Infrared Image Super-Resolution via Heterogeneous Convolutional WGAN

Yongsong Huang[1] , Zetao Jiang[1(✉)] , Qingzhong Wang[2,3] , Qi Jiang[1],
and Guoming Pang[4]

[1] Guilin University of Electronic Technology University, Guilin, China
`zetaojiang@guet.edu.cn`
[2] City University of Hong Kong, Kowloon, Hong Kong SAR, China
`qingzwang2-c@my.cityu.edu.hk`
[3] Baidu Research, Beijing, China
[4] ZTE Corporation, Shenzhen, China
`pang.guoming@zte.com.cn`

Abstract. Image super-resolution is important in many fields, such as surveillance and remote sensing. However, infrared (IR) images normally have low resolution since the optical equipment is relatively expensive. Recently, deep learning methods have dominated image super-resolution and achieved remarkable performance on visible images; however, IR images have received less attention. IR images have fewer patterns, and hence, it is difficult for deep neural networks (DNNs) to learn diverse features from IR images. In this paper, we present a framework that employs heterogeneous convolution and adversarial training, namely, heterogeneous kernel-based super-resolution Wasserstein GAN (HetSRW-GAN), for IR image super-resolution. The HetSRWGAN algorithm is a lightweight GAN architecture that applies a plug-and-play heterogeneous kernel-based residual block. Moreover, a novel loss function that employs image gradients is adopted, which can be applied to an arbitrary model. The proposed HetSRWGAN achieves consistently better performance in both qualitative and quantitative evaluations. According to the experimental results, the whole training process is more stable.

Keywords: Super-resolution · Infrared image · Image processing · Heterogeneous kernel-based convolution · Generative adversarial networks

1 Introduction

Image super-resolution (SR) reconstruction is a very active topic in computer vision as it offers the promise of overcoming some of the limitations of low-cost imaging sensors. Infrared (IR) image super-resolution plays an important role in the military and medical fields and many other areas of vision research. A major problem with IR thermal imaging is that IR images are normally low resolution since the size and precision of IR sensors can be limited. Image super-resolution is a promising and low-cost way to improve the resolution and quality

of IR images. Generally, image super-resolution methods based on deep learning can be classified into two categories, namely, models based on generative adversarial networks (GANs) [16,20] and models based on deep neural networks (DNNs) [5,6,8,12,18,21,26,27], both of which have achieved satisfying results on visible images. These methods can achieve a good peak signal-to-noise ratio (PSNR). However, they do not consider the visual characteristics of the human eye. The human eye is more sensitive to contrast differences with a lower spatial frequency. The sensitivity of the human eye to differences in brightness contrast is higher than its sensitivity to color, and the perception of a region by the human eye is affected by the surrounding areas. Situations in which the results of the evaluation are inconsistent with the subjective feeling of a viewer therefore often occur. We recommend using the structural similarity index (SSIM). The learning-based SISR algorithm learns a mapping between low-resolution (LR) and high-resolution (HR) image patches. The prior knowledge used is either explicit or implicit, depending upon the learning strategy. The super-resolution convolutional neural network (SRCNN) [4] algorithm introduced deep learning methods to SISR. A faster model, the faster super-resolution convolutional neural network (FSRCNN) [6], improved upon the SRCNN model and has also been applied to SISR. The efficient subpixel convolutional neural network (ESPCN) algorithm [21] and information multi-distillation network (IMDN) [12] were also proposed to further improve the computational efficiency. A significant advance in the generation of visually pleasing results is the super-resolution generative adversarial network (SRGAN) [16]. A large number of SR methods have been presented, most of which are designed for natural images. Fewer methods have been designed for infrared images. GANs provide a powerful framework for generating plausible-looking natural images. However, they have problems with instability [11,25]. Wasserstein generative adversarial networks (WGAN) [1] was proposed as a solution to this problem. Given the issues that there are few infrared image features and that super-resolution reconstruction is difficult, the building units of the neural network and the loss functions that provide better constraints each play an important role in improving the performance of the GAN.

