**Computer Life**
ISSN: 1819-4818 | Vol. 13, No. 2, 2025

# Image Super-Resolution Reconstruction Method Based on Lightweight CNN-Transformer

**Wenqiang Xi [1, *], Zairila Juria Zainal Abidin[2], Cheng Peng [3], Tadiwa Elisha Nyamasvisva [4]**

[1] Center for Postgraduate Studies, Infrastructure University Kuala Lumpur, 43000, Malaysia
[2] Faculty of Architecture and Built Environment, Infrastructure University Kuala Lumpur, 43000, Malaysia
[3] Faculty of Physics and Electrical Engineering, Weinan Normal University, 714000, China
[4] Department of Computing, Faculty of Engineering Science and Technology, Infrastructure University Kuala Lumpur, 43000, Malaysia

[*] **Corresponding author**: Wenqiang Xi (Email: 213923103@s.iukl.edu.my)

**Abstract:** Existing Transformer-based image super-resolution reconstruction methods suffer from excessive parameters and high training costs. To address these issues, we propose a lightweight CNN-Transformer-based image super-resolution reconstruction method. A CNN-Transformer module is designed using weight sharing, and a channel attention module is used to fully fuse image information, which improves the reconstruction of local and global features of the image. Meanwhile, depth-wise separable convolutions are used and the covariance matrix of cross-channel self-attention is calculated, which effectively reduces the number of parameters in the Transformer and lowers the computational cost. Then, a High-Frequency Residual Block (HFRB) is introduced to further focus on the texture and detail information in the high-frequency range. Finally, the choice of activation function required for Transformer to generate self-attention is discussed. Analysis shows that the GELU activation function can better promote feature aggregation and improve network performance. Experiments show that the method in this paper can effectively reconstruct more texture and details of the image while maintaining lightweight.

**Keywords:** Image Super-Resolution Reconstruction; Deep Learning; Transformer.

## 1. Introduction

Image super-resolution (SR) reconstruction, which aims to reconstruct high-resolution (HR) images from single or multiple low-resolution (LR) images, is one of the hottest research topics in computer vision. It has been widely applied in various fields such as surveillance [1], medical diagnosis [2], and remote sensing [3]. Due to the ill-posed nature of image super-resolution reconstruction and the increasing severity of artifacts, edge blurring, and pixel loss with the increase of the upscaling factor, image super-resolution reconstruction remains a challenging problem.

With the breakthrough advancements of deep learning in computer vision, researchers have introduced it into the field of image super-resolution reconstruction, achieving significantly better visual results compared to traditional interpolation or reconstruction-based methods. To further improve reconstruction performance, researchers have proposed numerous deeper or wider network models. Inspired by the residual learning idea, Kim et al. designed VDSR [4], expanding a 3-layer convolutional network to 20 layers. Lai et al. proposed LapSRN (Laplacian Pyramid Super-Resolution Network) using a progressive reconstruction strategy. Zhang et al. proposed RCAN (Residual Channel Attention Networks) by utilizing residual connections and channel attention [5]. Zhang et al. also proposed RDN (Residual Dense Network) by combining residual connections and dense connections. Mei et al. designed NISA (Non-local Sparse Attention) using dynamic sparse attention and non-local sparse attention [6]. Wu et al. proposed MSNLAN (Multi-scale Non-local Attention Network) [7] by combining multi-scale ideas and non-local attention mechanisms.

Although increasing the depth or width of the network can improve reconstruction performance, it also leads to an increase in the number of parameters and higher memory consumption. Consequently, researchers have begun to explore reconstruction methods that reduce network size. Among these, introducing a recursive mechanism is one strategy for achieving lightweight networks. Kim et al. adopted the recursive idea and proposed DRCN (Deeply-Recursive Convolutional Network) [8]. Based on this, Tai et al. incorporated the residual learning concept and proposed DRRN (Deep Recursive Residual Network) [9].

