of houses and plants, most of the methods have serious information loss or low contrast, and the fusion result is unnatural. For example, the dustbin in front of the house and the man in the phone booth are obvious in our method. The foreground texture of the sky composed of the branches of the tree is obvious. What is more, our images look quite natural. In the more natural results from DenseFuse, the infrared details are lost. Furthermore our method retains a lot of obvious plant details, as well as house wall information.

In Fig.12 and Fig.13, our method is equally great. Our images retain not only sufficient infrared radiation information, but also the complex texture information of the background, and the display of small targets is also very clear and obvious. As an important point of subjective evaluation, the perception of our images is very good and looks natural.

At the same time, we selected and displayed two of the subjective evaluations on the RoadScene dataset, as shown in Fig.14 and Fig.15. Our result has more obvious texture information, and the image looks more natural.

Obviously, most of the results generated by other methods are similar to the interpolated images of infrared images and visible light images, and the generated results look greyish overall. Moreover, our method has a visually impressive effect. This is because we use the more perceptual SSIM loss and gradient loss to replace the simple and crude mean square error loss. SSIM loss focuses on the structural information of the image, including contrast information. Gradient loss ensures that the image maintains sufficient edge contours information. Last but not least, GAN loss makes the generated image have a better visually impressive effect.

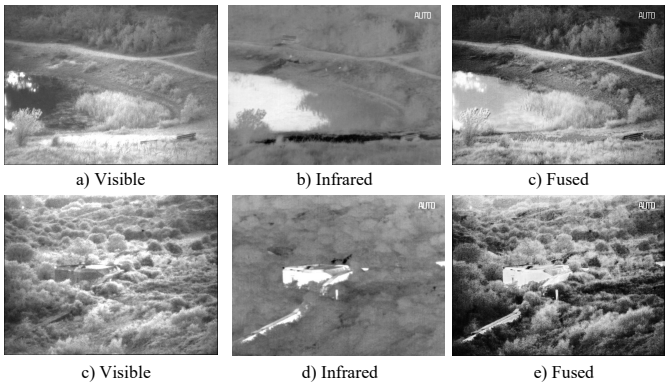
As shown in the Fig.7, there have been some failure cases. For example, the dividing line between the water surface and the grass is weakened due to decreased contrast. The bunker in the cloud is not as obvious as in the infrared image. We believe that our proposed method will focus on high brightness areas and improve the overall contrast of the image. Its disadvantage is that for areas where both images are highlighted, the dividing line may be lost.

4.3.2. *Objective Evaluation*

For the objective evaluation, we select twelve fusion evaluation indicators for comparison, and the results are shown in Table 3. The best value in the quality table is made bold with

Table 4: Objective evaluation of classic and latest fusion algorithms. This is the result on the RoadScene dataset.

Methods	EI	CE	SF	EN	Qabf	SCD	Nabf	MI	VIF	SD	DF	VIFF
CBF	67.6171	0.8217	16.4106	7.5322	0.4971	0.8784	0.0388	15.0645	0.6744	51.9820	7.9937	0.3594
CVT	59.7642	1.1498	14.7379	7.0159	0.4951	1.3418	0.0318	14.0319	0.6627	36.0884	6.9618	0.4009
DTCWT	57.3431	1.2475	14.7318	6.9211	0.4660	1.3329	0.0412	13.8421	0.6257	34.7264	6.7810	0.3673
GTF	37.2992	0.7934	10.1343	7.6346	0.3816	0.8072	0.0096	15.2693	0.4247	59.7582	4.3501	0.2492
LP	59.5437	1.1479	15.3634	7.0348	***0.6036***	1.3617	0.0250	14.0695	0.7799	37.3478	7.1062	0.4602
MSVD	36.0475	1.0734	11.3182	6.6960	0.3636	1.3458	0.0034	13.3919	0.3472	30.9643	5.0926	0.3211
RP	70.8057	1.1250	19.1529	7.0553	0.4907	1.2829	0.0773	14.1107	0.8366	38.4519	8.8084	0.4244
Wavelet	33.5603	1.0353	8.4818	6.6670	0.3743	1.3522	**0.0000**	13.3339	0.3422	30.5847	3.9395	0.3273
CNN	58.2838	0.9083	15.1313	7.2442	0.5884	1.4033	0.0241	14.4884	0.7666	45.0176	6.9601	0.4714
DeepFuse	**100.1552**	1.3153	***25.0937***	7.6088	0.2948	0.5462	0.2213	15.2177	0.7410	56.0832	**12.3650**	0.1969
DenseFuse	34.0135	1.0165	8.5541	6.6740	0.3814	1.3491	***0.0000***	13.3480	0.3476	30.6655	3.9885	0.3316
FusionGan	35.4048	1.8991	8.6400	7.1753	0.2737	0.8671	0.0168	14.3507	0.4256	42.3040	3.9243	0.2720
IFCNN	57.6653	1.1486	15.0677	6.9730	0.5150	1.3801	0.0315	13.9460	0.6249	35.8183	7.0401	0.3790
MDLatLRR	36.9468	0.9920	9.3638	6.7171	0.4493	1.3636	0.0000	13.4342	0.3920	31.3505	4.3216	0.3543
DDcGAN	98.1552	1.3153	***25.0937***	7.6088	0.2948	0.5462	0.2213	15.2177	0.7410	56.0832	**12.3650**	0.1969
ResNetFusion	39.4317	1.1539	8.4967	7.3401	0.1052	0.2179	0.0550	14.6801	0.2874	62.8924	3.8041	0.0483
NestFuse	54.4953	1.2820	14.4674	7.3731	0.4923	1.2583	0.0459	14.7461	0.9263	49.7441	6.3461	0.4902
NVCE	81.5118	0.9634	20.5063	7.3636	0.4599	1.3535	0.1255	14.7271	***1.3781***	47.7500	9.7719	0.6362
FusionDN	68.1690	1.3324	16.7138	7.5323	0.4392	1.1882	0.0780	15.0646	1.0086	55.0559	7.7925	***0.6387***
HybridMSD	62.2138	1.1032	16.7475	7.0256	0.5376	1.2642	0.0460	14.0512	0.8064	37.1333	7.5600	0.4791
PMGI	47.2067	1.6395	10.9368	7.3493	0.4248	1.0989	0.0146	14.6986	0.6461	49.3262	5.1288	0.4518
IFEVIP	47.6046	**0.5854**	13.1474	6.6331	0.4898	***1.4049***	0.0227	13.2661	0.6806	39.2868	5.5453	0.3760
StructAware	59.6608	***0.6626***	14.8652	***7.6551***	**0.6242**	0.7933	0.0149	***15.3102***	0.8035	57.5360	6.7881	0.3988
U2Fusion	66.2529	1.4806	15.8242	7.1969	0.4805	1.3551	0.0671	14.3938	0.8317	42.9368	7.5930	0.5462
MEF-GAN	50.5294	1.2160	11.9159	7.1613	0.1960	1.0967	0.0824	14.3226	0.6073	**69.2043**	5.1724	0.2839
Ours	***99.1008***	1.1223	**26.6063**	**7.6897**	0.3563	**1.4524**	0.2284	**15.3793**	**1.5125**	***63.6454***	12.3176	**0.8260**



