Super-resolution reconstruction of infrared images based on dual attention guidance

Yunhao Shu[a], Guanying Zhang[a] , Huan Chen[b*], Wenming Zhu[a], Jianxun Ma[a]
[a]State Grid Changzhou Power Supply Company, Changzhou, Jiangsu, 213000, China; [b]College of Information Science and Engineering, Hohai University, Changzhou, Jiangsu, 213000, China

ABSTRACT

To address the issues of low resolution and blurry edges in infrared images of substation scenes, a super-resolution reconstruction method based on a dual-attention guidance mechanism is proposed. Specifically, during the deep feature extraction of infrared images, the spatial and channel transformer group (SCTG) is proposed to extract global spatial similarity features, fully utilizing the long-range dependencies of non-local pixels and enhancing detailed information. Subsequently, a Spatial Frequency Information Fusion Module (SFIFM) utilizes the extracted high-frequency information, reducing artifacts and mosaic effects that occur during the super-resolution process. The overall quality of the reconstructed images is improved and the detailed contour information is refined. Finally, ablation and comparative experiments on a self-made dataset demonstrate that the proposed method outperforms state-of-the-art methods.

Keywords: Super-resolution reconstruction, dual-attention guidance mechanism, spatial frequency, substation.

1. INTRODUCTION

Infrared imaging technology[1] can convert invisible thermal radiation signals into images forms and achieve non-touchable temperature measurement. Considering the strong information expression capability and the long effective range, high interference resistance in the images Despite the rapid advancements in infrared imaging technology, infrared images still suffer from low spatial resolution, lack of detailed textures, and blurred targets due to the physical limitations of sensors.

Super-resolution reconstruction[2] is a significant field in image processing, aimed at converting low-resolution inputs into high-resolution images and enhancing image details. With the continuous advancement of deep learning, super-resolution technology has also been evolving. However, challenges still remain in reconstructing and enhancing the details when applied for infrared images. To address this issue, numerous studies have integrated self-attention mechanisms[3] to capture global information. SASRGAN[4] and AFiLM[5] were proposed to utilize self-attention mechanisms to strengthen spatial dependencies, thereby improving the structural quality of super-resolution. Nevertheless, many models often overlook multi-scale detail extraction, resulting in incomplete global information capture. The SCTANet[6] model incorporated a mixed complementary spatial attention mechanism during the feature extraction process, aiding pixel-level detail reconstruction and reducing computational complexity, but neglects depth and high-frequency information. In contrast, the MAGSR[7] employed a multi-scale hybrid attention mechanism to extract richer multi-scale depth and high-frequency details, resulting in clearer reconstructed infrared images. Despite these advancements, the global features are still ignored, which may restrict the potential for detailed information reconstruction. By combining spatial self-attention mechanisms along channel and spatial frequency dimensions, global spatial similarity features and long-range dependencies of infrared images can be fully leveraged. Thus, in this way, detail information in infrared images could be reconstructed, thereby improving the quality of the reconstructed images.

The proposed study introduces a super-resolution reconstruction network for infrared images based on dual attention guidance. The main research content of the study is as follows:

1) A Spatial and Channel Transformer Group (SCTG) is constructed to capture global information in both spatial and channel dimensions. The design enables the model to identify and enhance important spatial features while effectively refining and utilizing inter-channel global information.

* 231323030008@hhu.edu.cn

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 133950A · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3049860

2) A Spatial Frequency Domain Information Fusion Module (SFIFM) based on Fast Fourier Transform (FFT)[8] is proposed to extract high-frequency information from infrared images. The receptive field is extended in the frequency domain to enhance high-frequency details, reduce artifacts and obtain clearer contour edges.

c) Through the combination of the developed SCTG and SFIFM modules, a rational network framework is constructed, which can improve the resolution through the processing of infrared images, and the experiments in the dataset prove the advancement of constructing a network.