Recursive structures can reduce the number of network weights to some extent, they still fail to significantly decrease the computational cost of the network. Therefore, balancing the number of network weights and reconstruction performance, and constructing lightweight networks have become the mainstream of current deep learning research. Hui et al., based on the idea of information distillation, proposed IMDN (Lightweight Information Multi-Distillation Network) [10]. Zha et al., utilizing dense connections and attention mechanisms, proposed LDCASR (Lightweight Dense Connected Approach with Attention for Single Image Super-Resolution) [11]. Furthermore, Lan et al., combining the idea of multi-scale processing, proposed MADNet[12]. Peng et al., using skip residual connections and channel attention, designed LCRCA (Lightweight Skip Concatenated Residual Channel Attention Network)[13]. Feng et al. proposed a lightweight image super-resolution network based on regional complementary attention and multi-dimensional attention (RCA-MDA) [14]. Gao et al., utilizing lightweight residual blocks and convolutional blocks, proposed VLESR (Very Lightweight and Efficient Image Super-Resolution Network) [15].

With the successful application of Transformers [2] in natural language processing, researchers have introduced them into the field of computer vision and achieved breakthrough progress. Unlike traditional CNNs that mainly

extract local features, the core idea of Transformers is to extract global feature information through self-attention mechanisms. In particular, in the field of image super-resolution reconstruction, researchers have combined CNNs and Transformers to extract richer feature information. Wang et al. introduced multi-scale processing into Transformers and proposed MSTN (Multi-scale Multi-stage Single Image Super-Resolution Reconstruction Algorithm Based on Transformer) [16]. Lu et al. fused lightweight CNNs and Transformers and proposed ESRT (Efficient Super-Resolution Transformer) [17]. Fang et al. introduced an enhanced spatial attention mechanism and proposed HNCT (Hybrid Network of CNN and Transformer) [18]. Li et al., based on Restormer, proposed DLGSANet (Effective Lightweight Dynamic Local and Global Self-Attention Network) [19]. The aforementioned networks, combining CNNs and Transformers, have achieved significant improvements in reconstruction performance and visual effects.

The reconstruction methods based on CNNs and Transformers mainly adopt a single-branch network structure, which has problems such as insufficient utilization and fusion of extracted information. Moreover, during the process of extracting global information, Transformers tend to ignore the texture details in high-frequency regions. Therefore, this paper proposes an Image Super-Resolution Reconstruction Method Based on Lightweight CNN-Transformer (LCT).

# 2. Image super-resolution reconstruction method based on lightweight CNN-Transformer

The Image Super-Resolution Reconstruction Method Based on Lightweight CNN-Transformer (LCT) has a structure as shown in Figure 1. LCT includes a shallow feature extraction module, a CNN-Transformer module, a Deep Feature Fusion (DFF) module, and an upsampling reconstruction module.

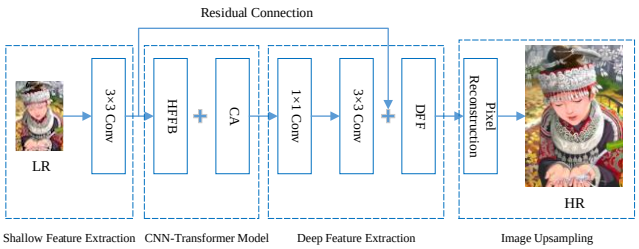


**Figure 1.** Structure of LSCT

First, given the input low-resolution image $I_{LR}$, shallow features $F_0$ are extracted through a 3 × 3 convolution. Then, $F_0$ is passed to the CNN-Transformer module for deep feature extraction, which yields deep features: $F_n = F_{\text{CNN-Transformer}}(F_0)$. where $F_{\text{CNN-Transformer}}(\cdot)$ represents the CNN-Transformer module. Then, $F_n$ is passed to the DFF module to smooth, refine and densely fuse the deep features to obtain densely fused features: $F_{DFF} = Conv_{3\times3} * (Conv_{1\times1} * F_n) + F_0$. Finally, $F_{DFF}$ is input into the upsampling reconstruction module to obtain the reconstructed image: $I_{SR} = F_{UP}(Conv_{3\times3} * F_{DFF})$.

Where $F_{UP}$ is the sub-pixel convolution.