a) Visible b) Infrared c) Fused

c) Visible d) Infrared e) Fused

Figure 7: some failure cases of image fusion.

the bold red font, and the second-best value in the black box is given in italic. It can be seen that our proposed method has nine best values (EI, SF, EN, and SCD, MI, VIF, SD, DF, VIFF). It has the best value in EI and SF, which means that our proposed algorithm's fusion image extracts more edge information and detail information. The best value is also obtained in EN, which means that the fused image contains more information. In the SCD and MI indicators, two first optimal values are obtained, indicating that the fused image sufficiently retains the information of the source images, and the fused image is more similar to the source images. It has the best value in VIF and VIFF, which means that the image is more realistic or natural. The two optimal values of DF and SD indicate that our images and have high image contrast.

The bad results of $Q^{ab/f}$ and $N^{ab/f}$ mean that our fusion image may have more noise. We believe that the reason for this result is that the loss function lacks the pixel-level mean square error loss and the use of adversarial loss. Because the mean square error loss allows the network to learn an average distribution of all sample distributions, the generated image will not have too much noise. In the generation of the GAN network, the generator is easier to generate a "real" sample distribution. And at the same time, it is also guided by the gradient loss, the network is easier to generate images with noise. Of course, there is another reason that the visible light images in the data set have much noise, and the discriminator uses this noise as a criterion.

We also make an objective evaluation on the RoadScene[56] data set as shown in Table 4, and also obtained similar results. We obtained six best values and two second best values.

4.4. Ablation Studies

4.4.1. Network Structure

To prove the effectiveness of our network structure, we make some ablation modifications to the network, as shown in Fig .9. We keep or remove the concatenation of visible light images and denseblock alternately, as shown in the Fig .9(a)(b). If we do not keep both of them, the generator network is just a multi-layer convolution network, as shown in the Fig .9 (c).

We also conduct subjective and objective evaluations of these ablation networks. As shown in the Fig .8, when we remove

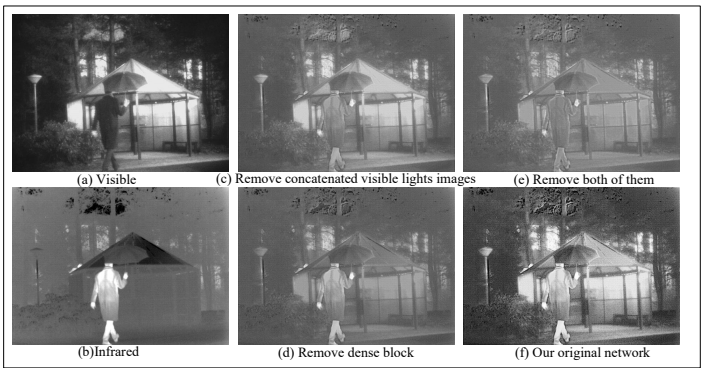


Figure 8: Images of ablation networks.

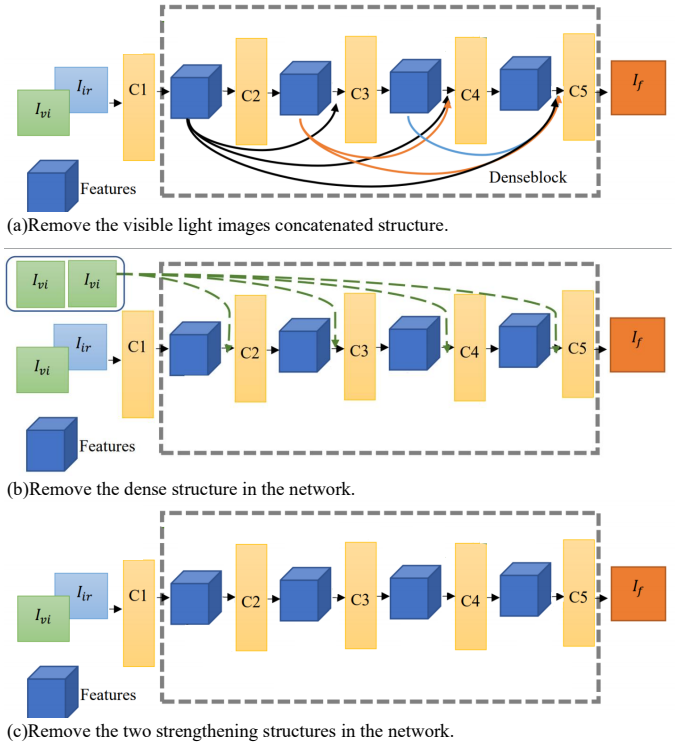


Figure 9: Remove dense connections or infrared light image connections, or both of them.

a part of the network structure, the image quality deteriorates significantly. The images lose details of the vegetation, and the radiation information of the person is not enough. The overall perception of the images is satisfactory.

As shown in the Table .5, we conducted ablation experiments on the network structure and evaluated them with objective indicators. Here, the network the concatenated visible light image is called "without viscat". It's easy to understand that "without Dense" is the network with the dense block removed. And, the network which removed both of them is called "without both". It can be observed through the table that these ten evaluation indicators become worse no matter which part is removed.