2. PROPOSED MODEL

The structure of the super-resolution reconstruction network for infrared images based on dual-attention guidance is illustrated in Figure 1. The network consists of three main components: a Shallow Feature Extraction Module (SFEM), a Deep Feature Extraction Module (DFEM), and a High-Quality Reconstruction Module (HQRM). The SFEM employs a 3×3 convolutional layer to extract shallow features from the input low-quality infrared image. The DFEM comprises five Dual Attention Guided Transformer Blocks (DATB) and one Spatial Frequency Domain Information Fusion Module (SFIFM). This module enhances the deeper features of the image, compensating for the loss of high-frequency detail information during the infrared image reconstruction process. The HQRM includes a Pixel Shuffle[9] and two 3×3 convolution blocks. By integrating the extracted shallow features with the deeper features using residual connections, the HQRM effectively fuses information from both feature levels, resulting in a high-quality infrared image after the final super-resolution reconstruction.

To address the issue of insufficient detail information provided by the self-attention mechanism, the study incorporates a global spatial and channel self-attention mechanism module. The module simultaneously applies spatial and channel self-attention mechanisms during the extraction of deep features, effectively refining and utilizing global information between channels to achieve more precise and efficient super-resolution reconstruction. Additionally, a spatial frequency domain information fusion module is integrated to further enhance network performance. The fusion module improves high-frequency details of infrared images, resulting in superior quality of reconstructed image contours and edges. The network thus enhances both the resolution and clarity of infrared images, providing high-quality data for subsequent bird target detection technology.

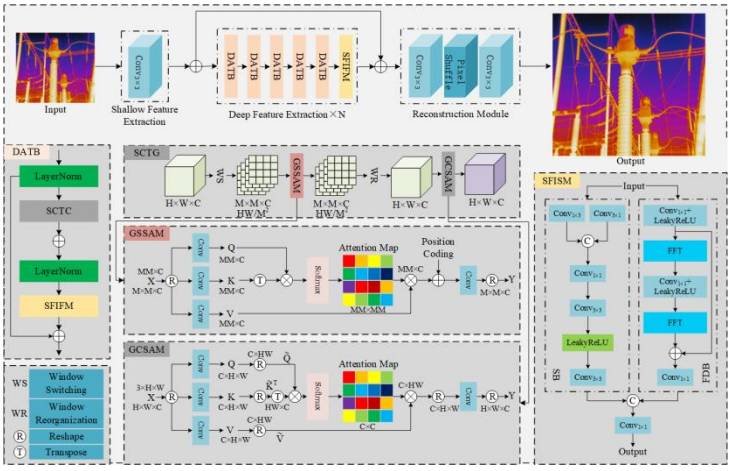

Figure 1. The structure of the super-resolution reconstruction network for infrared images based on dual-attention guidance.

2.1 Shallow feature extraction module (SFEM)

The main task of SFEM is to extract initial shallow features on the input low-quality infrared image. The shallow features $F_0 \in \mathbb{R}^{H\times W\times C}$ are extracted from the input image $I_{LQ}^{H\times W\times C}$, where I represents the infrared image, LQ represents the low-quality image, H and W represent the height and width, and C is the number of channels, set to 3 in this study. The extraction process follows the formula:

$$F_0 = Conv_{3\times 3}\left(I_{LQ}\right) \tag{1}$$

Where $Conv_{3\times3}(\cdot)$ denotes the 3×3 convolution.

2.2 Deep feature extraction module (DFEM)

The DFEM consists of five DATB and one SFIFM. Each DATB includes a SCTG with residual connections and one SFIFM. By integrating the SFIFM after each SCTG, the model's ability to capture high-frequency details and reconstruct image content is significantly enhanced, improving the model's understanding of image details. The overall process of DATB can be represented as:

$$X = SCTG\left(LayerNorm\left(X\right)\right) + X \tag{2}$$

$$X = SFIFM\left(LayerNorm\left(X\right)\right) + X \tag{3}$$

Where X is the input image feature and $LayerNorm$ is the normalization layer. By introducing the SFIFM after each layer of DATB, the model's ability to capture high-frequency and reconstruct image details are significantly enhanced. This structural design enables the network to effectively reconstruct image content at a deeper level, providing, robust support for high-quality image reconstruction and feature extraction.

2.1.1 Spatial and channel transformer group (SCTG). Existing Transformer-based methods focus only on extracting spatial features and ignore the global information in the channel dimension. To alleviate this problem, the Spatial and Channel Transformer Group (SCTG) is constructed to capture global information in both the spatial and channel dimensions, as shown in Figure 1.

