IET Image Processing

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Lightweight Multi-Stage Holistic Attention-Based Network for Image Super-Resolution

Aatiqa Bint E Ghazali[1] | Ahsan Fiaz[1,2] | Muhammad Islam[3]

[1]Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad, Pakistan | [2]Faculty of Computer Science, Government College University, Faisalabad, Pakistan | [3]College of Science and Engineering, James Cook University, Cairns, Queensland, Australia

Correspondence: Muhammad Islam (muhammad.islam1@my.jcu.edu.au)

Received: 23 November 2024 | **Revised:** 15 January 2025 | **Accepted:** 1 February 2025

Funding: The authors received no specific funding for this work.

ABSTRACT

High-resolution images are crucial for many applications, but factors such as environmental conditions can reduce image quality. Super-resolution (SR) techniques address this by generating high-resolution images from low-resolution inputs. While deep learning SR models have made significant progress, they can be computationally expensive and struggle with differentiating between various image scales. Lightweight SR methods, suitable for resource-constrained devices, often compromise image quality. This study introduces a multi-stage holistic attention-based network, using Gaussian Laplacian pyramids to decompose images and apply holistic attention modules at each level. This approach reduces parameters and computational costs while maintaining image quality, achieving a PSNR score of 28 and SSIM of 0.91 with only 29,000 parameters. The model demonstrates the potential for efficient and high-quality image reconstruction. Future work will focus on improving quality while minimizing costs and exploring other advanced techniques. The code will be made available upon request

1 | Introduction

Single image super-resolution (SISR) aims to reconstruct high-resolution images from low-resolution inputs. This type of reconstruction has numerous applications in medical imaging, satellite surveillance, and digital photography, where detailed information extracted from images is critical. However, in real-time scenarios, environmental conditions and technical limitations often degrade the resolution of an image; therefore, such lost details can only be recovered using super-resolution (SR) techniques. Nearest neighbour and bicubic interpolation are traditional SR approaches; they are computationally simple but fail to do well in complex scenarios where high-frequency details are missing [1–3]. More advanced techniques that include sparse coding-based methods have tried to enhance image reconstruction through the analysis of image patches and hand-crafted features [4, 5]. Even though better performances were obtained by these methods, they performed poorly for detailed recovery scenarios.

Deep learning has significantly improved the performance of SISR-based schemes. The models were dominated by CNNs, and then SRCNN [6] and SRGAN [7] have been explored and show comparative superiority over traditional approaches. Such models are known to extract deeper features while achieving better reconstructions. Attention mechanisms, residual learning, feature pyramids, and dense connections are employed in [8–10] for enhancing reconstruction quality. A deep holistic attention based network is proposed in [11] which used a combination of layer and channel attention as holistic attention to reconstruct a high resolution image with better imager quality. The capacity of single image super-resolution (SISR) to improve image quality and recover fine features from low-resolution inputs has made it a crucial method that has been widely used in many different

fields. SISR techniques, such as those put forth in [12–15], use multimodal multi-head convolutional attention with different kernel sizes in medical imaging to enhance the diagnostic quality and resolution of medical scans. EDiffSR, an effective diffusion probabilistic model created to handle the particular difficulties of super-resolving satellite and aerial pictures, was first presented in the field of remote sensing in [16]. Researchers in [17] created a dual-path deep fusion network for face image hallucination in the surveillance domain, which allowed for the reconstruction of high-resolution facial features for better recognition. To further improve the precision and detail of depth maps, authors in [18] showed how to employ recurrent structure attention guidance for depth super-resolution. These uses highlighted the SISR's adaptability and importance in addressing domain-specific problems and promoting image processing breakthroughs.

Super-resolution approaches have been used in medical imaging to enhance diagnostic image clarity, which helps with illness analysis and identification [19]. Similar to this, sophisticated SR techniques in satellite surveillance have been used to improve low-resolution satellite footage for better object tracking and scene comprehension by employing deformable convolution and temporal grouping [20]. Deep fusion networks have also been used in digital photography for face image hallucination, which allows for the reconstruction of high-quality facial photographs from inputs with poor resolution [21].

Such models are known to extract deeper features while achieving better reconstructions. Recent works have introduced further innovations in the field, enhancing super-resolution performance through more sophisticated techniques. Some of these methods explore multi-task interaction learning [22], while others focus on model-informed multi-stage networks [23] for enhanced efficiency. Techniques such as deep fusion networks and holistic attention mechanisms have also been implemented in various applications, from hyperspectral image super-resolution to medical and satellite imagery [24, 25]. These advancements demonstrate the growing capability of deep learning in super-resolution tasks.

Vision transformer-based networks are proposed in [26] to better extract fine details from low-resolution images. An enhanced laplacian pyramid-based approach is employed in [27] as a generator of generative adversarial networks (GANS) which gradually reconstructs HR pictures at several levels of the pyramid. However, deep networks are limited on resource-constrained devices, and the methods often fail to incorporate fine texture details due to augmentation, which might deteriorate the visual quality of super-resolved images. Moreover, the absence of parameter sharing frequently results in an excessive number of network parameters.

To ease the computational constraints of deep models, lightweight models have been proposed. For instance, in [28] the concept of residual block-based SR-CNN is introduced to achieve a significantly reduced computational demand preserving quality. In a similar improvement towards efficiency, a dual encoder–decoder network with a scalable guided module is introduced in [29]. The other approaches used in this paper include channel attention models with dense connections, as proposed in [30], and conditional weighting strategies for the reduction of expensive convolution operations, as provided in [31]. Although such progress was made, the lightweight models do so at the cost of image quality since they reduce their parameters thereby developing a form of efficiency-accuracy trade-off.

This paper presents a novel multi-stage holistic attention-based network that addresses the issues mentioned above. We first decompose an image with a pyramid decomposition using the Gaussian–Laplacian pyramid; subsequently, we perform an efficient hierarchical feature extraction process. Recursive blocks and layer attention module from holistic attention network is applied at each level for better feature extraction and reconstruction of images while reducing computations. HAN [11] takes an approach that is more comprehensive to highlight informative features across several dimensions, in contrast to conventional attention mechanisms that concentrate on either the spatial or the channel dimensions alone. This guarantees the model's ability to acquire global contextual dependencies, which are essential for tasks such as image reconstruction and super-resolution. As such, the model is compatible with resource-constrained devices without compromising image quality. The key contributions of this work are the following:

1. A Laplacian pyramid-based network is introduced that decomposes the low-resolution image into different frequency bands using Gaussian and Laplacian pyramids. This multi-level decomposition enables the isolation of high-frequency details and focused feature extraction at each level to further enhance the overall clarity and sharpness of the reconstruction HR image.
2. A layer attention modules taken from holistic attention-based mechanism [11] is enforced at each level of the pyramid. The model focuses on important features while avoiding unnecessary computations. The attention module allows longer-range dependencies in pixels by the network, hence enabling the network to regain fine details across multiple scales at a minimal increase in computational loads.

Such efforts would help in overcoming the drawbacks of existing techniques by giving an efficient solution with high quality and being resource-effective towards the generation of HR images from LR inputs.