4.4.2. Hyperparameters Comparison

Referring to Equation 6-10, our loss function includes the loss of LSGAN and the loss of content. We change the formula to a different style in order to make it easier to demonstrate the effect of the value of the hyperparameters in the content loss function, as shown in Equation 12.

$$L_{content} = \lambda_1 \left[1 - \left(SSIM\left(I_f, I_{vi}\right) + SSIM\left(I_f, I_{ir}\right)/2\right)\right] + \lambda_2 \left[Gradient\left(I_f, I_{vi}\right) + \lambda_3 Gradient\left(I_f, I_{ir}\right)\right] \quad (12)$$

The hyperparameters used to balance them are: λ_1 is taken from $\{0, 1, 10, 100\}$, λ_2 is taken from $\{0, 10, 100, 1000, 10000\}$, and λ_3 is taken from $\{0.5, 1, 2\}$. Through these ablation experiments, as shown in Table .6, Table .7 and Table .8 respectively, we obtain suitable hyperparameter settings.

Table 5: Objective evaluation of classic and latest fusion algorithms.

Methods	EI	Cross Entropy	SF	EN	SCD
Ours	88.3545	1.0680	21.8903	7.7660	1.6564
without viscat	41.3922	2.0567	10.6261	6.6566	1.6207
without Dense	33.7982	1.4966	8.1133	6.6582	1.5699
without both	40.4124	1.8372	9.8722	6.5489	1.5176
Methods	MI	VIF	SD	DF	VIFF
Ours	15.5319	2.3350	68.8495	11.6163	1.1901
without viscat	13.3131	0.5046	31.4112	5.4468	0.3710
without Dense	13.3165	0.5467	32.4376	4.0958	0.4209
without both	13.0977	0.5214	26.3253	5.1762	0.3698

Table 6: The Weight of SSIM Loss: λ_1

λ_1	EI	CrossEntropy	SF	EN	SCD	MI
0	37.63	1.91	9.14	6.78	1.11	13.56
10	39.39	2.05	9.73	6.72	1.39	13.44
100	***89.06***	***1.50***	***22.28***	***7.78***	***1.63***	***15.57***
1000	40.68	2.16	9.88	6.55	1.57	13.10

Table 7: The Weight of Gradient Loss: λ_2

λ_2	EI	CrossEntropy	SF	EN	SCD	MI
0-0	34.55	1.71	8.54	6.72	1.70	13.43
5-10	36.90	***1.35***	8.96	6.92	***1.73***	13.83
50-100	40.08	1.78	10.10	6.76	1.69	13.52
500-1000	***89.06***	1.50	***22.28***	***7.78***	1.63	***15.57***
5000-10000	39.11	2.41	9.85	6.44	1.41	12.88

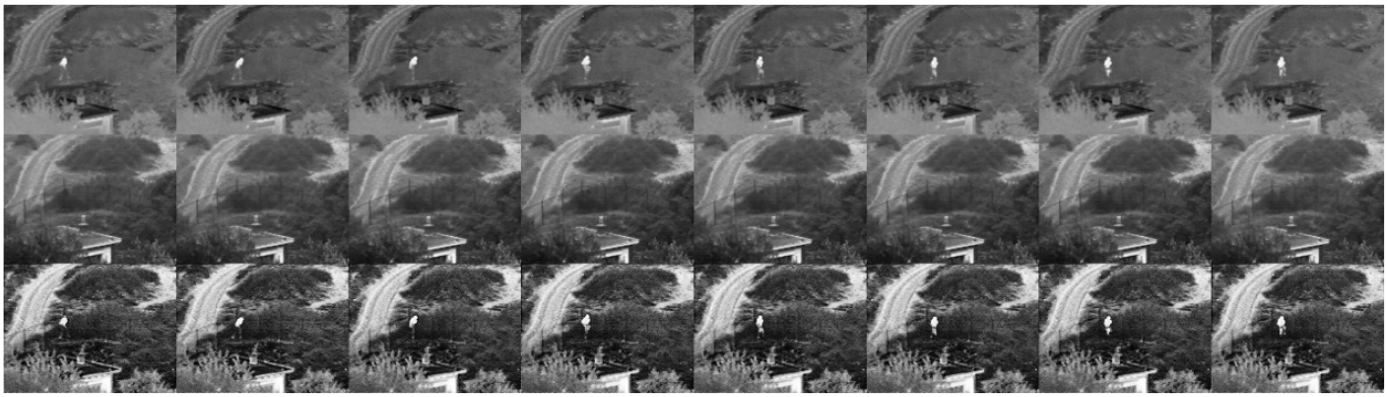


Figure 10: Image fusion of video sequences.The three rows are, respectively, an infrared video sequence, a visible video sequence and a fused image sequence.

Table 8: The Weight Between Gradient Loss of Infrared and Visible Light Images: λ_3

λ_3	EI	CrossEntropy	SF	EN	SCD	MI
0-0	34.55	1.71	8.54	6.72	1.70	13.43
5-10	36.90	***1.35***	8.96	6.92	***1.73***	13.83
50-100	40.08	1.78	10.10	6.76	1.69	13.52
500-1000	***89.06***	1.50	***22.28***	***7.78***	1.63	***15.57***
5000-10000	39.11	2.41	9.85	6.44	1.41	12.88

4.5. Video image fusion

Since the generator only has five layers, the image fusion speed is very fast. A set of images in the TNO data is selected, and some of these are shown in Fig.10. The pixel size of the test images is 360×270, and an image file size is 95.9 kb. In the experiment, 32 images were copied, and the number was expanded to 1000. On the NVIDIA GTX1080, using a simplified network to read and fuse 1000 frames takes 21.52 seconds. FPS can reach 46, faster than most image fusion algorithms, and it can meet real-time requirements. In Fig. 10, the first row and the second row are, respectively, an infrared video sequence and a visible video sequence, and the third row is a fused image sequence. It can be clearly seen that the fused video retains both the details of the visible details and the primary radiation target information of the infrared video.