In this paper, we propose a novel approach for infrared image super-resolution. We revisited the key components of SRGAN and improved the model in two ways. First, we improved the network structure by introducing the heterogeneous kernel-based residual block, which has fewer parameters than previous algorithms, and it is easier to train. HetConv enables multiscale extraction of image features by combining convolutional kernels of different sizes. Second, we developed an improved loss function: the gradient cosine similarity loss function. The traditional loss function does not consider the characteristics of infrared images, and the gradient cosine similarity loss function takes the image gradient as an important feature for better-supervised training. The experimental datasets are publicly available [10], and the experimental effects can be validated.

The remainder of this paper is organized as follows. The related works are presented in Sect. 2. We describe the HetSRWGAN architecture and the gradient cosine similarity loss function in Sect. 3. A quantitative evaluation of new

datasets, as well as visual illustrations, is provided in Sect. 4. The paper concludes with a conclusion in Sect. 5.

2 Related Works

2.1 Generative Adversarial Networks

Generative adversarial networks [7] were proposed by Goodfellow, based on game theory. In a pioneering work, C. Ledig et al. [16] used SRGAN to learn the mapping from LR to HR images in an end-to-end manner, achieving performance superior to that of previous work. A low-resolution image I^{LR} is input to a generator network to generate the reconstructed image I^{SR}, while a discriminator network takes the high-resolution images I^{HR} and I^{SR} as input to determine which is the real image and which is the reconstructed image.

2.2 HetConv: Heterogeneous Kernel-Based Convolutions

The heterogeneous kernel-based convolutions algorithm was proposed by Pravendra Singh [22]. Pravendra Singh et al. presented a novel deep learning architecture in which the convolution operation uses heterogeneous kernels. Compared to standard convolution operations, the proposed HetConv reduces the number of calculations (FLOPs) and parameters while still maintaining the presentation efficiency. HetConv is especially different from the depthwise convolutional filter used to perform depthwise convolution (DWC) [3], the pointwise convolutional filter used to perform pointwise convolution (PWC) [24] and the groupwise convolutional filter used to perform groupwise convolution (GWC) [15]. In HetConv, a variable P is used to control how much of the normal convolution kernel is retained in the operation. In addition, the total reduction is R for $K \times K$ kernels. The number of calculations of HetConv is compared with that of the normal convolution, as shown in Eq. 1.

$$R_{HetConv} = \frac{1}{P} + \frac{(1 - 1/P)}{K^2} \tag{1}$$

According to the characteristics of the heterogeneous kernel-based convolutions, we used a skip connection when designing the generator network structure. The HetSRWGAN structure is shown in Fig. 1.

3 HetSRWGAN

3.1 HetSRWGAN Architecture

Our main goal was to improve the overall visual performance of SR. In this section, we describe our improved network architecture. The main difference between the GAN and WGAN [1] is that the sigmoid function and batch normalization (BN) [13] layer of the discriminator network are removed. The entire
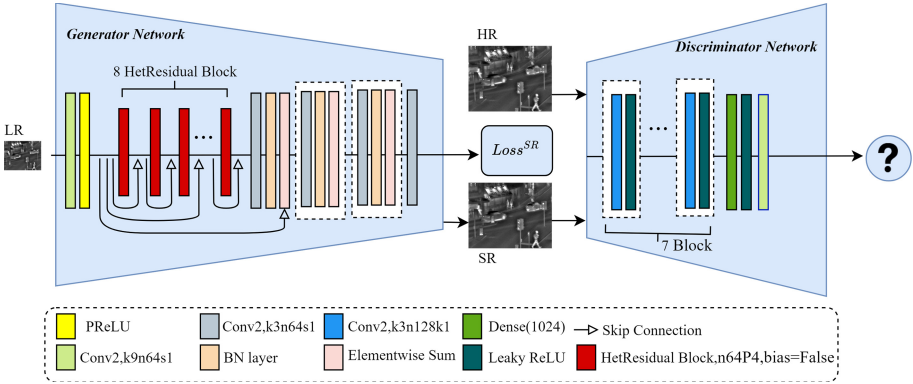
Fig. 1. Architecture of heterogeneous kernel-based super-resolution Wasserstein GAN with the corresponding kernel size (k), number of feature maps (n), stride (s) for each convolutional layer, padding (p) and number of the normal convolution kernel (P) (Best viewed in color). (Color figure online)