2 | Related Work

Single Image Super Resolution (SISR) is a field that has been under intense research studies recently. Early methods in SISR were stuck in the traditional interpolation methods [32] and learning-based approaches like those in [33, 34]. Later on deep learning-based approaches changed the game. Dong et al. [6] were the first to employ CNNs for SISR-related problems with significant performance improvements. This inspired continued research into even deeper and more efficient architectures for such SISR problems.

A multi-stage architecture-based image restoration network is proposed in [35] for different tasks where initial stages integrate the encoder–decoder networks, and the final stage employs a

network that works on the original input resolution. A coarse-to-fine technique has been used to divide feature extraction and image reconstruction into two different blocks in [36], to allow the information to flow from shallow layers to deep layers. The residual learning together with traditional bicubic upsampling has been added to a CNN-based model with guided filters [37]. Although this let the information flow from shallow layers to deep layers and further saved many parameters of the guided filters, it still increased the computational cost of the model caused by bicubic interpolation. In [38], a multiscale skip connection-based network was designed that works thoroughly on multi-scale image features by combining dilated and standard convolutions to improve feature extraction. Recently, lightweight SISR models have attracted a lot of interest due to the demand for efficiency in resource-constrained devices. The s-LWSR network in [28] removes parameters by using inverted residual blocks. Additionally, it removes some activation layers at the end for feature degradation, thus rendering the model highly efficient. Similarly, a scalable guided encoder-decoder network [29] can be pruned after training to form a lightweight version usable for real-time applications. An example of such high-performance-based residual dense connection and attention blocks is Mad-Net [30], which however increases model complexity with its dense connections.

Recently, pyramid-based networks have been extensively used to reduce the computational costs. Decomposition of an image at various scales, as discussed in [39] and [40], reduces the computational effort for feature extraction. A deep Laplacian pyramid-based network was proposed in [41] with the concept of CGAN. Its aim is to achieve improved results by establishing connections through long and short skip connections throughout the network.

Moreover, attention mechanisms have played a vital role in progressing SISR performance. A spatial attention-based model [42], incorporates bottleneck modules to boost spatial resolution. The residual channel-spatial attention module proposed in [43] combined with a dense sampling method outperforms their previous results. Holistic attention network [11], remedied the issue of smoothing in super-resolved images by making layer-specific attention weights preserve fine details and improve feature representation. Some models focus on spatial details through non-local residual groups and improve the reconstruction of images significantly as presented in [10]. Despite these improvements from attention-based models in various sectors, many of these, including those in [44], do not consider layer-to-layer correlations that result in smooth yet less detailed outputs. Pyramid-based approaches further advanced the multi-scale image representation for SISR. In [45], and [46], pyramid structures were employed for extracting features at various scales. Encoder–decoder sub-networks were used in this regard to prevent information loss.

In [47] an image formation-guided network is proposed which uses pixel substitution to enforce constraints. The approach relies heavily on enforcing image formation constraints during training. In [48] a multistage architecture is employed to reconstruct the image gradually. The subnetwork used at different levels of the proposed network revolves Recursive networks and transformer-based designs have both been used in recent SISR developments to obtain better performance. around encoder–decoder and attention-based methods. A transformer-based technique also entered into the scene with the pioneering SRFormer [49], which used the mechanism of self-attention to reduce the computational complexity without losing the quality of images. The DiVA mechanism that captures long-range spatial correlations, while taking advantage of inter-channel dependencies to extract more discriminative features also sparked as a renewal [50]. The TTST (Top-K Token Selective Transformer), which was proposed by Xiao et al. [51], effectively processes high-resolution characteristics in remote sensing images by using a token-selective method. This technique shows how transformers can optimize computational overhead while capturing long-range dependencies. However, Jiang et al. [52] presented a hierarchical dense recursive network that uses recursive connections to efficiently take use of hierarchical feature representations. In order to maintain fine-grained details during reconstruction, this architecture places a strong emphasis on dense connections. Even while these techniques produce amazing outcomes, they frequently call more intricate network architectures or higher processing expenses. In contrast, our proposed lightweight multi-stage holistic attention-based network achieves comparable reconstruction quality with significantly fewer parameters, making it more suitable for resource-constrained environments.

In brief, SISR methods have developed from traditional techniques to quite sophisticated models led by CNNs, mechanisms of attention, and transformers. Though deep learning has been significantly contributing to improving the quality of images, big challenges exist in designing models that balance computational costs and are performance-oriented. Pyramid networks and the attention-based method continue to shave off the trade-off of these models, thus making it possible for them to be useful for real-time environments with resource constraints.

3 | Methodology

The proposed network model for single-image super-resolution is shown in Figure 1. The model works by decomposing a low-resolution image into a Laplacian pyramid and then building a sub-network for each pyramid level. Each sub-network has the same architecture and has its loss function which is used in the training. The super-resolved image is generated as a Gaussian pyramid by the network. The bottom level of the Gaussian pyramid is the final reconstructed image. So overall the model passes through three stages, image decomposition in the form of Laplacian pyramid's levels, sub-network, and image reconstruction using Gaussian pyramids. The proposed lwMHAN uses five sub-networks which are trained independently using their loss function. The architecture of these sub-networks is the same. Figure 2 shows the structure of the sub-network that is used in this study. After the decomposition of the image, a single image goes as input to each sub-network. At first shallow features are extracted using a convolutional layer and for deep feature extraction five recursive blocks are used. Inside each recursive block image is convoluted using three convolutions and the LReLU activation function is used in between. The recursive blocks strategy is adapted for sharing parameters among different layers and to reduce cost. After recursive blocks, the information is passed to a layer attention module (LAM) taken from HAN

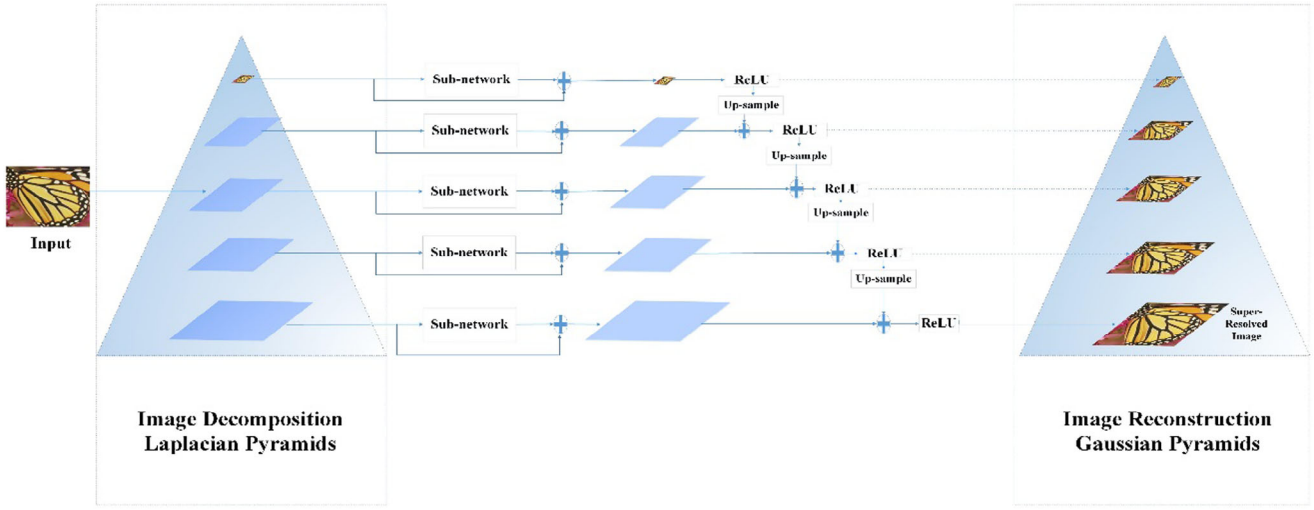

FIGURE 1 | Architecture of the proposed network.