Compared with the traditional methods, it is easier for us to use CUDA (Compute Unified Device Architecture by NVIDIA) to perform matrix operations on the GPUs to accelerate the operation speed of the deep network, instead of matrix optimization from the bottom.

5. Acknowledgement

This work was supported by the National Natural Science Foundation of China (62020106012, U1836218, 61672265), and the 111 Project of Ministry of Education of China (B12018).

6. Conclusions

In this paper, based on a simple end-to-end network, we add denseblock to improve the fusion effect. In addition, we innovatively concatenate visible images at each layer, so that the fused image retains more visible information. We design new loss functions, which leads the end-to-end network to generate the fusion image better and more stable. The SSIM loss function is added to maintain fused image structure similar to the source images, and the gradient loss is added to generate more edge information. We did not use the general mean square error loss, because the mean square error makes the data distribution learned by the network more "average", which does not fit the function of the GAN network-that is to generate sufficiently close to the "real" data distribution. The discriminator classifies the visible image to make the generated image more realistic and natural, which is more in line with human perception. The test model is the generator model, so the fused network is very simple. It is an end-to-end network that does not require feature extraction or design fusion strategies. Therefore, videos can be fused in real time. Our methods can fuse more source image details than classic and including the latest methods. The fused image is more natural and satisfies the visual aesthetics of human beings. In terms of objective evaluation, our fusion method achieves the best value based on several evaluation indicators. In the future work, we will try to replace the new discriminator network and strengthen the discriminator function. Furthermore, we will design a new generator loss function to expand the application area of image fusion. We will also continue to study simplified network to improve its results; and will try to make the GAN solve other problems in the field of image fusion, such as multi-focus image fusion, medical image fusion and address other issues.

References

[1] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: A survey, Information Fusion 45 (2019) 153–178.

[2] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: A survey of the state of the art, information Fusion 33 (2017) 100–112.

[3] T. Mertens, J. Kautz, F. Van Reeth, Exposure fusion: A simple and practical alternative to high dynamic range photography, Computer Graphics Forum 28 (2009) 161–171.

[4] Z. Zhang, R. S. Blum, A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application, Proceedings of the IEEE 87 (1999) 1315–1326.

[5] K. P. Upla, M. V. Joshi, P. P. Gajjar, An edge preserving multiresolution fusion: Use of contourlet transform and mrf prior, IEEE Transactions on Geoscience and Remote Sensing 53 (2014) 3210–3220.

[6] J.-j. Zong, T.-s. Qiu, Medical image fusion based on sparse representation of classified image patches, Biomedical Signal Processing and Control 34 (2017) 195–205.

[7] Q. Zhang, Y. Fu, H. Li, J. Zou, Dictionary learning method for joint sparse representation-based image fusion, Optical Engineering 52 (2013) 057006.

[8] Y. Bin, Y. Chao, H. Guoyu, Efficient image fusion with approximate sparse representation, International Journal of Wavelets, Multiresolution and Information Processing 14 (2016) 1650024.

[9] H. Li, X.-J. Wu, Multi-focus image fusion using dictionary learning and low-rank representation, in: International Conference on Image and Graphics, Springer, 2017, pp. 675–686.

[10] D. P. Bavirisetti, G. Xiao, G. Liu, Multi-sensor image fusion based on fourth order partial differential equations, in: 2017 20th International Conference on Information Fusion (Fusion), IEEE, 2017, pp. 1–9.

[11] N. Cvejic, D. Bull, N. Canagarajah, Region-based multimodal image fusion using ica bases, IEEE Sensors Journal 7 (2007) 743–751.

[12] J. Mou, W. Gao, Z. Song, Image fusion based on non-negative matrix factorization and infrared feature extraction, in: 2013 6th International Congress on Image and Signal Processing (CISP), volume 2, IEEE, 2013, pp. 1046–1050.

[13] X. Zhang, Y. Ma, F. Fan, Y. Zhang, J. Huang, Infrared and visible image fusion via saliency analysis and local edge preserving multi-scale decomposition, Journal of The Optical Society of America A-optics Image Science and Vision 34 (2017) 1400–1410.

[14] Mengfanjie, Songmiao, Guobaolong, Shiruixia, Shandalong, Image fusion based on object region detection and non-subsampled contourlet transform, Computers & Electrical Engineering (2017).

[15] W. Kong, L. Zhang, Y. Lei, Novel fusion method for visible light and infrared images based on nsst–sf–pcnn, Infrared Physics & Technology 65 (2014) 103–112.

[16] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Information Fusion 24 (2015) 147–164.

[17] S. Yin, L. Cao, Q. Tan, G. Jin, Infrared and visible image fusion based on nsct and fuzzy logic, in: 2010 IEEE International Conference on Mechatronics and Automation, IEEE, 2010, pp. 671–675.

[18] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, X. Wang, Deep learning for pixel-level image fusion: Recent advances and future prospects, Information Fusion 42 (2018) 158–173.

[19] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Information Fusion 36 (2017) 191–207.

[20] A. Azarang, H. Ghassemian, A new pansharpening method using multi resolution analysis framework and deep neural networks (2017) 1–6.

[21] H. Li, X.-J. Wu, Densefuse: A fusion approach to infrared and visible images, IEEE Transactions on Image Processing 28 (2018) 2614–2623.

[22] H. Li, X. Wu, J. Kittler, Infrared and visible image fusion using a deep learning framework (2018) 2705–2710.

[23] H. Li, X. Wu, T. S. Durrani, Infrared and visible image fusion with resnet and zero-phase component analysis, Infrared Physics & Technology 102 (2019) 103039.

[24] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, Fusiongan: A generative adversarial network for infrared and visible image fusion, Information Fusion 48 (2019) 11–26.

[25] H. Zhang, Z. Le, Z. Shao, H. Xu, J. Ma, Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion, Information Fusion 66 (2020) 40–53.

[26] H. Xu, J. Ma, X.-P. Zhang, Mef-gan: Multi-exposure image fusion via generative adversarial networks, IEEE Transactions on Image Processing (2020).

[27] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, J. Jiang, Pan-gan: An unsupervised learning method for pan-sharpening in remote sensing image fusion using a generative adversarial network, Information Fusion (2020).