neural network is stabilized by gradient punishment [1]. It has been shown that removing the BN layer improves performance and reduces complexity [18,25]. Further, the removal of the BN layer contributes to improving the robustness of the network and reduces the computational complexity and memory consumption. We replaced the original basic block with a heterogeneous kernel-based residual block (HetResidual block), which includes HetConv, as depicted in Sect. 3.2. The HetResidual block is the basic network building unit. This block requires fewer parameters than the original basic block, improves network performance, and reduces computational complexity. More parameters may lead to a higher probability of mode collapse [11,25], so reducing the total number of parameters is beneficial. For the discriminator network, we deepened the network structure and experimentally demonstrated that this modification improves image quality. The detailed experimental results are given in Sect. 4. According to the characteristics of the heterogeneous kernel-based convolutions, we used a skip connection when designing the generator network structure.

3.2 Heterogeneous Kernel-Based Residual Block

Kaiming He et al. [9] first proposed the residual block structure and solved some of the problems caused by deep neural networks by introducing a skip connection and combination. The heterogeneous kernel-based residual block is shown in detail in Fig. 2. The relevant formula is analyzed as follows:

$$\mathbf{y}_i = h\left(\mathbf{x}_i\right) + \mathcal{F}\left(\mathbf{x}_i, \mathcal{W}_i\right) \tag{2}$$
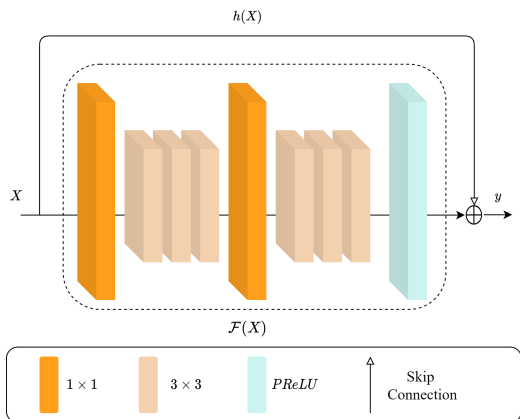
Fig. 2. Architecture of heterogeneous kernel-based residual block.

$$\mathbf{x}_{i+1} = \mathcal{F}\left(\mathbf{x}_i, \mathcal{W}_i\right) + h\left(\mathbf{x}_i\right) \tag{3}$$

where $\mathcal{F}$ stands for the heterogeneous kernel-based residual block processing. Since $h\left(\mathbf{x}_l\right)$ is an identity map, Eq. 3 can be derived:

$$\mathbf{x}_{i+1} = \mathcal{F}\left(\mathbf{x}_i, \mathcal{W}_i\right) + \mathbf{x}_i \tag{4}$$

3.3 Gradient Cosine Similarity Loss Function

To make the reconstructed image I^{SR} obtained from the generator network closer to the high-resolution image I^{HR}, it is necessary to provide a neural network loss function with effective constraints. We chose the spatial gradient of the image as the feature that measures the similarity between two images. When there is an edge in the image, there must be a high gradient value. Conversely, when there is a relatively smooth region in an image, the gray value changes little, and the corresponding gradient is also small. Using the gradient as a feature not only captures contours, images, and some texture information but also further weakens the effects of lighting. The gradient of an image at a pixel point (x, y) is a vector with direction and size. G_x is the gradient of I in direction X, and G_y is the gradient of I in direction Y direction. The gradient vector $\boldsymbol{v}$ can be expressed as Eq. 5.

$$\boldsymbol{v} = \left[\boldsymbol{G_x}, \boldsymbol{G_y}\right]^{\boldsymbol{T}} \tag{5}$$

The infrared images in the dataset are RGB images, which are three-channel images [2]. The gradient between the high-resolution three-channel image I^{HR} and the super-resolution reconstructed three-channel image I^{SR} can be expressed as Eqs. 6 and 7.

$$\mathbf{I}_G^{HR} = \left(\mathbf{I}_{G_r}^{HR}, \mathbf{I}_{G_g}^{HR}, \mathbf{I}_{G_b}^{HR}\right) \tag{6}$$