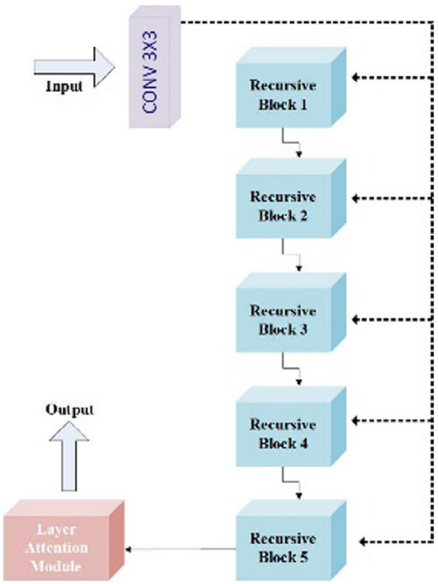

FIGURE 2 | Subnetwork.

[11] to extract the information that exists between different layers. HAN [11] takes an approach that is more comprehensive to highlight informative features across several dimensions, in contrast to conventional attention mechanisms that concentrate on either the spatial or the channel dimensions alone. This guarantees the model's ability to acquire global contextual dependencies, which are essential for tasks such as image reconstruction and super-resolution. Residual learning is utilized to facilitate the training of the model. The proposed network is comprised of three main modules: Laplacian pyramids, sub-network, and image reconstruction.

3.1 | Laplacian Pyramids

The Laplacian pyramid is a band-pass image decomposition derived from the Gaussian pyramid (GP). The Gaussian pyramid (GP) is a multi-resolution picture representation created through a recursive reduction (low-pass filtering and decimation) of the image data set. A modified version of the Laplacian pyramid [39] is used here which was created for the image-deraining process. The low-resolution images passed to the network are first decomposed into different levels of the Laplacian pyramid through Equation (1).

$$Ln(X) = Gn(X) - up\,sample\,(Gn + 1(X)), \tag{1}$$

where G_n is the Gaussian pyramid, $n = 1 \ldots N - 1$. With $G_1(X) = X$ and $L_N\ (X) = G_N(X)$, the function $G_n(X)$ is computed by down-sampling $G_{n-1}(X)$ with a Gaussian kernel. The network will be able to take advantage of sparsity at all levels because of this decomposition strategy. The main reason for choosing the combination of Laplacian and Gaussian pyramids is their low cost in addition to multi-scale feature representation which will help to better reconstruct the image.

3.2 | SUB-Network

After decomposition, a sub-network is designed for each pyramid level independently.

Each sub-network has a similar architecture with a different number of filters. The sub-network consists of five recursive blocks inside which residual learning is used on three convolutional layers and the LReLU activation function.

3.2.1 | Feature Extraction

Two types of feature extraction strategies are used here. One is shallow to extract course features and the other is deep extraction to extract fine or high-level features. For the first one, a single convolutional layer is defined using Equation (2) as mentioned in HAN [11]

$$F_0 = \mathrm{Conv}\,(I_L R) \tag{2}$$

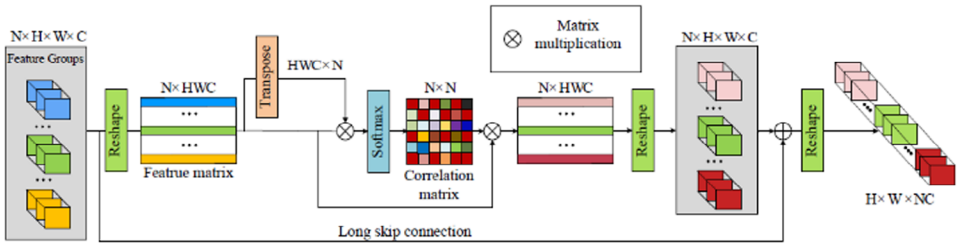

FIGURE 3 | LayerAttention Module [11].

and for deep feature extraction of recursive and residual blocks with holistic attention are used.

3.2.2 | Recursive and Residual Blocks

It is well-known that deep architecture can extract better information as compared to shallow networks. This is because they use more layers for that purpose. However, training these networks is not a piece of cake. Moreover, more layers bring more cost making the model computationally expensive. In recursion, different layers share parameters by calling themselves again and again. To gain the benefits of this strategy a group of recursive blocks are used in this model which consist of three convolutional layers with LReLU activation function. He initialization [53] is used to initialize weights for these layers. Filters of size 3 × 3 are used inside each layer. All recursive blocks are linked with each other using residual connections which can facilitate training.

3.2.3 | Layer Attention

To extract deep features from images at each level of the pyramid Holistic attention network HAN [11] is studied in detail which aims to improve feature representations' representational quality for super-resolution. HAN introduced a novel type of attention mechanism comprised of two types of attention modules that is layer attention module (LAM) to learn parameters of hierarchical layers and channel-spatial attention module (CSAM) to focus on channel and spatial interdependencies of layers. After extensive experiments, the authors chose LAM from HAN [11] as a subnetwork for our proposed scheme. The reason behind choosing LAM is discussed in the section of ablation studies. Layer attention module as shown in Figure 3 is utilized to treat the issues of feature correlation among different layers. It is used in the subnetwork of the proposed model. The concatenated features from recursive blocks are reshaped into a 2D matrix in dimension NxHWC and transposed in the layer attention module. After this, the variable storing values of previous operations are multiplied as shown in Equation (3) of LAM [11].

$$w_{\{i,j\}} = \delta(\phi(F_r))_i \cdot (\phi(F_r))_j^T, \quad i, j = 1, 2, \ldots N, \tag{3}$$

where δ and φ represent the softmax and reshape operations, respectively. The predicted correlation matrix is then multiplied with the reshaped feature map as shown in Equation (4) from HAN [11].

$$F_{\text{LAM}} = \alpha \sum w_{i,j} F_r \tag{4}$$

In the end, all the results are added to the input of LAM. In this way attention weights are applied to the features of several layers and softmax is applied before element element-wise sum of the input and attentive features. LAM learns the association between features of varying depths, hence improving the ability to represent features automatically. The overall module LAM [11] is represented by Equation (5).

$$\text{LAM} = F_{\text{LAM}} + (F_r)_j \tag{5}$$

3.3 | Image Reconstruction

The last layer of the Gaussian pyramid is the final output image and to obtain that output, the Laplacian pyramid is used which can be defined by using Equation (6) from the reconstruction phase described in [39].

$$\begin{aligned} G_N(Y) &= \max(0, L_N(Y)), \\ G_n(Y) &= \max(0, L_n(Y) + \text{upsample}(G_{n+1}(Y))), \end{aligned} \tag{6}$$

where $n = 1 \ldots N - 1$. To correct the outputs rectified linear units(ReLU) are used which is represented as max(0,x) and it returns the values between 0 and 1 while avoiding the issue of dead neurons.