SCTG is mainly composed of GSSAM and GCSAM. After the layer normalization operation, the input feature $F \in R^{H\times W\times C}$ is obtained, and a segmentation operation with a window size of M is performed to segment the feature F in the spatial dimension. The input feature is segmented into n chunks $\{F_1, F_2, ..., F_n\}$ of size HW/M^2 and non-overlapping, where $F_i \in R^{M\times M\times C}$, and each chunk F_i is fed into the GSSAM to obtain global similarity information in the spatial dimension. The global spatial feature output $F_{GSSAM} \in \mathbb{R}^{H\times W\times C}$, is then reorganized into windows and fed into GCSAM to capture global information in the channel dimension. The whole process can be defined as:

$$\{F_i\} = WindowPartition\left(F\right), i = 1, ...n \tag{4}$$

$$F_{GSSAM} = WindowReverse\left(GSSAM\left(F_i\right)\right), i = 1, ..., n \tag{5}$$

$$F_{GCSAM} = GCSAM\left(F_{GSSAM}\right) \tag{6}$$

The main processes of GSSAM and GCSAM are as follows:

a) Global Spatial Self-Attention Module (GSSAM)

As shown in Fig. 1, each non-overlapping chunk after the window partitioning operation is first reshaped, followed by three 1×1 convolutions to obtain the query matrix Q, the key matrix K, and the value matrix V, where $\left(Q、K、V\right) \in \mathbb{R}^{M^2\times C}$ respectively. The global spatial information matrix within a localized window can be computed by the self-attention mechanism:

$$Attention\left(Q, K, V\right) = Softmax\left(QK^T/\sqrt{d} + P\right)V \tag{7}$$

where P is the positional encoding that splits the attention into N heads to learn the separate attention matrices, $d = \frac{C}{N}$, in a parallel fashion. The results then undergo a 1×1 convolution and reshaping operation to obtain the output features of the $F_i^{GSSAM} = GSSAM\left(F_i\right), i = 1, ..., n$.

b) Global Channel Self-Attention Module (GCSAM)

After GSSAM extracts the global spatial features, GCSAM further capture more global features in the channel dimension by weighting the extracted spatial features. As shown in Fig. 1, given the input features $F_{GSSAM} \in \mathbb{R}^{H\times W\times C}$ from GSSAM, a reshaping and linear projection operation is first performed on the query matrix Q, the key matrix K, and the value matrix V, where $(Q、K、V)\in\mathbb{R}^{M^2\times C}$. This is followed by a reshaping operation on Q and V to obtain $(\tilde{Q},\tilde{V})\in\mathbb{R}^{C\times HW}$. A reshaping and transposition operation is then performed on the key matrix K to obtain $(\tilde{K})^T\in\mathbb{R}^{HW\times C}$, and the global channel attention matrix is computed as follows:

$$Attention(\tilde{Q},\tilde{K},\tilde{V}) = Sofmax\left(\tilde{Q}(\tilde{K})^T/\alpha\right)\tilde{V} \tag{8}$$

where α is a learnable scaling parameter. Linear projection and reshaping operations are then performed to obtain $F_{GCSAM}\in\mathbb{R}^{H\times W\times C}$.

2.1.2 Spatial and frequency information fusion module (SFISM). As shown in Figure 1, the Spatial Frequency Domain Information Fusion Module (SFIFM) consists of two main branches: the frequency domain branch and the spatial branch. The frequency domain branch processes the FFT-transformed image to capture global information, while the spatial branch processes image features in the original spatial domain to maintain sensitivity to local details. Ccombining these two branches allows the module to comprehensively utilize both global and local information for a more thorough feature representation. The input feature T is fed to both branches to obtain the frequency domain feature $T_{frequency}$ and the spatial feature $T_{spatial}$, respectively. In the frequency domain branch, the input features are first processed using a 1×1 convolutional layer $Conv_{1\times1}$ and a $LeakyReLU$ activation function $L(\cdot)$,