Algorithm 1. Gradient Cosine Similarity Loss Function

Input: I^{SR} , I^{HR}
Output: Gradient Cosine Similarity
1: Infrared images can be processed into RGB images [2].
2: **while** not convergent **do**
3:     $I^{HR} \longrightarrow \left(\mathbf{I}^{HR}_{G_r}, \mathbf{I}^{HR}_{G_g}, \mathbf{I}^{HR}_{G_b} \right)$
4:     $I^{SR} \longrightarrow \left(\mathbf{I}^{SR}_{G_r}, \mathbf{I}^{SR}_{G_g}, \mathbf{I}^{SR}_{G_b} \right)$ ▷ Gradient matrix.
5:     $\mathbf{X}' = \left[\mathbf{I}^{HR}_{G_r}, \mathbf{I}^{HR}_{G_g}, \mathbf{I}^{HR}_{G_b} \right]_{1 \times m}$
6:     $\mathbf{Y}' = \left[\mathbf{I}^{SR}_{G_r}, \mathbf{I}^{SR}_{G_g}, \mathbf{I}^{SR}_{G_b} \right]_{1 \times m}$ ▷ Matrix compression.
7:     $F_{\cos}\left(\mathbf{X}', \mathbf{Y}' \right) = \frac{\mathbf{X}' \cdot \mathbf{Y}'^{\mathbf{T}}}{\|\mathbf{X}'\| \cdot \|\mathbf{Y}'\|}$ ▷ Cosine similarity.
8: **return** $F_{\cos}\left(\mathbf{X}', \mathbf{Y}' \right)$

$$\mathbf{I}^{SR}_G = \left(\mathbf{I}^{HR}_{G_r}, \mathbf{I}^{HR}_{G_g}, \mathbf{I}^{HR}_{G_b} \right) \tag{7}$$

$\mathbf{I}^{HR}_G$ indicates the gradient vector of the high-resolution image. The subscript of G_g indicates the *green* channel of the high-resolution image. Other subscripts indicate different image channels of *red* and *blue*. For super-resolution reconstructed images I^{SR}, the subscript indicates the same. We use the cosine similarity to measure the similarity between these two vectors, as shown in Eq. 8.

$$\cos_{sim}(\mathbf{X}, \mathbf{Y}) = \frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\| \cdot \|\mathbf{Y}\|} \tag{8}$$

$\mathbf{X}$ and $\mathbf{Y}$ represent two matrices that can be multiplied by points. The high-resolution image gradient $\mathbf{I}^{HR}_G$ and the SR image gradient $\mathbf{I}^{SR}_G$ can be calculated according to Algorithm 1.

We calculate the cosine similarity by stretching the two matrices into a one-dimensional vector. Likewise, the similarity between the high-resolution image gradient $\mathbf{I}^{HR}_G$ and the SR image gradient $\mathbf{I}^{SR}_G$ can be calculated according to Algorithm 1. The generator loss function of the SRGAN and WGAN includes content loss and adversarial loss. The generator loss function of HetSRWGAN is shown in Eq. 9:

$$\text{Loss}^{SR} = l^{SR}_X + \lambda l^{SR}_{Gen} + \mu \left(1 - F_{cos} \right) \tag{9}$$

where l^{SR}_X and l^{SR}_{Gen} represent the content loss and adversarial loss, respectively.

Table 1. Quantitative evaluation of SR algorithms: Average PSNR/SSIM for scale factors $\times 4$. SRGAN[1] has model collaps.

Algorithm	Params ↓	FLOPs ↓	PSNR/dB ↑		SSIM ↑	
			fusionA-22	fusionC-22	fusionA-22	fusionC-22
HetSRWGAN(Ours)	0.496M	0.095G	30.302	31.987	0.858	0.883
SRMD [27]	1.552M	0.063G	33.210	33.850	0.834	0.852
IMDN [12]	0.893M	91.70G	29.725	30.057	0.735	0.751
DPSR [26]	2.995M	0.052G	32.692	31.662	0.825	0.810
DBPN [8]	10.41M	0.106G	17.438	17.934	0.816	0.842
SRWGAN	0.956M	0.132G	28.319	28.520	0.799	0.805
SRGAN[1] [16]	0.956M	0.132G	5.150	30.444	0.278	0.871
SRCNN [5]	0.148M	0.182M	29.437	30.170	0.754	0.789
FSRCNN [6]	0.013M	0.077M	30.624	31.094	0.797	0.822
ESPCN [21]	0.061M	0.001G	30.814	31.607	0.789	0.819