3.3.1 | Loss Function

Since each pyramid level has its sub-network a separate loss function is used to train each sub-network. Mean square loss smooths the edges of images so it can degrade the quality of reconstructed images. So each sub-network is trained using the l1 loss function as shown in Euation (7).

$$\text{l}(x, y) = 1/M \sum_{(i=1)}^{M} (I_{\text{SR}} - I_{\text{HR}}) \tag{7}$$

l1 loss does not over-penalize larger errors and can therefore preserve structures and edges. In contrast, the commonly used l2 loss frequently produces over-smoothed results because it penalizes larger errors while tolerating smaller ones. Therefore, l2 faces challenges to preserve the image's deeper structures compared to l1.

4 | Experiments

4.1 | Experimental Setup

All the experiments are performed in Python 3.6 using tensor flow 1.0.1 on Google collaboratory Pro. In each training batch, 16 patches of size 128 × 128 are selected randomly. 50 iterations of back-propagation constitute an epoch. The learning rate 1×10^{-5} is used for training the model with a batch size of 16 and it is optimized using Adam optimizer. As batch normalization blurs the results of SISR it is not utilized in the model as a regularization parameter. Data augmentation is performed to make data sufficient for the model. Each image in the dataset is rotated and translated with an angle of 90°. In the Gaussian pyramid reconstruction feature maps are divided by different values to restore a better version of the image. To check the quality of reconstructed images two well-known evaluation metrics PSNR and SSIM are used.

4.1.1 | PSNR

Peak signal-to-noise ratio (PSNR) is a traditional method used to check the quality of images. It is defined as the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation It is like mean square error (MSE)and can be calculated by taking the log of MSE as shown in Equation (8) [54].

$$\mathrm{PSNR} = 10\log 10(\max/\mathrm{MSE})^2 \tag{8}$$

4.1.2 | SSIM

Structure similarity index works by considering three factors that is, luminance, structure, and contrast. Structural information is more important as it works on the idea of pixel inter-dependencies especially when they are spatially close to each other. As the SSIM formula is based on three comparison measurements it can be calculated as shown in Equation (9) [54]. Whereas Equations (10)–(12) describe its parts:

$$\mathrm{SSIM}\,(x,y) = l\,(x,y)\cdot c\,(x,y)\cdot s\,(x,y) \tag{9}$$

where

$$l\,(x,y) = \frac{2\mu x\mu y + c1}{\mu x2 + uy2 + c1} \tag{10}$$

$$c\,(x,y) = \frac{2\partial \upsilon xy + c2}{\partial x2 + \partial y2 + c2} \tag{11}$$

$$s\,(x,y) = \frac{2\partial xy + c3}{\partial x\partial y + c3}. \tag{12}$$

4.2 | Datasets

The DIV2K dataset is used to train the proposed model which consists of 1000 images out of which 800 are for training and 200 for validation and testing as shown in Table 1. This dataset contains all images in PNG format which are low in resolution and these LR images are specifically for the task of single images super-resolution. High-resolution images are also given as ground truths. For testing two well-known datasets are used. URBAN100 contains challenging urban scene images with varying frequency band details, and MANGA109 is a Japanese manga dataset. The complete dataset organization is shown in Table 2.

TABLE 1 | Training dataset organization.

Dataset	Training samples	Validation samples
DIV2K	800	200

TABLE 2 | Test datasets organization.

Dataset	Samples
Urban 100	100
Manga109	109

TABLE 3 | Model with HAN and LAM.

Methods	PSNR	SSIM	Parameters
With HAN	16.70	0.55	122714
With LAM	28	0.91	29k

4.3 | Ablation Study and Discussion

We present a detailed examination of the different components of the model and their contributions to the overall architecture and performance. Our analysis encompasses a thorough review of critical layers, connections, and specific design choices that play a significant role in shaping the model's outcomes. By carefully assessing each of these elements, we aim to understand their impact on the model's behavior and overall success. Through this ablation study, we systematically modify or remove various parts of the model to determine their influence on performance. This analysis enables us to identify the key drivers of the model's success and potential areas for optimization which can guide future improvements and enhancements in model design.

4.3.1 | Holistic Attention Module as a Sub-Network

The holistic attention network (HAN) is employed as a sub-network within the proposed model. However, as indicated in Table 3, the model's performance is sub-optimal when using a pyramid-based structure. This approach incurs additional costs, resulting in thousands of extra parameters, without providing significant improvements in image quality as measured by PSNR and SSIM. We found that adding HAN [11] results in a considerable increase in the number of parameters in the model (122,714) and a significant decrease in performance measures (PSNR: 16.70, SSIM: 0.55). This implies that the lightweight pyramid-based structure of lwMHAN is incompatible with the dense connections of HAN, notwithstanding their effectiveness in other settings. The ensuing inefficiencies emphasize how crucial it is

TABLE 4 | Model with and without layer attention module.

Methods	PSNR	SSIM	Parameters
With LAM	28	0.91	28.309K
Without LAM	25	0.85	29k

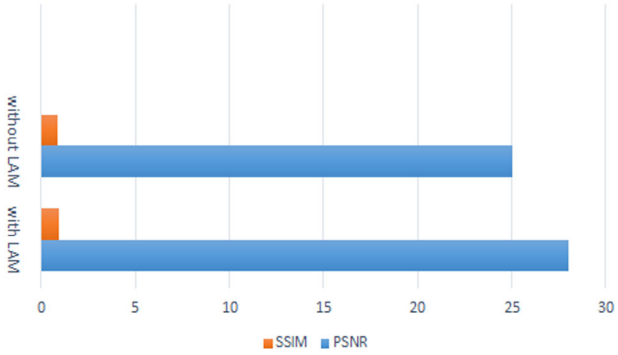

FIGURE 4 | Effect of adding LAM on quality and cost.

to develop components that complement our model's lightweight nature. One possible explanation for this is that the proposed model adopts a lightweight approach based on a divide-and-conquer strategy. Furthermore, integrating a dense model as a sub-network may not only disrupt the gradients but also hinder the model's ability to converge effectively, thereby compromising its overall performance. Consequently, while the pyramid-based structure theoretically expands the model's capacity to capture long-range dependencies, the practical implementation presents challenges that outweigh its benefits. Additionally, the increase in parameters leads to higher computational costs during both training and inference, making the model less efficient for real-world applications.