$$T^1_{frequency} = L\left(Conv_{1\times1}(T)\right) \tag{9}$$

Then the frequency-domain features $T^1_{frequency}$ are converted to spatial features $T^2_{frequency}$ using the fast Fourier transform $FFT(\cdot)$, a 1×1 convolutional layer, an $LeakyReLU$ activation function and the fast Fourier inverse transform $InvFFT(\cdot)$,

$$T^2_{frequency} = InvFFT(L(Conv_{1\times1}(FFT(T^1_{frequency})))) \tag{10}$$

Finally, $T^1_{frequency}$ and $T^2_{frequency}$ are residually concatenated and fed into a 1×1 convolutional layer, which adjusts the number of channels to obtain the final features $T^{final}_{frequency}$:

$$T^{final}_{frequency} = Conv_{1\times1}(T^1_{frequency} + T^2_{frequency}) \tag{11}$$

In the spatial branch, the input features T are fed to a 1×3 convolutional layer and a 3×1 convolutional layer, respectively. The outputs of these convolutional layers are then concatenated and further processed by a 1×1 convolutional layer:

$$T^1_{spaial} = Conv_{1\times1}\left(Cat\left(Conv_{1\times3}(T), Conv_{3\times1}(T)\right)\right) \tag{12}$$

where $Cat(\cdot)$ denotes a concatenation operation(T^1_{spaial} and $T^{final}_{spatial}$ relationship). The features obtained from the frequency domain branch $T^{final}_{frequency}$ and the spatial branch $T^{final}_{spatial}$ are concatenated, and then fed into a 1×1 convolutional layer to adjust the number of channels, resulting in the final spatial-frequency domain fusion features T_{fusion}:

$$T_{fusion} = Conv_{1\times1}\left(Cat\left(T^{final}_{frequency}, T^{final}_{spatial}\right)\right) \tag{13}$$

2.3 High quality reconstruction module (HQRM)

In the final image reconstruction stage, the High Quality Reconstruction Module (HQRM) integrates the extracted shallow and deep feature information using residual connections. This fusion ensures that the shallow detail awareness is preserved and the deep semantic information is fully utilized. HQRM consists of a Pixel Shuffle and two 3×3 convolutional blocks. The Pixel Shuffle rearranges the elements of the input feature to increase the spatial resolution while maintaining the image's integrity and continuity. This process produces the final high-quality, super-resolution infrared image.

2.4 Loss function

In this study, the loss function is *Charbonnier* loss. The model is optimized based on the high-quality image I_{SHQ} and the corresponding ground truth image I_{HQ} generated by the super-resolution reconstruction model. The loss function calculation formula is shown below:

$$L(\eta) = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(M(I_{SHQ}^{i},\eta) - I_{HQ}^{i})^{2} + \delta} \tag{14}$$

where M denotes the model proposed in this study, N denotes the number of image pairs in the training dataset, η denotes the model parameters, and δ is a constant.

3. EXPERIMENTAL DETAILS

3.1 Dataset and experimental setup

Super-resolution reconstruction experiments were conducted using a self-made infrared image dataset of substation scenes. The proposed network is built on the PyTorch deep learning framework. An adaptive optimizer (Adaptive Moment Estimation, Adam) was used for model optimization, with an initial learning rate set to 0.0001, and the MultiStepLR scheduler was employed to dynamically adjust the learning rate during training. The original resolution of the infrared images is 640×512, and the experimental magnification factors were chosen to be 2 and 3.

3.2 Evaluation criteria

Two commonly used super-resolution reconstruction metrics including Peak Signal-to-Noise Ratio (PSNR)[10] and Structural Similarity (SSIM)[11] are adopted to evaluate the performance of different algorithms.

3.3 Ablation experiment

To evaluate the impact of the Spatial and Channel Transformer Group (SCTG) and the Spatial and Frequency Information Fusion Module (SFIFM) on the performance of infrared image super-resolution reconstruction algorithms, ablation experiments were conducted on the self-made infrared image dataset. The experiments focused on 2×and 3× super-resolution cases.

Table 1. Quantitative results of ablation experiments.