4 Experiments and Evaluations

4.1 Training Details

Following SRGAN, all experiments were performed with a scaling factor of (4, applied to the 2×2 image) between LR and HR images. We used the PSNR and structural similarity index (SSIM) to evaluate the reconstructed images. Super-resolved images were generated using the reference methods, including SRMD, IMDN, DPSR, DBPN, SRCNN, FSRCNN, ESPCN, SRGAN, and super-resolution Wasserstein GAN (SRWGAN). The generator was trained using the loss function presented in Eq. 9 with $\lambda = 0.001$ and $\mu = 0.001$. The learning rate was set to 0.0001. We observed that a larger batch size benefits training a deeper network. We set the batch size to 64. For optimization, we used Adam [14] with $\beta_1 = 0.9$ in the generator. For the WGAN, we used the Asynchronous SGD (ASGD) [19] in the discriminator. We implemented our models with the PyTorch framework and trained them using NVIDIA TITAN X (Pascal) GPUs.

For training, we primarily used the CVC-09: FIR Sequence Pedestrian Dataset [23]. In CVC-09, a sequence is composed of two sets of images, the day and night sets, a designation which refers to the time of day at which they were acquired. The first set contains 5990 frames, the second set contains 5081 frames, and each sequence was divided into training and testing sets. We performed experiments on two datasets, namely, fusionA-22 and fusionC-22, which contain images obtained by fusing infrared and visible light, using the methods of literature [17] and literature [28], respectively [10]. An image after the fusion of IR and visible light images will have better visual quality, and it will be easier to distinguish details such as characters in the image. The fused image also maintains significant information from the infrared image but makes the performance of the algorithm more easily visualized.
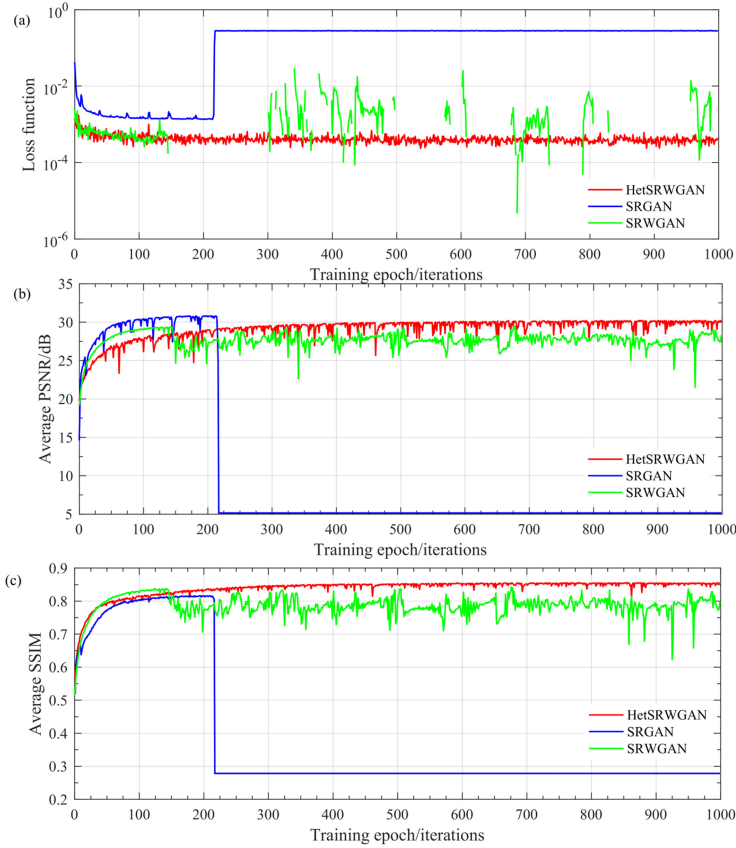
Fig. 3. (a): Changes of loss function with the number of iterations on the dataset CVC-09-1K (b): CVC-09-1K Dataset Training Average PSNR, (c): CVC-09-1K Dataset Training Average SSIM