4.3.2 | Layer Attention Module as a Sub-Network

The results presented in Table 3 demonstrate that the model struggles when using a pyramid-based structure with holistic attention network. This approach increases costs by adding thousands of extra parameters without yielding significant improvements in image quality, as measured by PSNR and SSIM. Instead of using HAN model fully in pyramids, only the layer attention module from this model is integrated with recursive blocks in the proposed model to assess the outcomes. As indicated in Table 4, removing this module from the model resulted in reductions in both PSNR and SSIM. On the other hand, its inclusion brought minimal additional parameters to the network. This can be attributed to the layer attention's foundation on channel attention, which leverages the information between feature channels and also manages inter-layer information. Figure 4 illustrates the shifts in cost and quality of the model when adding and removing LAM. With only a small increase in parameters (29K), the integration of LAM significantly boosts performance (PSNR: 28, SSIM: 0.91). Both PSNR and SSIM drop when LAM is removed (PSNR: 25, SSIM: 0.85), indicating how important LAM is for improving feature representation across layers. This module successfully strikes a compromise between image quality and processing efficiency. Based on this analysis, LAM was chosen as part of the recursive blocks to be employed as a sub-network within proposed model (lwMHAN). This selection not only optimized the model's performance but also minimized the increase in computational costs, thus enhancing its efficiency for real-world applications.

TABLE 5 | Model with and without recursive blocks.

Methods	PSNR	SSIM	Parameters
With recursive blocks	28	0.91	29K
Without recursive blocks	22	0.70	50K

4.3.3 | Recursive Blocks for Reducing Parameters

To thoroughly assess the effectiveness of recursive blocks, we conducted an experiment where these blocks were removed from the model. We then evaluated the performance in terms of parameters, PSNR, and SSIM on the Manga109 dataset. This careful analysis was aimed at understanding how the removal of recursive blocks would impact both the model's cost and its accuracy. The results, as detailed in Table 5, demonstrate a notable decrease in the model's overall cost, which can be attributed to the sharing of parameters across the layers of the recursive blocks. This sharing significantly reduces the parameter count and computational overhead. Despite this cost reduction, the model maintains its performance quality, with no detrimental effect on PSNR and SSIM. By allowing parameter sharing, recursive blocks allow the model to have 29K parameters while still performing well (PSNR: 28, SSIM: 0.91). When these blocks are removed, metrics significantly decrease (PSNR: 22, SSIM: 0.70), but parameters increase to 50K. This illustrates how well the blocks work to create a simplified design without sacrificing output quality. These findings highlight the efficiency gains that can be achieved through the strategic use of recursive blocks in the model. By optimizing the network structure and reducing unnecessary complexity, we can achieve a more streamlined model that retains high-quality output without excessive computational expense.

4.4 | Results

In this section the impact of different parameter settings on the proposed model's performance is presented. The heading presents a comprehensive analysis of how variations in the loss function, pyramid structure, and learning rate affect the model's efficiency, convergence, and overall image quality. By experimenting with different configurations, the study identifies optimal parameter choices that balance high-quality output with minimized resource consumption. We provide detailed insights into the effectiveness of specific parameter-tuning strategies and their role in enhancing the performance of the proposed model.

4.4.1 | Loss Function

Since the l2 loss function works well with SSIM in many image restoration techniques, it is also tested for SISR. However, it gives

FIGURE 5 | Error curve during training on DIV2K dataset.

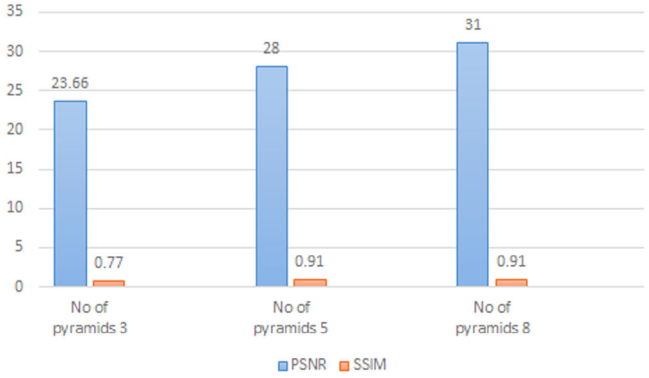

FIGURE 6 | Quality metrics analysis with different number of pyramids.

dark areas in the output image and the model took more time to converge on this loss function. More time in convergence means the model might get stuck in finding the optimal solution and will remain pending. As compared to this function l1 loss is tried and tested in this study and it gives efficient results and helps the network to converge fast. The change in L1 loss over training is seen in Figure 5. In the early phases of training, the loss first begins at a relatively low value before rapidly increasing. Weight changes when the model adjusts its parameters could be the cause of this increase. The loss then steadies and progressively drops, suggesting that the model is learning successfully. The overall trend indicates convergence despite early oscillations, indicating the efficiency of our suggested methodology in reducing error over time.

4.4.2 | Pyramid Structure

The number of pyramids decides the decomposition factor of input images. The proposed model is tested with different numbers of pyramids and qualitative measures along with network parameters are calculated with each one. As shown in Figure 6 reducing the number of pyramids reduces the total number of parameters. However, it decreases the quality of output images. Similarly adding more pyramid numbers improves the performance of the model but it brings more computational cost. Experiments show that five pyramids best performed as compared to other numbers in terms of parameters and PSNR.

TABLE 6 | Model complexity of proposed lwMHAN.

Model parameters	Flops (M)	Inference time (ms)
29K	58	20

4.4.3 | Learning Rate

Learning rate is an important parameter which is the step size that the model will take during training in the process to minimize loss and reach global minima. Two learning rates are tested in the proposed study. When the learning rate is 1×10^{-4} PSNR suddenly starts decreasing with upcoming iteration. This is because if the value of the learning rate is too large concerning the model then results are unstable due to overshooting of minima and it will not be able to converge. To this end learning rate with the value of 1×10^{-5} is tested and it works well with the results. Figure 7 shows the PSNR curve with different learning rates.

4.5 | Model Complexity

The complexity of the proposed model is being evaluated in terms of number of parameters m flops and Inference time. Table 6 shows the overall complexity of proposed model.

4.6 | Comparison With State-Of-the-Art Methods

Selected comparison schemes are retrained on the DIV2K dataset and tested on benchmark datasets. The results of the proposed method are compared to the following image restoration methods, LESRCNN [55], s-LWSR [28], swinIR [26], shuffleMixer [56] and HAN [11]. Results are compared by conducting extensive experiments using two benchmark datasets URBAN100 and MANGA109. Figures 8 and 9 show a comparison between proposed lwMHAN and state-of-the-art models in terms of visual quality. The results are obtained by conducting experiments on the Urban100 dataset for $x2$ scale. Figure 10 shows the comparison between different models and our proposed model in terms of network parameters which clearly shows that our proposed model is lightweight among all other models. As more trainable parameters bring more cost so in the proposed model total number of parameters is reduced using recursive blocks and by skipping the pre-processing step of up-sampling. Table 7 shows the results based on two quality metrics that is, PSNR and SSIM with a total number of parameters, flops and inference time in comparison of our proposed model with s-LwSR [28], LESRCNN [55], and HAN [11] whereas Table 8 compares proposed model in terms of parameters with advanced models such as swinIR [26], shuffle-Mixer [56] and IFIN [57]. It can be seen that our method has less number of parameters, flops and takes less time as compared to other schemes and it has achieved a good PSNR and SSIM score as compared to the two models. Bringing more pyramid levels into the architecture might beat the other models as well. However, it will not be a good choice for resource-constrained devices. Moreover, our proposed model demonstrates its superiority in terms of computational efficiency, making it