## 2.1. CNN-Transformer Module

The CNN-Transformer module consists of a Hybrid Feature Fusion Block (HFFB)[20] and a Channel Attention Block (CA)[21]. To reduce the number of computational parameters, the two HFFBs share weight values. The specific structure of the CA block is illustrated in Figure 2.

Specifically, the output feature $F_n$ of the CNN-Transformer module is the concatenation of the feature outputs of each HFFB in the upper and lower branches. The output feature after the i-th HFFB in the upper branch is: $F_{\text{HFFB}}^{U,i} = f_{\text{HFFB}}^{U,i}(F_{\text{HFFB}}^{U,i-1}), i = 2,3,\cdots,n$.

Where $f_{\text{HFFB}}^{U,i}(\cdot)$ represents the operation of the i-th HFFB in the upper branch. The output feature after the i-th HFFB in the lower branch is:

$$F_{\text{HFFB}}^{D,i} = f_{\text{HFFB}}^{D,i}(F_{\text{HFFB}}^{D,i-1} + f_{\text{CA}}^{i}(F_{\text{HFFB}}^{U,i})), i = 2,3,\cdots,n.$$

where $f_{\text{HFFB}}^{D,i}(\cdot)$ represents the operation of the i-th HFFB in the lower branch, and $f_{\text{CA}}^{i}(\cdot)$ represents the operation of the i-th CA.

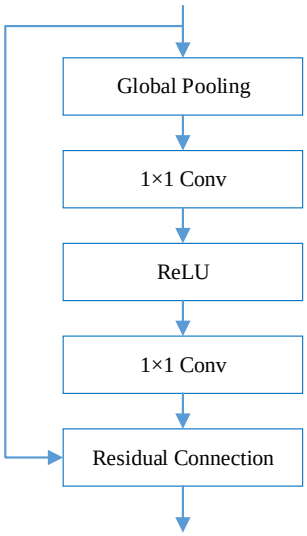


**Figure 2.** Structure of CA.

he Channel Attention (CA) block first applies global average pooling to compute the mean value of all elements within each channel. Then, 1×1 convolution are used to reduce the dimensionality of the channel feature map. After being activated by the ReLU function, the weights of channel attention are obtained by the second 1×1 convolution and the Sigmoid function. Finally, the weights are multiplied by the input features to obtain the weighted feature map. To reduce the number of weights in network training and maintain the lightweight nature of the network, the i-th HFFB weights of the upper and lower branches are shared, expressed as:

$$f_{\text{HFFB}}^{U,i}(\cdot) = f_{\text{HFFB}}^{D,i}(\cdot), i = 1,2,\cdots,n.$$

## 2.2. Hybrid Feature Fusion Module (HFBB)

The Hybrid Feature Fusion Block (HFFB) integrates the ideas of attention mechanism, dense connection, Transformer, and network lightweighting. It designs multiple densely connected Attention Feature Blocks (AFB), lightweight Transformer modules, and HFRBs[22] to achieve the

extraction of local and global feature information, as well as the enhancement of information in the high-frequency region. The specific structure of the HFFB is shown in Figure 3.

To enhance the capability of local feature extraction by integrating channel attention and enhanced spatial attention, Attention Feature Block (AFB) is designed. To maintain the lightweight nature of the network model, a 1 × 1 group convolution is introduced before the AFB to halve the number of channels, thereby reducing the number of parameters. Furthermore, to achieve deep fusion of multi-level feature information, the features extracted by all AFB are densely connected.