[28] I. Goodfellow, J. Pougetabadie, M. Mirza, B. Xu, D. Wardefarley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets (2014) 2672–2680.

[29] J. Ma, H. Xu, J. Jiang, X. Mei, X. Zhang, Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion, IEEE Transactions on Image Processing 29 (2020) 4980–4995.

[30] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, J. Jiang, Infrared and visible image fusion via detail preserving adversarial learning, Information Fusion 54 (2020) 85–98.

[31] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, S. P. Smolley, Least squares generative adversarial networks (2017) 2813–2821.

[32] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1125–1134.

[33] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[34] A. L. Maas, A. Y. Hannun, A. Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proc. icml, volume 30, 2013, p. 3.

[35] A. Toet, et al., Tno image fusion dataset, Figshare. data (2014).

[36] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv: Learning (2014).

[37] C. Xydeas, , V. Petrovic, Objective image fusion performance measure, Electronics letters 36 (2000) 308–309.

[38] B. S. Kumar, Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform, Signal, Image and Video Processing 7 (2013) 1125–1143.

[39] A. M. Eskicioglu, P. S. Fisher, Image quality measures and their performance, IEEE Transactions on communications 43 (1995) 2959–2965.

[40] J. W. Roberts, J. A. van Aardt, F. B. Ahmed, Assessment of image fusion procedures using entropy, image quality, and multispectral classification, Journal of Applied Remote Sensing 2 (2008) 023522.

[41] V. Aslantas, E. Bendes, A new image quality metric for image fusion: the sum of the correlations of differences, Aeu-international Journal of electronics and communications 69 (2015) 1890–1896.

[42] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 1226–1238.

[43] H. R. Sheikh, A. C. Bovik, Image information and visual quality, IEEE Transactions on Image Processing 15 (2006) 430–444.

[44] Y.-J. Rao, In-fibre bragg grating sensors, Measurement science and technology 8 (1997) 355.

[45] X. Desheng, Research of measurement for digital image definition, Journal of Image and Graphics (2004).

[46] Y. Han, Y. Cai, Y. Cao, X. Xu, A new image fusion performance metric based on visual information fidelity, Information Fusion 14 (2013) 127–135.

[47] P. J. Burt, E. H. Adelson, The laplacian pyramid as a compact image code, IEEE Transactions on Communications 31 (1983) 671–679.

[48] A. Toet, Image fusion by a ration of low-pass pyramid., Pattern Recognition Letters 9 (1989) 245–253.

[49] L. J. Chipman, T. M. Orr, L. N. Graham, Wavelets and image fusion 3 (1995) 3248.

[50] J. J. Lewis, R. J. Callaghan, S. G. Nikolov, D. R. Bull, N. Canagarajah, Pixel-and region-based image fusion with complex wavelets, Information fusion 8 (2007) 119–130.

[51] F. Nencini, A. Garzelli, S. Baronti, L. Alparone, Remote sensing image fusion using the curvelet transform, Information fusion 8 (2007) 143–156.

[52] V. Naidu, Image fusion technique using multi-resolution singular value decomposition, Defence Science Journal 61 (2011) 479.

[53] K. R. Prabhakar, V. S. Srikar, R. V. Babu, Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs., in: ICCV, 2017, pp. 4724–4732.

[54] H. Li, X.-J. Wu, T. Durrani, Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models, IEEE Transactions on Instrumentation and Measurement (2020).

[55] Z. Zhou, M. Dong, X. Xie, Z. Gao, Fusion of infrared and visible images for night-vision context enhancement, Applied optics 55 (2016) 6480–

6490.

[56] H. Xu, J. Ma, Z. Le, J. Jiang, X. Guo, Fusiondn: A unified densely connected network for image fusion., in: AAAI, 2020, pp. 12484–12491.

[57] Z. Zhou, B. Wang, S. Li, M. Dong, Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters, Information Fusion 30 (2016) 15–26.

[58] H. Zhang, H. Xu, Y. Xiao, X. Guo, J. Ma, Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity., in: AAAI, 2020, pp. 12797–12804.

[59] Y. Zhang, L. Zhang, X. Bai, L. Zhang, Infrared and visual image fusion through infrared feature extraction and visual information preservation, Infrared Physics & Technology 83 (2017) 227–237.

[60] W. Li, Y. Xie, H. Zhou, Y. Han, K. Zhan, Structure-aware image fusion, Optik 172 (2018) 1–11.

[61] H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2fusion: A unified unsupervised image fusion network, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