4.2   Performance of the Final Networks

We compared the performance of three different super-resolution reconstruction algorithms based on generative adversarial networks. Since the GAN cannot simply use the loss function to judge the network training situation, we selected the image after the end of each batch of training to calculate the PSNR and SSIM values. When there are too many model parameters, mode collapse will occur. As the number of iterations increased, SRWGAN was more robust. SRGAN experiences mode collapse. Although the SRWGAN introduces gradient punishment to solve the problem that the network cannot be trained in the later stages, using cross-entropy as a loss function requires considerable time to adjust parameters and still cannot guarantee the stability of the model. Therefore, the loss function will have a negative value, which will cause the curve to be discontinuous. There
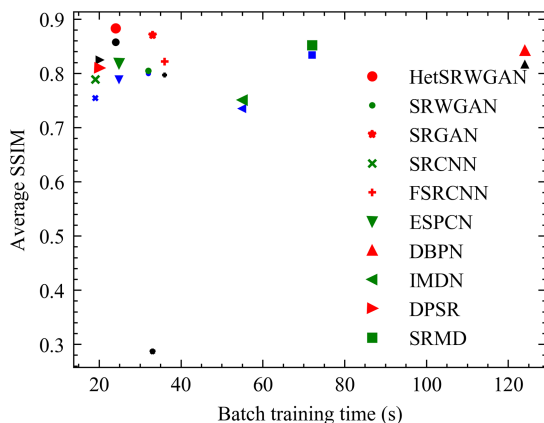
Fig. 4. Time efficiency comparison of all reconstruction methods. The same colour means the same method. The horizontal axis represents the time required for one training session, and the vertical axis represents an objective indicator after the model converges.

was no situation where convergence or instability was not possible. The results are shown in Fig. 3.

The total number of parameters for HetSRWGAN was reduced by 496657 compared to that for SRGAN, a reduction of 52% (Table 1). The significantly reduced total number of parameters helps to reduce the computational complexity of the model and improve robustness.

The SRMD model obtains better performance based on the PSNR; however, it has a large number of parameters, resulting in long training and inference times and greater memory consumption (Table 1, Fig. 4). The objective evaluation indices of the average PSNR and average SSIM were calculated. DNNs have a good effect in reconstructing visible images, but because of the features of single-frame infrared images with few features and high redundancy, the reconstruction effect is not good (Table 1).

SRGAN does not provide control of the generation process, and there is mode collapse (see Fig. 1). The new loss function and HetResidual block make the models faster to train and converge. The HetSRWGAN takes 24 s to train each batch, and the average SSIM is 0.858 and 0.883 (see Table 1). Compared with other methods, HetSRWGAN has the best time efficiency and average SSIM. Figure 5 shows the reconstructions produced by different algorithms.

Figure 5 shows that our proposed HetSRWGAN outperformed previous approaches in both sharpness and amount of detail. Previous GAN-based methods sometimes introduce artifacts. For example, SRGAN adds noise to the entire image. HetSRWGAN removes these artifacts and produces natural results.
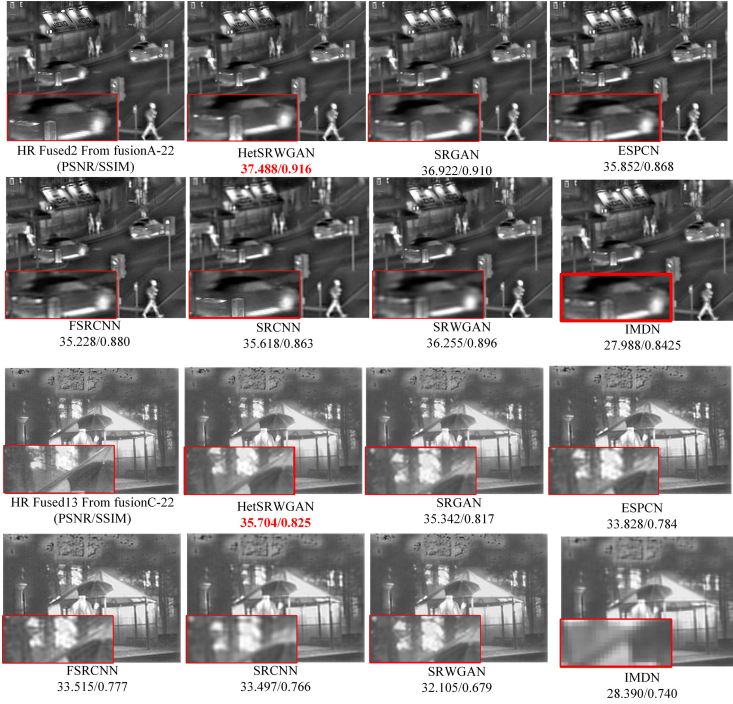
Fig. 5. Super-resolution image reconstruction effect comparison schematic diagram. From left to right: original HR image, HetSRWGAN, SRGAN, ESPCN, FSRCNN, SRCNN, SRWGAN, IMDN, Corresponding PSNR and SSIM are shown below the figure. Red indicates the best. [×4 upscaling] (Color figure online)