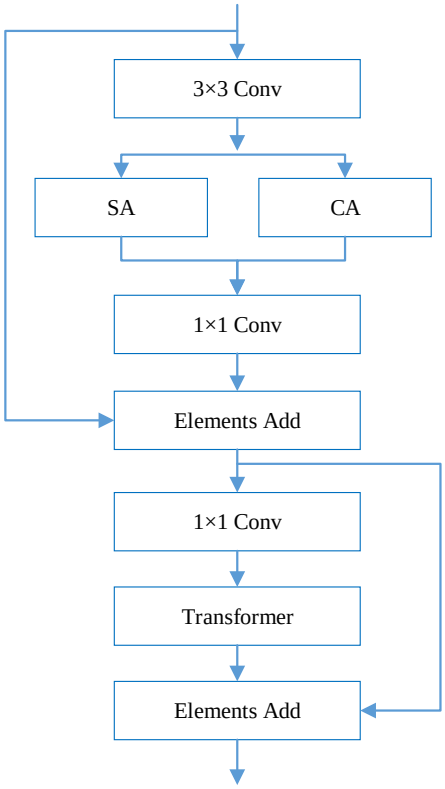


**Figure 3.** Structure of HFFB

Furthermore, to capture richer global feature information while maintaining a lightweight design, LCT employs a lightweight Transformer architecture based on depthwise separable convolutions. Additionally, to enhance compensation for high-frequency region details, a High-Frequency Residual Block (HFRB) is introduced in the residual connections.

# 3. Lightweight Transformer Module

Unlike CNN architectures, Transformers leverage the self-similarity properties of images and employ self-attention mechanisms to capture global information. However, due to the inner product operations, Transformers face challenges related to high memory consumption and significant GPU resource demands. To address this, LCT designs a lightweight Transformer module based on the cross-channel covariance matrix of self-attention and the GELU activation function. The specific structure is illustrated in Figure 4.

For an input feature with a kernel size of k × k and C channels (where the number of input and output channels remains the same), the parameter count $C$ of a standard convolution is k × k × C × C. In contrast, depthwise separable convolution consists of a depthwise convolution (with a kernel size of k × k × 1 per channel) ollowed by a 1 × 1 point convolution. The number of parameters for depthwise convolution and point convolution are k × k × 1 × C and 1 × 1 × C × C. respectively, that is:

$$\frac{dsc}{c}=\frac{k\times k\times 1\times C+1\times 1\times C\times C}{k\times k\times C\times C}=\frac{1}{C}+\frac{1}{k^2}.$$

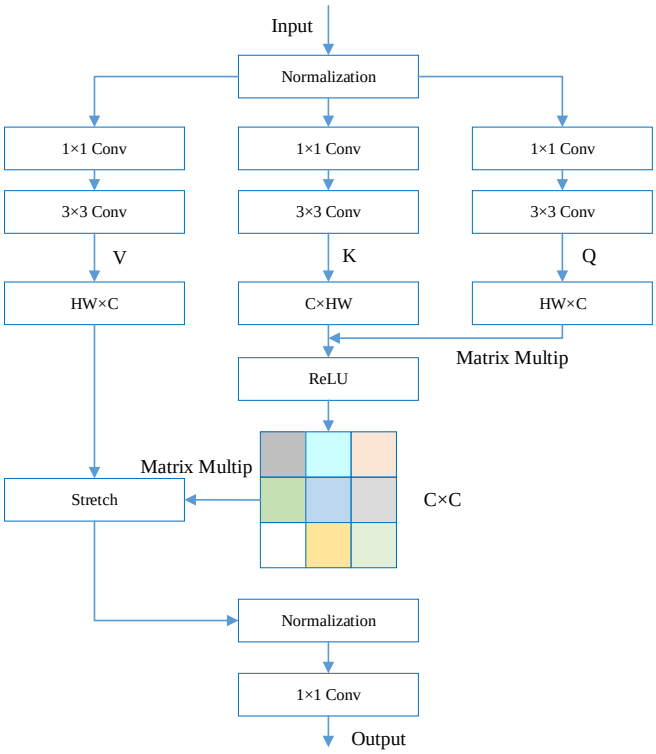


**Figure 4.** Structure of lightweight Transformer block

From the above formula, it can be seen that the number of parameters of the depthwise separable convolution kernel is approximately $\frac{1}{k^2}$ of that of the ordinary convolution kernel, which greatly reduces the number of parameters.

Existing Transformers use Softmax as the activation function, retaining the similarity of all tokens between the query (Q) and the key (K) for feature aggregation. However, not all tokens in Q are related to the tokens in K, and using all similarities does not effectively promote feature aggregation.

Considering that the GELU function has better sparsity than Softmax, this paper adopts GELU as the activation function for generating self-attention, so that pixels in sparse regions can interact and select the pixels with the highest similarity, thereby effectively promoting feature aggregation.