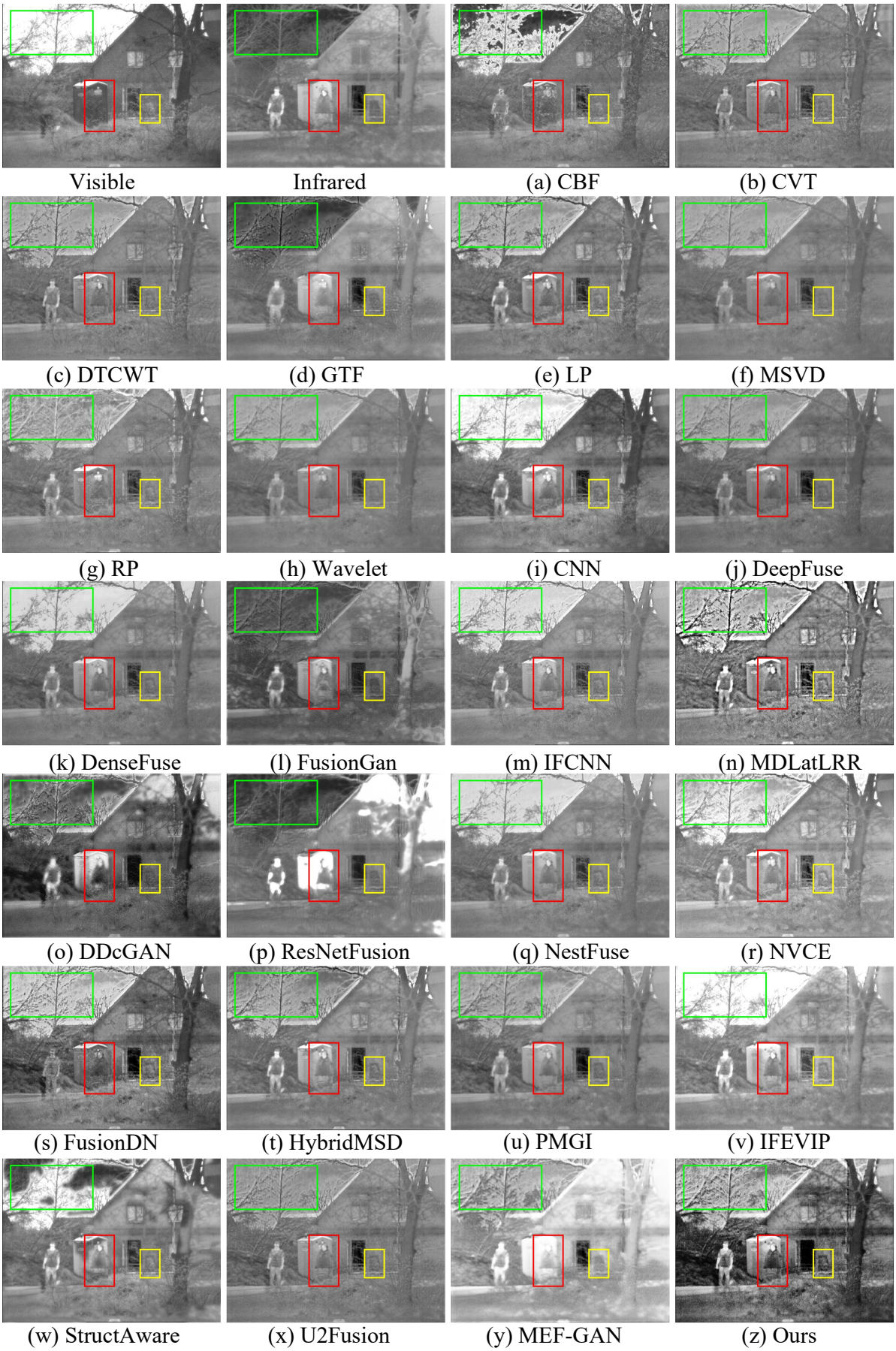


Figure 11: Subjective experiments of various methods on the "men and house" images. The image is elected from TNO data.

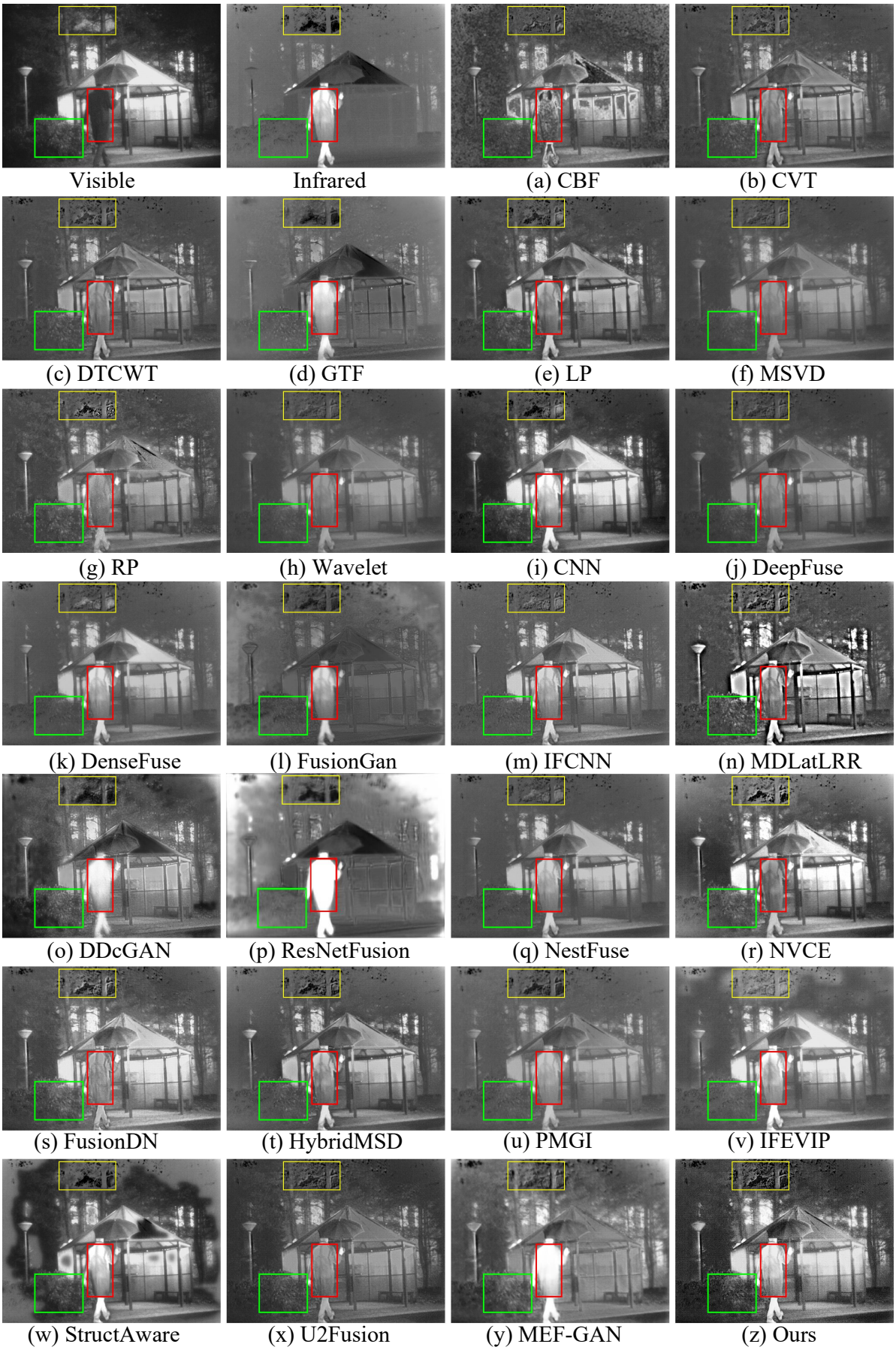


Figure 12: Subjective experiments of various methods on the "umbrella" images. The image is elected from TNO data.