5   Conclusions

Our proposed HetSRWGAN method can be well used for infrared image super-resolution reconstruction. We proposed a novel architecture composed of several heterogeneous kernel-based residual blocks without BN layers. A gradient cosine similarity loss function was developed, which can provide stronger supervision of image details, such as edges, and the reconstructed high-resolution images contain more details and realistic textures.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. arXiv:1701.07875 [cs, stat], January 2017
2. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media Inc., Sebastopol (2008)

3. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)

4. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2016). https://doi.org/10.1109/TPAMI.2015.2439281

5. Dong, C., et al.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2015)

6. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 391–407. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_25

7. Goodfellow, I.J., et al.: Generative Adversarial Networks. arXiv:1406.2661 [cs, stat], June 2014

8. Haris, M., et al.: Deep back-projection networks for super-resolution. In: Proceedings of the IEEE Conference on CVPR, pp. 1664–1673 (2018)

9. He, K., Zhang, X., Ren, S., Sun, J.: Identity Mappings in Deep Residual Networks. arXiv:1603.05027 [cs], March 2016

10. Huang, Y.: Hetsrwgan-dataset, September 2019. https://figshare.com/articles/dataset/HetSRWGAN-dataset/9862184/2

11. Huang, Y., Jiang, Z., Lan, R., Zhang, S., Pi, K.: Infrared image super-resolution via transfer learning and PSRGAN. IEEE Signal Process. Lett. **28**, 982–986 (2021)

12. Hui, Z., et al.: Lightweight image super-resolution with information multi-distillation network. In: Proceedings of the 27th ACM MM, pp. 2024–2032 (2019)

13. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:1502.03167 [cs], February 2015

14. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs], December 2014

15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)

16. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE Conference on CVPR, pp. 4681–4690 (2017)

17. Liu, Y., Chen, X., Cheng, J., Peng, H., Wang, Z.: Infrared and visible image fusion with convolutional neural networks. Int. J. Wavelets Multiresolut. Inf. Process. **16**(03), 1850018 (2017)

18. Nah, S., et al.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on CVPR, pp. 3883–3891 (2017)

19. Odena, A.: Faster Asynchronous SGD. arXiv:1601.04033 [cs, stat], January 2016

20. Radford, A., et al.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)

21. Shi, W., et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on CVPR, pp. 1874–1883 (2016)

22. Singh, P., Verma, V.K., Rai, P., Namboodiri, V.P.: HetConv: Heterogeneous Kernel-Based Convolutions for Deep CNNs. arXiv:1903.04120 [cs], March 2019

23. Socarrás, Y., Ramos, S., Vázquez, D., López, A.M., Gevers, T.: Adapting pedestrian detection from synthetic to far infrared images. In: ICCV Workshops, vol. 3 (2013)
24. Szegedy, C., et al.: Going Deeper With Convolutions, pp. 1–9 (2015)
25. Wang, X., et al.: ESRGAN: enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018)
26. Zhang, K., Zuo, W., Zhang, L.: Deep plug-and-play super-resolution for arbitrary blur kernels. In: IEEE Conference on CVPR, pp. 1671–1681 (2019)
27. Zhang, K., et al.: Learning a single convolutional super-resolution network for multiple degradations. In: Proceedings of the IEEE Conference on CVPR, pp. 3262–3271 (2018)
28. Zhang, Y., Zhang, L., Bai, X., Zhang, L.: Infrared and visual image fusion through infrared feature extraction and visual information preservation. Infrared Phys. Technol. **83**, 227–237 (2017)