In summary, the Transformer module utilizes depthwise separable convolution and computes the covariance matrix of self-attention across channels, effectively reducing the number of parameters and computational cost, thus achieving lightweight networks.

# 4. Experiments and results analysis

## 4.1. Datasets and Metrics

In this work, we select the first 800 images from the DIV2K[23] dataset and apply data augmentation techniques such as rotation and horizontal flipping to construct the training set. Meanwhile, the Set5[24], Set14[25], BSD100[26], Urban100[27], and Manga109[28] datasets are used as test sets.

The reconstruction performance of the network is evaluated using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM)[29]. Additionally, the model complexity is assessed using the number of parameters and Floating-Point Operations per Second (FLOPs).

## 4.2. Comparison of Different Activation Functions in Transformer

Existing Transformer models primarily use Softmax as the activation function to generate self-attention. However, different activation functions may influence self-attention generation in various ways. For instance, DLGSANet (Effective Light-weight Dynamic Local and Global Self-Attention Network) employs ReLU as the activation function, effectively filtering out redundant self-attention information and achieving better reconstruction performance compared to Softmax.

In the experiment, we compare the performance of Softmax, Sigmoid, ReLU, and GELU activation functions across five datasets. The results are presented in Table 1, where bold numbers indicate the best values.

From the table, it is evident that using GELU as the activation function for self-attention in the Transformer consistently achieves the highest PSNR and SSIM values across almost all datasets. The reason for this improvement lies in the smoothness of the Gaussian Error Linear Unit (GELU) compared to ReLU and other activation functions. Its smoother nature allows for faster convergence during training and enhances feature aggregation, leading to improved reconstruction performance.

## 4.3. Comparisons with Advanced SISR Models

we conduct a quantitative comparison with several state-of-the-art super-resolution methods, including SRCNN (Super-Resolution CNN), ESPCN (Efficient Sub-Pixel CNN), Lap-SRN, DRCN, MADNet, LCRCA, VLESR, HNCT, IDN (Information Distillation Network), PAN (Pixel Attention Network), MRMDN (Model-Driven Recursive Multi-Scale Denoising Network), SMSR (Sparse Mask Super-Resolution), RiRSR (ResNet in ResNet Architecture), and the Lightweight Inverse Separable Residual Information Distillation Network (LIRDN).

When the super-resolution image reconstruction scale factor is set to 3×, LCT achieves the best reconstruction performance across all five datasets. The primary reason for this superiority lies in LSCT's effective integration of CNN and Transformer, leveraging their distinct feature extraction capabilities. The symmetric network structure enhances the ability to capture both local and global features, while the introduction of the High-Frequency Residual Block (HFRB) further improves the extraction of high-frequency region details, leading to superior reconstruction quality.

Figure 5 presents the comparison results when the image reconstruction scale factor is set to 2×. In the high-resolution (HR) image, the diagonal pattern in the lower half is oriented toward the bottom-right.

The reconstructed images from ESPCN, DRCN, RiRSR, and HNCT exhibit a grid-like artifact in this region. Meanwhile, the reconstructions from MRMDN and VLESR show the diagonal pattern incorrectly oriented toward the bottom-left. In contrast, only LCT successfully restores the diagonal pattern with the correct orientation and clarity, demonstrating its superior ability to preserve structural details.

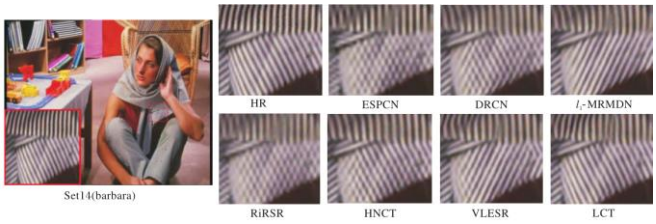


**Figure 5.** Visual comparison with other SISR models. It is obvious that LSCT can reconstruct correct photorealistic SR images.

Figure 6 presents the comparison results for image reconstruction with a scaling factor of 3×. The high-resolution (HR) image exhibits a grid-like texture, while other networks suffer from varying degrees of distortion and blurring in the right half of the reconstructed images.