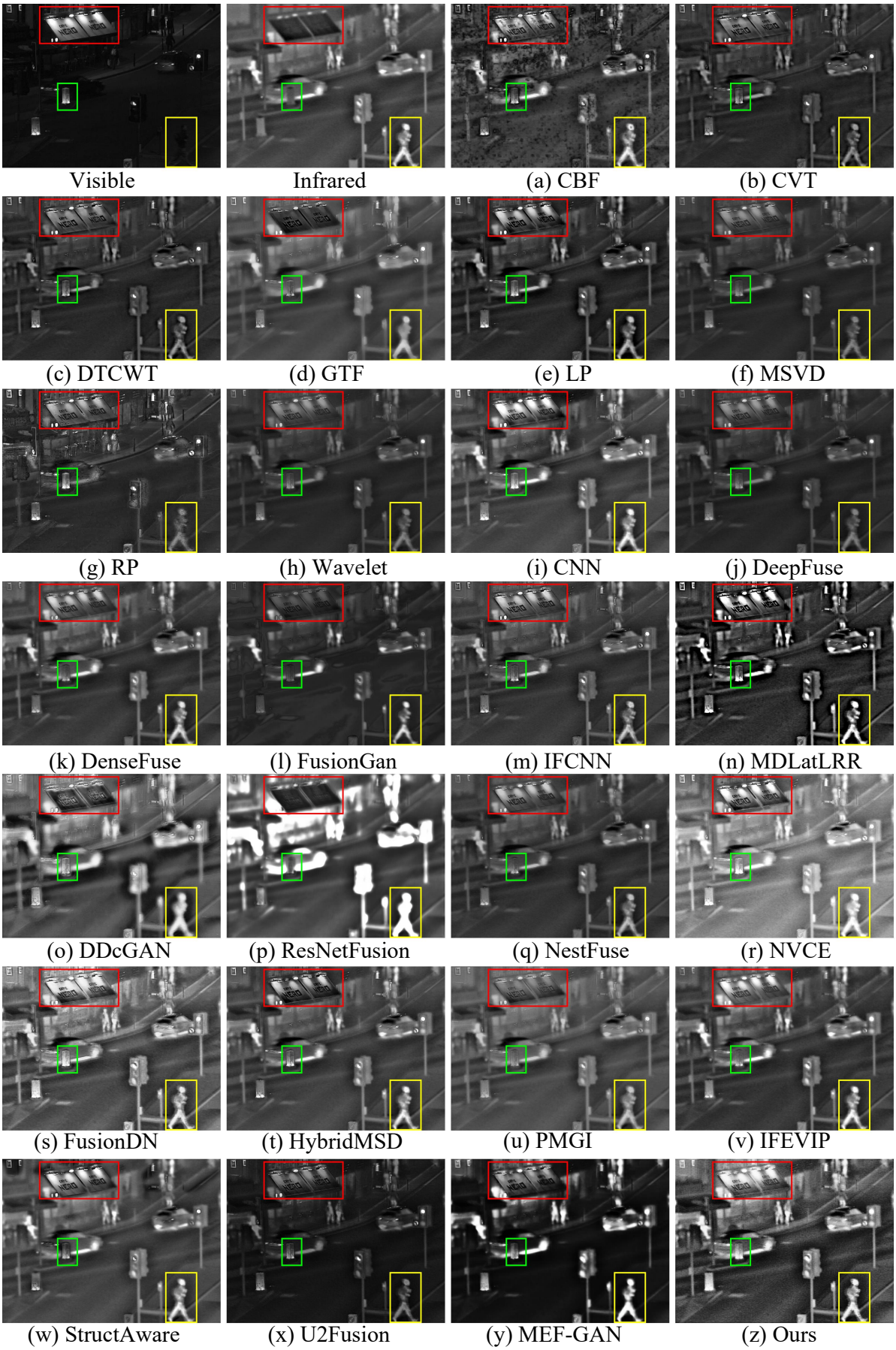


Figure 13: Subjective experiments of various methods on the "street" images. The image is elected from TNO data.