SCTG	SFIFM	2×		3×	
		PSNR(db)	SSIM(%)	PSNR(db)	SSIM(%)
		39.9426	0.9512	34.4231	0.9198
	√	40.1188	0.9587	34.7219	0.9213
SCTG	SFIFM	2×		3×	
		PSNR(db)	SSIM(%)	PSNR(db)	SSIM(%)
√		40.3251	0.9688	34.9854	0.9299
√	√	40.4961	0.9703	35.5011	0.9368

Table 1 shows that using only SCTG slightly reduces network performance compared to using both SCTG and SFIFM. For 2× magnification, PSNR and SSIM drop by 0.171 dB and 0.15%, respectively. When only SFIFM is used, performance decreases more, with PSNR and SSIM dropping by 0.3773 dB and 1.16%. Without both SCTG and SFIFM, performance drops significantly, with reductions of 0.5535 dB in PSNR and 1.91% in SSIM. Similar trends are observed for 3× magnification, highlighting the importance of SCTG for enhancing global detail capture in infrared images and its effective collaboration with SFIFM to maintain high image quality.

3.4 Comparison experiment

Table 2 presents the results of the quantitative experiments on the self-made infrared image dataset, with bold font indicates the best metrics and underlined font indicates the second best metrics.

Table 2. Quantitative results of comparative experiments.

Methodologies	**2×**		**3×**	
	PSNR(db)	**SSIM(%)**	**PSNR(db)**	**SSIM(%)**
BICUBIC	34.23	94.60	30.64	90.59
EDSR	35.7442	94.84	31.0493	91.14
RCAN	36.9504	95.24	32.4436	91.87
SAN	38.3923	95.83	33.6378	92.03
SwinIR	39.9426	96.12	34.4331	92.29
CAT	40.3706	97.17	34.6257	93.76
OURS	40.5678	97.83	35.5211	94.13

The experimental data show that at 2× magnification, the proposed method achieved the highest PSNR and SSIM, with improvements of 0.1972 dB and 0.66% over the second-best algorithm, respectively. These results demonstrate the effectiveness of the proposed super-resolution algorithm for infrared images. At 3× magnification, although all methods showed a decrease in metrics, the proposed method still led in both indicators, demonstrating its ability at high magnification.

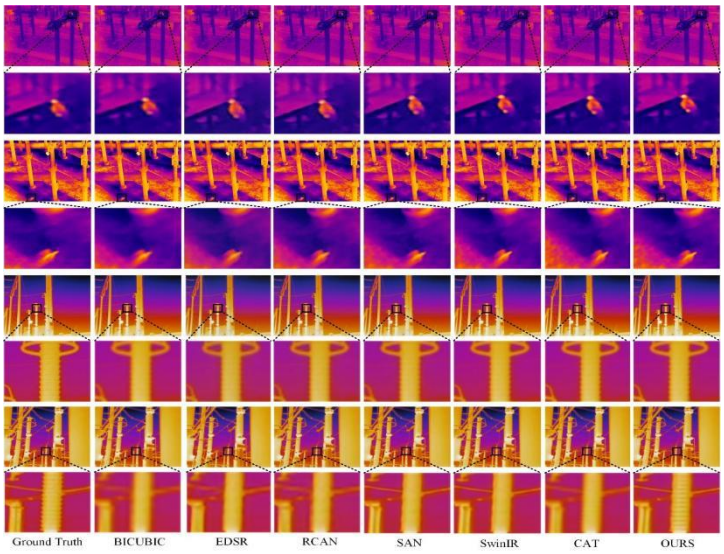

Figure 2. The visualization comparison result of birds and insulators in substation scenes.

Figure 2 shows the visualization comparison result of birds and insulators in substation scenes. For both 2× or 3× magnification, the BICUBIC[12], EDSR[13], and RCAN[14] algorithms produced generally mediocre results, with magnified images appearing blurry and lacking visual clarity. At 2× magnification, SAN[15] and SwinIR[16] achieved some improvements, but the contour edges remain blurry and jagged. In contrast, both CAT[17] and the proposed algorithm clearly magnified the texture of the stacked shapes and maintain the hierarchical detail and clarity of the edge contours. However, at 3× magnification, CAT failed to display the stacked shapes and jagged edges of the insulator clearly, while the proposed algorithm reconstructed more texture details, demonstrating good super-resolution reconstruction capability and effectively preserving the texture details and edge contours.