In contrast, LCT successfully reconstructs more detailed and sharper textures. The fundamental reason for this improvement lies in LCT's ability to effectively combine the advantages of both CNN and Transformer architectures. Additionally, the introduction of HFRB enhances the network's focus on high-frequency region details, enabling the reconstructed images to retain clearer textures and finer details.

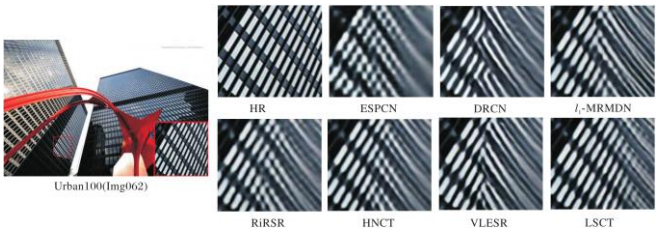


**Figure 6.** Visual comparison with other SISR models, LSCT reconstructs more and clearer textures.

## 4.4. Comparison on Computational Cost

we provide a more detailed comparison of each model. LCT achieves the highest PSNR while utilizing fewer parameters.

The key reason behind this efficiency is LCT's lightweight design, which is achieved through depthwise separable convolutions and weight sharing. Additionally, LCT effectively integrates both local and global information during the reconstruction process while enhancing high-frequency region features. This balanced approach allows LCT to optimize both model complexity and reconstruction performance, achieving superior results with reduced computational cost.

**Table 1.** Comparison of the complexity of each network

| Network Model | Parameter quantity/K | FLOPs/G | PSNR/dB | SSIM |
|---|---|---|---|---|
| DRCN | 1774 | 9788.7 | 30.75 | 0.9133 |
| MRMDN | 1380 | 499.3 | 32.11 | 0.9283 |
| MADNet | 878 | 187.1 | 31.59 | 0.9234 |
| SMSR | 985 | 224.1 | 32.19 | 0.9284 |
| LIRDN | 1171 | 221.4 | 32.22 | 0.9287 |
| LSCT | 842 | 423.1 | 32.4 | 0.9302 |

# 5. Conclusion

To address the challenges of large network parameters and high computational costs in existing image super-resolution reconstruction networks, this paper proposes a lightweight

CNN-Transformer-based image super-resolution reconstruction method (LCT).

First, a CNN-Transformer module is designed using weight sharing, where the Channel Attention (CA) mechanism is employed to fully integrate extracted information, enhancing the reconstruction of both local and global features. By leveraging depthwise separable convolutions and computing the self-attention cross-channel covariance matrix, LCT effectively reduces the number of Transformer parameters, lowering computational costs and memory consumption, thereby achieving a lightweight network design.

To mitigate the issue of high-frequency information loss during the Transformer's feature extraction process, a High-Frequency Residual Block (HFRB) is introduced to focus on high-frequency region details, capturing more texture information. Furthermore, this paper explores the impact of activation function selection in Transformer-based self-attention generation and finds that the GELU activation function effectively promotes feature aggregation and enhances performance.

Extensive experiments demonstrate that LCT can effectively reconstruct images with richer textures and sharper edge details while maintaining a lightweight network structure. However, from a visual perspective, LCT-reconstructed high-resolution images still exhibit artifacts and slight blurring. Future research will further address these limitations, striving to enhance the quality and visual fidelity of reconstructed images while maintaining network efficiency.

# Acknowledgment

This work was supported in part by the Faculty of Engineering Science and Technology, Infrastructure University Kuala Lumpur (IUKL). This research was also supported by the Education Department of Shaanxi Province, China, under Project No. 24JK0437, titled "Research on Super-Resolution Technology for Complex Industrial X-ray Images," and by the 2024 Qin Chuang Yuan Scientific Research Special Project of Shaanxi Province under Project No. 2024QCY-YJ17, titled "Research on Deep Learning-Based Super-Resolution Technology for Industrial X-ray Images." We thank Dr. Tadiwa Elisha Nyamasvisva, the doctoral advisor, for his valuable guidance and continuous support throughout this research.

# References

[1] A. Adler, Y. Hel-Or, and M. Elad, "A shrinkage learning approach for single image super-resolution with overcomplete representations," in Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11, Springer, 2010, pp. 622–635.