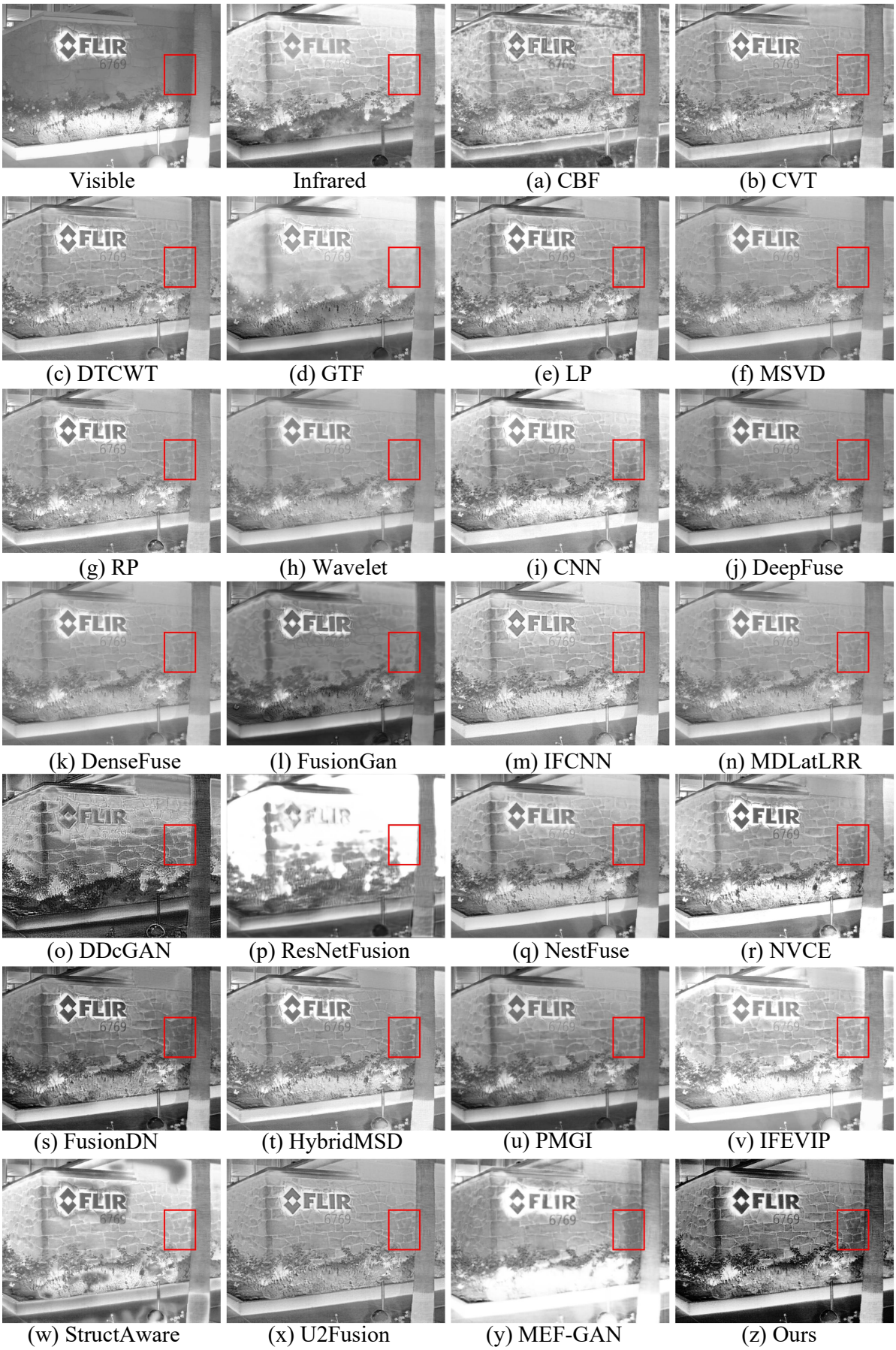


Figure 14: Subjective experiments of various methods on the "sign" images. The image is elected from RoadScene data.

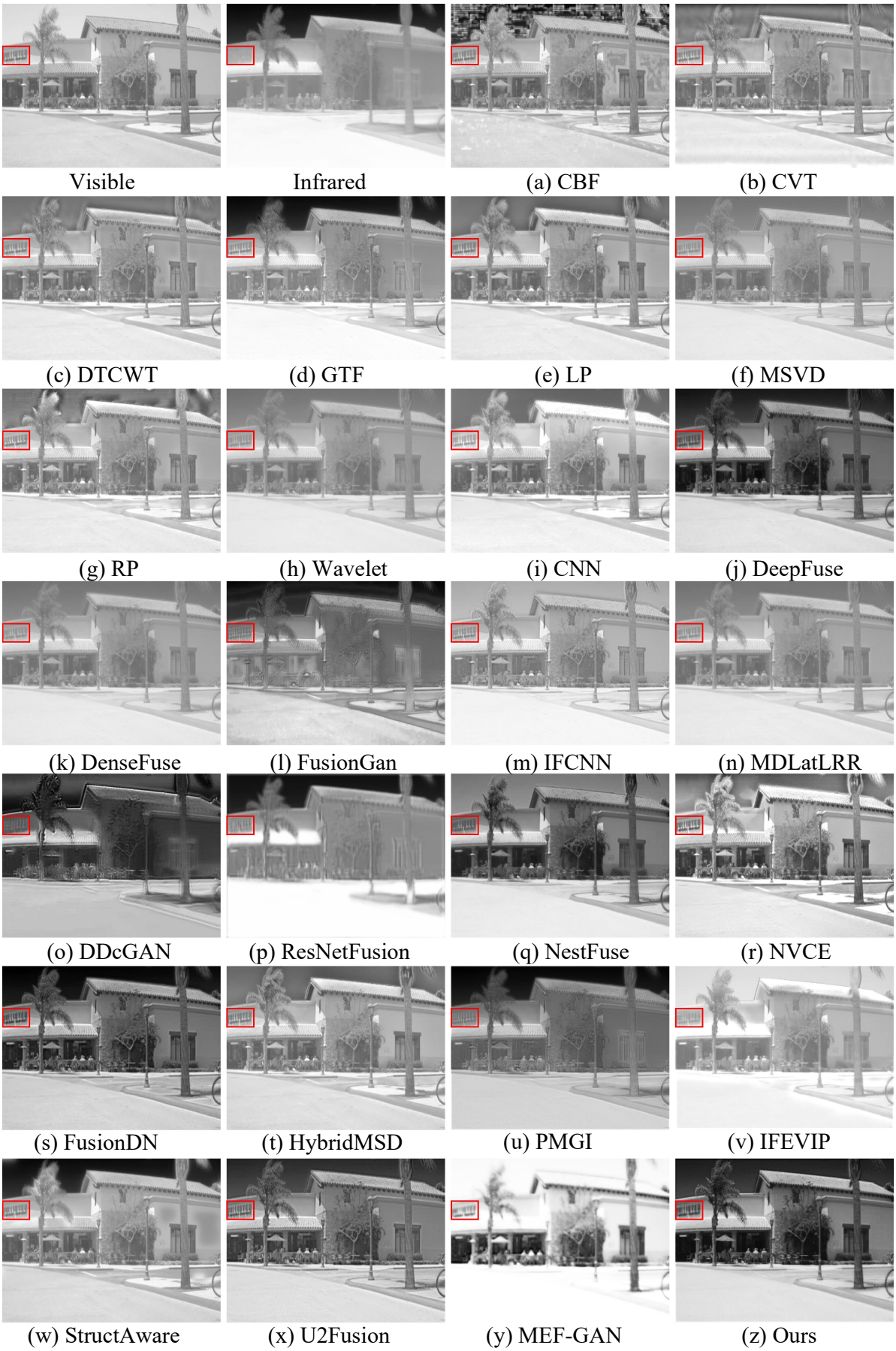


Figure 15: Subjective experiments of various methods on the "crossing" images. The image is elected from RoadScene data.