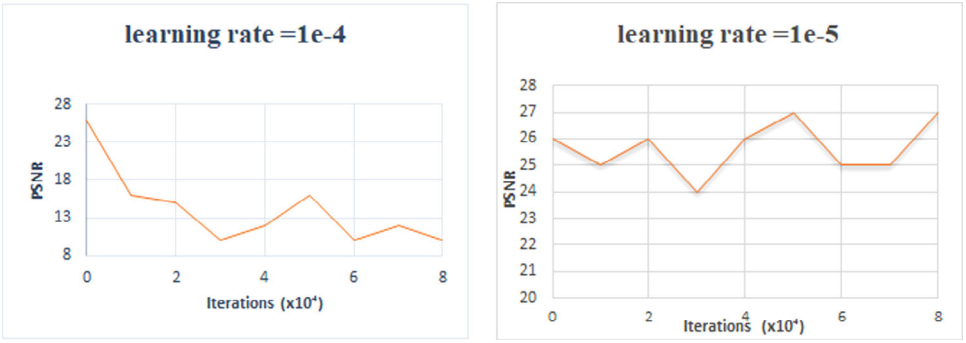

FIGURE 7 | PSNR value with different learning rates.

TABLE 7 | Comparative analysis based on PSNR, SSIM, parameters, and time with flops.

Methods	PSNR	SSIM	Parameters	Flops (M)	Inference time ms
LESRCNN	25.77	0.77	0.515m	1030	100
sLwSR	26.87	0.79	2.277m	4500	250
HAN	31.42	0.91	15.6m	31200	400
lwMHAN	28	0.91	0.029m	58	20

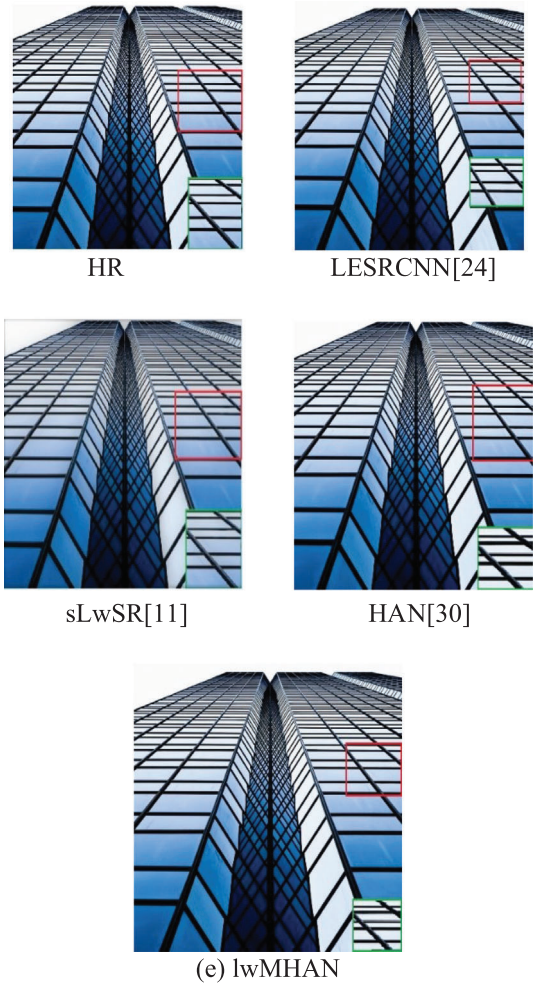

FIGURE 8 | Visual results of proposed model On Urban100 × 2 dataset.

TABLE 8 | Comparative analysis based on PSNR, SSIM, and parameters.

Methods	PSNR	SSIM	Parameters
swinIR	30.92	0.9151	897K
Shuffle-Mixer	30.65	0.909	411K
IFIN	28.57	0.860	980K
lwMHAN	28	0.91	29K

an ideal choice for real-world applications where computational resources are limited.

5 | Limitations and Future Work

Although the suggested lightweight multi-stage holistic attention-based network shows significant gains in image quality with reduced computational costs, there are still certain restrictions. The performance of the model has mostly been assessed using benchmark datasets, which might not accurately represent the variety and complexity of real-world situations. The network may encounter difficulties when handling excessive ambient noise or reconstructing highly textured regions with fine-grained details. Another drawback is the fixed pyramid structure, which can make it more difficult to adapt to changing input resolutions or shifting computing needs. To verify the model's resilience in a variety of scenarios, future studies will concentrate on resolving these problems through comprehensive real-world testing. Reconstruction quality and efficiency could be further increased by adding features like adjustable pyramid levels, noise-aware modules, and sophisticated generative approaches.

FIGURE 9 | Visual results of proposed model on Manga109 × 2 dataset.

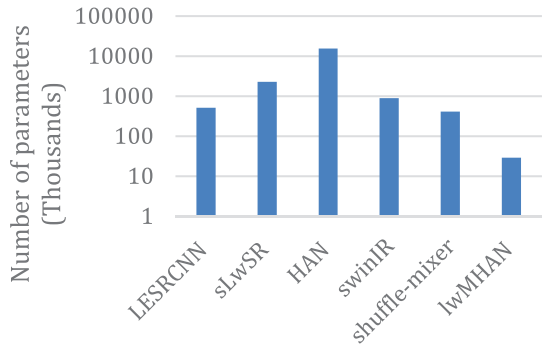

FIGURE 10 | Comparison with state of the art models.

6 | Conclusion

Single image super-resolution is a valuable sub-domain of image restoration that is required for numerous practical applications. It is a difficult task because numerous environmental and technical factors affect the image quality. Numerous super-resolution approaches based on deep learning have attempted to address this issue. However, they incur additional computational costs, and to reduce these costs, compression units are sometimes employed, which can negatively impact output quality. A multi-stage holistic attention-based model is proposed in this study. The model receives a low-resolution image as input and returns a super-resolved one as output. After taking the input, the model decomposes the input image into five pyramid levels using the Laplacian decomposition strategy. A sub-network is used for each pyramid level which is trained with its own loss function. Each sub-network has a similar structure and different numbers of kernels. As a sub-network layer attention module from holistic attention network HAN [11] with different parameter settings is used. To get a better generalization capability of the proposed method, extensive experiments are done on the DIV2K dataset and results are compared with other state-of-the-art methods. Experimental analysis of the proposed work demonstrates that using recursive blocks for parameter sharing reduced the total number of parameters which is equal to 29K only which is a smaller number as compared to other SISR schemes. Moreover, the super-resolution images restored from our model are of good quality and achieved a PSNR score of 28 and an SSIM index of 0.91.

Author Contributions

All authors contributed equally to accomplish this study. In addition, all authors read and approved the final manuscript.

Acknowledgements

Open access publishing facilitated by James Cook University, as part of the Wiley - James Cook University agreement via the Council of Australian University Librarians.

Ethics Statement

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data will be available upon request to the corresponding author.

References

1. Z. Wang, J. Chen, and S. C. H. Hoi, "Deep Learning for Image Super-Resolution: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, no. 10 (2020): 3365–3387, https://doi.org/10.1109/tpami.2020.2982166.