[2] J. Chen et al., "Transunet: Transformers make strong encoders for medical image segmentation," ArXiv Prepr. ArXiv210204306, 2021.

[3] Z. Guo et al., "Super-resolution integrated building semantic segmentation for multi-source remote sensing imagery," IEEE Access, vol. 7, pp. 99381–99397, 2019.

[4] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646–1654.

[5] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 286–301.

[6] Y. Mei, Y. Fan, and Y. Zhou, "Image super-resolution with non-local sparse attention," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3517–3526.

[7] X. Wu, K. Zhang, Y. Hu, X. He, and X. Gao, "Multi-scale non-local attention network for image super-resolution," Signal Process., vol. 218, p. 109362, 2024.

[8] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1637–1645.

[9] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3147–3155.

[10] [Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in Proceedings of the 27th acm international conference on multimedia, 2019, pp. 2024–2032.

[11] L. Zha, Y. Yang, Z. Lai, Z. Zhang, and J. Wen, "A lightweight dense connected approach with attention on single image super-resolution," Electronics, vol. 10, no. 11, p. 1234, 2021.

[12] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "MADNet: A fast and lightweight network for single-image super resolution," IEEE Trans. Cybern., vol. 51, no. 3, pp. 1443–1453, 2020.

[13] C. Peng, P. Shu, X. Huang, Z. Fu, and X. Li, "LCRCA: image super-resolution using lightweight concatenated residual channel attention networks," Appl. Intell., pp. 1–15, 2022.

[14] H. Feng, L. Wang, Y. Li, and A. Du, "LKASR: Large kernel attention for lightweight image super-resolution," Knowl.-Based Syst., vol. 252, p. 109376, 2022.

[15] D. Gao and D. Zhou, "A very lightweight and efficient image super-resolution network," Expert Syst. Appl., vol. 213, p. 118898, 2023.

[16] W. Wang, Y. Zhu, D. Ding, J. Li, and Y. Luo, "Multi-Scale Multi-Stage Single Image Super-Resolution Reconstruction Algorithm Based on Transformer," in 2022 21st International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), IEEE, 2022, pp. 111–114.

[17] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 457–466.

[18] J. Fang, H. Lin, X. Chen, and K. Zeng, "A hybrid network of cnn and transformer for lightweight image super-resolution," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 1103–1112.

[19] X. Li, J. Dong, J. Tang, and J. Pan, "Dlgsanet: lightweight dynamic local and global self-attention networks for image super-resolution," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 12792–12801.

[20] Y. Zheng, S. Liu, H. Chen, and L. Bruzzone, "Hybrid FusionNet: A hybrid feature fusion framework for multi-source high-resolution remote sensing image classification," IEEE Trans. Geosci. Remote Sens., 2024.

[21] H. Gu, L. Su, Y. Wang, W. Zhang, and C. Ran, "Efficient Channel-Temporal Attention for Boosting RF Fingerprinting," IEEE Open J. Signal Process., 2024.

[22] A. Li, L. Zhang, Y. Liu, and C. Zhu, “Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 12514–12524.

[23] E. Agustsson and R. Timofte, “NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA: IEEE, Jul. 2017, pp. 1122–1131. doi: 10.1109/CVPRW.2017.150.

[24] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, “Low-complexity single-image super-resolution based on nonnegative neighbor embedding,” 2012.

[25] R. Zeyde, M. Elad, and M. Protter, “On single image scale-up using sparse-representations,” in Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7, Springer, 2012, pp. 711–730.

[26] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, “Contour detection and hierarchical image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 5, pp. 898–916, 2010.

[27] J.-B. Huang, A. Singh, and N. Ahuja, “Single image super-resolution from transformed self-exemplars,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5197–5206.

[28] Y. Matsui et al., “Sketch-based manga retrieval using manga109 dataset,” Multimed. Tools Appl., vol. 76, pp. 21811–21838, 2017.

[29] X. Gao, W. Lu, D. Tao, and X. Li, “Image quality assessment based on multiscale geometric analysis,” IEEE Trans. Image Process., vol. 18, no. 7, pp. 1409–1423, 2009.