Electronic Research Archive

ERA, 34(7): 4387–4409.
DOI: 10.3934/era.2026194
Received: 07 April 2026
Revised: 24 April 2026
Accepted: 27 April 2026
Published: 20 May 2026

https://www.aimspress.com/journal/ERA

Research article

An infrared image super-resolution network fusing convolution and attention mechanisms

Sihang Luo, Yong Gan* and Xuan Wang

School of Computer Science and Artificial Intelligence, Zhengzhou University of Light Industry, Zhengzhou 450001, China

* **Correspondence:** Email: ganyong@zzuli.edu.cn.

Abstract: Infrared imaging technology plays an indispensable role in critical applications such as military surveillance, autonomous driving, and medical diagnostics. However, its inherent low-resolution and low-contrast characteristics often limit operational performance. While deep learning-based super-resolution (SR) techniques offer a software-driven solution, models face severe feature redundancy caused by simply stacking deep layers, and a lack of discriminative power in distinguishing critical textures from thermal noise. To address these issues, we proposed a novel Convolutional and Attention-based Super-Resolution Network (CASRNet). The novelty of our model lies in the synergistic fusion of a channel splitting (CS) strategy and a dual attention mechanism. First, the CS strategy decomposes feature maps into parallel streams, extracting diverse and less redundant representations. Second, a novel channel and spatial attention residual block (CSA_ResBlock) was designed to adaptively focus on informative feature channels and critical spatial boundaries. Quantitatively, CASRNet achieved superior performance on public benchmarks. Specifically, for the FLIR dataset (× 2), our model achieved a peak signal-to-noise ratio (PSNR) of 39.73 dB and structural similarity (SSIM) of 0.9639, outperforming the state-of-the-art infrared-specific model TherISuRNet by 0.48 dB and standard models like VDSR by 0.15 dB. Similar robust improvements (e.g., an exceptional 40.45 dB PSNR on the ThermalTau2 dataset) demonstrated the general applicability and high fidelity of CASRNet for real-world infrared enhancement tasks.

Keywords: infrared image; super-resolution; convolutional neural networks; attention mechanism

1. Introduction

Infrared imaging technology generates images by capturing the infrared radiation emitted by objects; this unique capability enables it to operate effectively under conditions where visible light is restricted, such as complete darkness, smoke, or adverse weather. Owing to these all-weather and non-contact characteristics, infrared imaging has become an indispensable tool across critical fields, with applications spanning pedestrian detection in autonomous driving, long-distance intrusion early warning in security monitoring, non-destructive testing of equipment in industrial sectors [1–3], and early lesion screening in medical diagnosis [4]. Moreover, the application of state-of-the-art (SOTA) deep learning techniques has revolutionized complex domains, demonstrating immense potential in extracting highly abstract representations. For instance, advanced machine learning architectures have shown profound usefulness in optimizing complex fundamental neural networks [5], advancing engineering informatics and non-destructive monitoring [6], and facilitating precise computer-based medical biology diagnostics [7]. Inspired by the remarkable success of SOTA deep learning approaches across these critical fields, we propose CASRNet to enhance infrared images, as the success of these application scenarios depends largely on the quality and the richness of details in the acquired images.

However, a long-standing technical bottleneck has constrained the full potential of infrared imaging technology. Compared with mature visible-light cameras, the spatial resolution of infrared imaging sensors is typically lower. This inherent resolution deficiency stems from factors such as the complexity of infrared detector manufacturing processes, high costs, and the physical diffraction limit of long-wave infrared radiation. Low-resolution images lead to the blurring or loss of critical details; for example, difficulty in distinguishing small distant objects, the inability to precisely identify minute cracks on industrial equipment, or missing subtle temperature anomalies in medical imaging, thereby severely affecting the accuracy of subsequent analysis and decision-making.

To overcome this hardware limitation, software-based image super-resolution (SR) technology [8] has emerged, aiming to reconstruct high-resolution (HR) corresponding images from one or more low-resolution (LR) images through algorithms. Furthermore, deep learning methods represented by convolutional neural networks (CNN) [9,10], by virtue of their powerful nonlinear feature learning capabilities, have transformed the field of image super-resolution. These methods are capable of automatically and end-to-end learning the extremely complex mapping relationships between LR and HR images [11,12], with performance that far exceeds traditional interpolation-based [13] or frequency domain-based [14] methods.