2. K. Li, S. Yang, R. Dong, X. Wang, and J. Huang, "Survey of Single Image Super-resolution Reconstruction," *IET Image Process* 14, no. 11 (2020): 2273–2290, https://doi.org/10.1049/iet-ipr.2019.1438.

3. W. Wang, Y. Hu, Y. Luo, and T. Zhang, "Brief Survey of Single Image Super-Resolution Reconstruction Based on Deep Learning Approaches," *Sensing and Imaging* 21, no. 1 (2020): 1–20, https://doi.org/10.1007/s11220-020-00285-4.

4. W. Yang, X. Zhang, Y. Tian, W. Wang, J. H. Xue, and Q. Liao, "Deep Learning for Single Image Super-Resolution: A Brief Review," *IEEE Transactions on Multimedia* 21, no. 12 (2019): 3106–3121, https://doi.org/10.1109/TMM.2019.2919431.

5. H. Zhu, C. Xie, Y. Fei, and H. Tao, "Attention Mechanisms in CNN-Based Single Image Super-resolution: A Brief Review and a New Perspective," *Electronics* 10, no. 10 (2021): 1187, https://doi.org/10.3390/electronics10101187.

6. C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, no. 2 (2016): 295–307, https://doi.org/10.1109/TPAMI.2015.2439281.

7. C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, and A. Acosta, "Photo-Realistic Single Image Super-resolution Using a Generative Adversarial Network," in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE, 2017), 105–114. https://doi.org/10.1109/CVPR.2017.19.

8. I. Guerrero-Ros and R. Valdor, "RCAN," in *Encyclopedia of Signaling Molecules*, ed. S. Choi (Springer, 2018), 4537–4546, https://doi.org/10.1007/978-3-319-67199-4_101514.

9. Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual Dense Network for Image Super-Resolution," in Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE, 2018), 2472–2481. https://doi.org/10.1109/CVPR.2018.00262.

10. J. Lyn and S. Yan, "Non-Local Second-Order Attention Network for Single Image Super Resolution," in *Machine Learning and Knowledge Extraction. CD-MAKE 2020. Lecture Notes in Computer Science*, Vol. 12279, eds. A. Holzinger, P. Kieseberg, A. Tjoa, E. Weippl (Springer, 2020).

11. B. Niu, et al., "Single Image Super-Resolution via a Holistic Attention Network," preprint, arXiv:2008.08767, August 20, 2020, https://arxiv.org/pdf/2008.08767.pdf.

12. G. E. Georgescu, I. I. Georgescu, and R. A. Tudoran, "Multimodal Multi-Head Convolutional Attention with Various Kernel Sizes for Medical Image Super-Resolution," *Sensors* 22, no. 6 (2022): 1–15, https://doi.org/10.3390/s22062075.

13. M. Usman, A. Mahmood, A. A. Abbasi, and O.-Y. Song, "Predictive Analytics Framework for Accurate Estimation of Child Mortality Rates for Internet of Things Enabled Smart Healthcare Systems," *International Journal of Distributed Sensor Networks* 16, no. 5 (2020).

14. M. I. Satti, J. Ahmed, and H. S. M. Muslim, "Ontology-Based News Linking for Semantic Temporal Queries," *Computers, Materials and Continua* 74, no. 2 (2023): 3913–3929.

15. A. Nadeem, M. Naveed, M. I. Satti, H. Afzal, T. Ahmad, and K.-I. Kim, "Depression Detection Based on Hybrid Deep Learning SSCL Framework Using Self-Attention Mechanism: An Application to Social Networking Data," *Sensors* 22, no. 24 (2022): 9775.

16. Y. Xiao, Q. Yuan, K. Jiang, J. He, C. W. Lin, and L. Zhang, "EDiffSR: An Efficient Diffusion Probabilistic Model for Remote Sensing Image Super-Resolution," *IEEE Transactions on Image Processing* 33 (2024): 1–12, https://doi.org/10.1109/TIP.2023.3349004.

17. K. Jiang, Z. Wang, P. Yi, and J. Jiang, "Dual-Path Deep Fusion Network for Face Image Hallucination," *Pattern Recognition* 107 (2020): 107475, https://doi.org/10.1016/j.patcog.2020.107475.

18. Q. Yuan, Z. Li, and H. Liu, "Recurrent Structure Attention Guidance for Depth Super-Resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, no. 8 (2022): 5004–5016, https://doi.org/10.1109/TPAMI.2021.3078354.

19. L. Cheng, Z. Wang, and Q. Li, "Image Super-resolution Using Progressive Generative Adversarial Networks for Medical Image Analysis," *IEEE Transactions on Medical Imaging* 39, no. 5 (2020): 1504–1515.

20. J. Xu, H. Li, Y. Zhang, and X. Wang, "Satellite Video Super-resolution via Multiscale Deformable Convolution Alignment and Temporal Grouping Projection," *IEEE Transactions on Geoscience and Remote Sensing* 59, no. 7 (2021): 5678–5692.

21. Y. Liu, C. Ma, and J. Zhang, "Dual-path Deep Fusion Network for Face Image Hallucination," *Pattern Recognition* 112 (2021): 107745.

22. J. He, Q. Yuan, J. Li, Y. Xiao, and L. Zhang, "A Self-supervised Remote Sensing Image Fusion Framework With Dual-stage Self-learning and Spectral Super-resolution Injection," *Isprs Journal of Photogrammetry and Remote Sensing* 204 (2023): 131–144.

23. Q. Ma, J. Jiang, X. Liu, and J. Ma, "Multi-task Interaction Learning for Spatiospectral Image Super-resolution," *IEEE Transactions on Image Processing* 31 (2022): 2950–2961.

24. J. Li, et al., "Enhanced Deep Image Prior for Unsupervised Hyperspectral Image Super-resolution," *IEEE Transactions on Geoscience and Remote Sensing (Early View)*: https://doi.org/10.1109/TGRS.2025.3531646.

25. J. Li, K. Zheng, L. Gao, L. Ni, M. Huang, and J. Chanussot, "Model-informed Multi-stage Unsupervised Network for Hyperspectral Image Super-resolution," *IEEE Transactions on Geoscience and Remote Sensing* 64 (2022): 5516117.

26. J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image Restoration Using swin Transformer," in Proceedings of the IEEE/CVF International Conference on Computer Vision (IEEE, 2021), 1833–1844.

27. X. Zhang, H. Song, K. Zhang, J. Qiao, and Q. Liu, "Single Image Super-resolution With Enhanced Laplacian Pyramid Network via Conditional Generative Adversarial Learning," *Neurocomputing* 398 (2020): 531–538.

28. B. Li, B. Wang, J. Liu, Z. Qi, and Y. Shi, "S-LWSR: Super Lightweight Super-Resolution Network," *IEEE Transactions on Image Processing* 29 (2020): 8368–8380, https://doi.org/10.1109/TIP.2020.3014953.

29. L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual Super-resolution Learning for Semantic Segmentation," in Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE, 2020), 3773–3782, https://doi.org/10.1109/CVPR42600.2020.00383.