4. CONCLUSION

To address the issues of low resolution and blurred target edges in infrared images of power equipment from existing infrared sensors, a super-resolution reconstruction network based on a dual attention-guided mechanism is proposed. The Spatial and Channel Transformer Group (SCTG) captures long-range dependencies, accurately focusing on crucial details such as target shapes, thereby significantly improving super-resolution reconstruction. Additionally, a spatial frequency domain information fusion module (SFISM) based on fast Fourier transform (FFT) is introduced, utilizing FFT's powerful high-frequency detail extraction to further enhance the detail expressiveness in infrared images. By emphasizing high-frequency details and edge information in the frequency domain, these residuals are integrated into the SCTG module, ensuring that the super-resolution process preserves the main structure while refining textures and edges. Extensive experiments demonstrate that this method surpasses existing state-of-the-art techniques, providing high-quality infrared image data for subsequent tasks such as substation bird target detection.

ACKNOWLEDGEMENT

This research was funded by the Incubation Project of State Grid Jiangsu Electric Power Co., Ltd., grant number JF2023012.

REFERENCE

[1] Kastberger, G. and Stachl, R., “Infrared imaging technology and biological applications,” Behavior Research Methods, Instruments, & Computers, 35, 429-439 (2003).
[2] Elad, M,, Feuer, A., “Super-resolution reconstruction of image sequences,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(9), 817-834 (1999).
[3] Edelman, B, L., Goel, S., Kakade, S., “Inductive biases and variable creation in self-attention mechanisms,” International Conference on Machine Learning. PMLR, 1, 5793-5831 (2022).
[4] Zong, L., Chen, L., “Single image super-resolution based on self-attention,” 2019 IEEE International Conference on Unmanned Systems and Artificial Intelligence (ICUSAI), IEEE, 56-60 (2019).
[5] Rakotonirina, N, C., “Self-attention for audio super-resolution,” 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP), 1-6(2021).
[6] Bao, Q., Liu, Y., Gang, B., “SCTANet: A spatial attention-guided CNN-transformer aggregation network for deep face image super-resolution,” IEEE Transactions on Multimedia, 25, 8554-8565 (2023).
[7] Liang, G., KinTak, U., Yin, H., “ Multi-scale hybrid attention graph convolution neural network for remote sensing images super-resolution,” Signal Processing, 207, 108954 (2023).
[8] Duhamel, P., Vetterli, M., “Fast Fourier transforms: a tutorial review and a state of the art,” Signal processing, , 19(4), 259-299(1990).
[9] Prasad M, Sudha K L. “Chaos image encryption using pixel shuffling,” CCSEA, 1, 169-179(2011).
[10] Rahman, A., Detection of Deepfake using computer vision and deep learning, Brac University, 2023.
[11] Wang, Z., Bovik, A, C. and Sheikh, H, R., “Image quality assessment: from error visibility to structural similarity,” IEEE transactions on image processing, , 13(4), 600-612(2004).
[12] Xiang. R., Yang, H. and Yan, Z., “Super-resolution reconstruction of GOSAT CO2 products using bicubic interpolation,” Geocarto International, 37(27), 15187-15211(2022).
[13] Jenefa, A, M. and Naveen, E, V., “EDSR: Empowering super-resolution algorithms with high-quality DIV2K images,” Intelligent Decision Technologies, 17(4), 1249-1263 (2023).

[14] Lin, Z., Garg, P. and Banerjee, A., "Revisiting rcan: Improved training for image super-resolution," arXiv preprint arXiv, 11279, 2201 (2022).
[15] Dai, T., Cai, J. and Zhang, Y., "Second-order attention network for single image super-resolution," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, IEEE, 11065-11074 (2019).
[16] Chen, K., Li, L., and Liu, H., "Swinfsr: Stereo image super-resolution using swinir and frequency domain knowledge," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 1764-1774 (2023).
[17] Buzzelli, M., Tchobanou, M, K., Schettini, R., "RGB illuminant compensation using spectral super-resolution and weighted spectral color correction," Color and Imaging Conference. Society for Imaging Science and Technology, 31, 33-37 (2023).