30. 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 Transactions on Cybernetics* 51, no. 3 (2021): 1443–1453, https://doi.org/10.1109/TCYB.2020.2970104.

31. C. Yu, et al., "Lite-HRNet : A Lightweight High-Resolution Network," in Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE, 2021), 10440–10450.

32. B. Fu, Y. Li, X. hai Wang, and Y. gong Ren, "Image Super-Resolution Using TV Priori Guided Convolutional Network," *Pattern Recognition Letters* 125 (2019): 780–784, https://doi.org/10.1016/j.patrec.2019.06.022.

33. Z. Zhang, X. Chen, Z. Zhang, et al., "Single Image Super Resolution via Neighbor Reconstruction," *Pattern Recognition Letters* 125 (2019): 157–165, https://doi.org/10.1016/j.patrec.2019.04.021.

34. B. Li, Y. Zhou, Y. Zhang, and A. Wang, "Depth Image Super-resolution Based on Joint Sparse Coding," *Pattern Recognition Letters* 130 (2020): 21–29, https://doi.org/10.1016/j.patrec.2018.07.023.

35. S. Waqas, Z. Aditya, A. Salman, and K. Munawar, "Multi-Stage Progressive Image Restoration Number of parameters (Millions)," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE, 2021), 14821–14831.

36. C. Tian, Y. Xu, W. Zuo, B. Zhang, L. Fei, and C. W. Lin, "Coarse-to-Fine CNN for Image Super-Resolution," *IEEE Transactions on Multimedia* 23 (2021): 1489–1502, https://doi.org/10.1109/TMM.2020.2999182.

37. M. Y. Abbass, K. C. Kwon, M. S. Alam, Y. L. Piao, K. Y. Lee, and N. Kim, "Image Super Resolution Based on Residual Dense CNN and Guided Filters," *Multimedia Tools and Applications* 80, no. 4 (2021): 5403–5421, https://doi.org/10.1007/s11042-020-09824-3.

38. J. Liu, J. Ge, Y. Xue, W. He, Q. Sun, and S. Li, "Multi-Scale Skip-Connection Network for Image Super-Resolution," *Multimedia Systems* 27, no. 4 (2021): 821–836, https://doi.org/10.1007/s00530-020-00712-2.

39. X. Fu, B. Liang, Y. Huang, X. Ding, and J. Paisley, "Lightweight Pyramid Networks for Image Deraining," *IEEE Transactions on Neural Networks and Learning Systems* 31, no. 6 (2020): 1794–1807, https://doi.org/10.1109/TNNLS.2019.2926481.

40. S. Anwar and N. Barnes, "Densely Residual Laplacian Super-Resolution," *IEEE Transactions on Pattern Analysis and Machine Intel-*

ligence, 44, no. 3, (2020): 1192–1204, https://doi.org/10.1109/tpami.2020.3021088.

41. X. Zhang, H. Song, K. Zhang, J. Qiao, and Q. Liu, "Single Image Super-Resolution With Enhanced Laplacian Pyramid Network via Conditional Generative Adversarial Learning," *Neurocomputing* 398 (2020): 531–538, https://doi.org/10.1016/j.neucom.2019.04.097.

42. W. Shi, H. Du, W. Mei, and Z. Ma, "(SARN)spatial-wise Attention Residual Network for Image Super-resolution," *Visual Computer* 37, no. 6 (2021): 1569–1580, https://doi.org/10.1007/s00371-020-01903-8.

43. Y. Zhu, C. Geiß, and E. So, "Image Super-Resolution With Dense-Sampling Residual Channel-Spatial Attention Networks for Multi-Temporal Remote Sensing Image Classification," *International Journal of Applied Earth Observation and Geoinformation* 104 (2021), https://doi.org/10.1016/j.jag.2021.102543.

44. J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, "Residual Feature Aggregation Network for Image Super-Resolution," in Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE, 2020), 2356–2365. https://doi.org/10.1109/CVPR42600.2020.00243.

45. W. Dong, T. Zhang, J. Qu, S. Xiao, J. Liang, and Y. Li, "Laplacian Pyramid Dense Network for Hyperspectral Pansharpening," *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022): 1–13, https://doi.org/10.1109/TGRS.2021.3076768.

46. H. Wang, C. Wu, J. Chi, X. Yu, Q. Hu, and H. Wu, "Image Super-Resolution Using Multi-Granularity Perceptioand Pyramid Attention Networks," *Neurocomputing* 443 (2021): 247–261, https://doi.org/10.1016/j.neucom.2021.03.010.

47. J. Pan, Y. Liu, D. Sun, et al., "Image Formation Model Guided Deep Image Super-resolution," *Proceedings of the AAAI Conference on Artificial Intelligence* 34, no. 07 (2020): 11807–11814.

48. S. W. Zamir, A. Arora, S. Khan, et al, "Multi-Stage Progressive Image Restoration," in Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, (IEEE, 2021), 14821–14831.

49. Y. Zhou, Z. Li, C. L. Guo, S. Bai, M. M. Cheng, and Q. Hou, "Srformer: Permuted Self-attention for Single Image Super-resolution," in Proceedings of the IEEE/CVF International Conference on Computer Vision (IEEE, 2023), 12780–12791.

50. P. Behjati, P. Rodriguez, C. Fernández, I. Hupont, A. Mehri, and J. Gonzàlez, "Single Image Super-resolution Based on Directional Variance Attention Network," *Pattern Recognition* 133 (2023): 108997.

51. Y. Xiao, Q. Yuan, K. Jiang, J. He, C. W. Lin, and L. Zhang, "TTST: A Top-k Token Selective Transformer for Remote Sensing Image Super-resolution," *IEEE Transactions on Image Processing* 33 (2024): 738–752.

52. K. Jiang, Z. Wang, P. Yi, and J. Jiang, "Hierarchical Dense Recursive Network for Image Super-resolution," *Pattern Recognition* 107 (2020): 107475.

53. Z. Guo, Y. Xiao, W. Liao, P. Veelaert, and W. Philips, "FLOPs-Efficient Filter Pruning via Transfer Scale for Neural Network Acceleration," *Journal of Computational Science* 55, (2021): 101459, https://doi.org/10.1016/j.jocs.2021.101459.

54. D. R. I. M. Setiadi, "PSNR vs SSIM: Imperceptibility Quality Assessment for Image Steganography," *Multimedia Tools and Applications* 80, no. 6 (2021): 8423–8444, https://doi.org/10.1007/s11042-020-10035-z.

55. C. Tian, R. Zhuge, Z. Wu, et al., "Lightweight Image Super-Resolution With Enhanced CNN," *Knowledge-Based Systems* 205 (2020): 106235, https://doi.org/10.1016/j.knosys.2020.106235.

56. L. Sun, J. Pan, and J. Tang, "Shufflemixer: An Efficient Convnet for Image Super-resolution," *Advances in Neural Information Processing Systems* 35 (2022): 17314–17326.

57. W. Li, X. Li, T. Wei, J. Peng, and R. Chen, "Lightweight Interactive Feature Inference Network for Single-image Super-resolution," *Scientific Reports* 14, no. 1 (2024): 11601.