Despite the tremendous success of deep learning-based image super-resolution models, network architectures generally face two core challenges. The first is the issue of feature

redundancy. To pursue superior performance, model designers tend to construct very deep networks. However, simply stacking convolutional layers leads to the network extracting a large amount of highly correlated or even repetitive features in the deeper layers. Such redundancy not only causes a waste of parameters and computational resources but, more importantly, potentially obstructs the effective transmission of core information crucial for image reconstruction. The second challenge is the lack of feature discriminative power.

To simultaneously address the aforementioned two challenges, we propose a novel network architecture tailored for infrared image super-resolution; the channel splitting and Attention Synergistic Infrared Image Super-Resolution Network (CASRNet). The core concept of our network design lies in the synergistic fusion of two powerful strategies. To promote feature diversity and suppress redundancy, we introduce a channel splitting mechanism. This mechanism decomposes feature maps into multiple parallel processing streams along the channel dimension, guiding different branches of the network to learn complementary and differentiated feature representations. To achieve intelligent screening and enhancement of features, we design a novel fundamental building block; the channel and spatial attention residual block (CSA_ResBlock). This module endows the network with the capability of dynamic focusing: It first answers what features are significant through the channel attention mechanism, and then locates where the features are significant through the spatial attention mechanism, thereby adaptively enhancing the most valuable information for image reconstruction.

The major contributions of this paper can be summarized as follows:

(1) We propose CASRNet, a novel architecture that synergistically integrates a channel splitting strategy and a dual attention mechanism to resolve the challenges of feature redundancy and noise in infrared images.

(2) We design the channel splitting and attention residual block (CARB), which significantly enhances the network's discriminative power.

(3) Comprehensive evaluations on public benchmarks demonstrate the quantitative superiority of our model. Compared to state-of-the-art models, CASRNet achieves a significant performance margin. For example, it outperforms the specialized TherISuRNet by 0.85 dB in PSNR on the complex ThermalTau2 (× 2) dataset while maintaining competitive computational efficiency.

2. Related works

Before delving into SR techniques, it is crucial to recognize the broader landscape of modern image analysis. Advancements have heavily focused on designing robust, adaptive architectures for complex visual tasks. For instance, sophisticated computational frameworks have been successfully deployed for generalized image analysis, knowledge-based feature extraction, and complex pattern recognition [15,16]. Furthermore, cutting-edge studies in computer vision continually emphasize the importance of advanced spatial-channel correlations and attention paradigms to handle diverse and challenging imaging conditions [17,18]. This study is deeply connected to these broad approaches by inheriting their core philosophy of dynamic feature

weighting and robust representation learning; however, it explicitly distinguishes itself by tailoring a unique channel-splitting and attention synergy optimized to combat the distinct challenges of the infrared domain; namely, the severe lack of textures and pervasive high thermal noise.

2.1. Deep learning-based image super-resolution technology

Very Deep Super-Resolution (VDSR) [19] significantly improved performance by increasing the network depth to 20 layers and introducing the concept of Residual Learning [20]. Residual learning enables the network to learn only the residues between LR and HR images, which drastically reduces the training difficulty and has since become a standard configuration for subsequent deep SR models. On this basis, Enhanced Deep Super-Resolution (EDSR) [21] constructed a more concise and efficient residual module by removing unnecessary Batch Normalization (BN) layers from traditional residual blocks, successfully training even deeper networks and further refreshing performance benchmarks. Beyond pursuing higher reconstruction fidelity, some researchers have focused on enhancing the perceptual quality of generated images. SRGAN [22] innovatively introduced Generative Adversarial Networks (GAN), generating images with more natural textures that better align with human visual perception through adversarial and perceptual losses. To intuitively demonstrate the differences in core technical architectures between the general super-resolution models and the CASRNet proposed in this study, a comparison with classical methods is presented in Table 1.

Table 1. Comparison of key technologies in general super-resolution networks.

Method	Residual block	Channel splitting	Attention mechanism
SRCNN	✗	✗	✗
VDSR	✓	✗	✗
EDSR	✓	✗	✗
CASRNet (ours)	✓	✓	✓

The advantage of SRCNN lies in its simple structure and low computational overhead, which validate the effectiveness of deep learning in the field of super-resolution. However, its deficiency is that the network layers are too shallow, and the receptive field is restricted, making it difficult to extract complex high-frequency texture details; as a result, the reconstruction effect tends to be overly smooth.

The advantage of VDSR/EDSR is the successful training of extremely deep networks by introducing residual learning, which significantly enhances the peak signal-to-noise ratio (PSNR) of image reconstruction. Their deficiency lies in the primary reliance on the simple stacking of convolutional layers to improve performance. As the depth increases, the features extracted in the deeper layers of the network often exhibit high redundancy, leading to a waste of computational resources and a lack of discriminative power regarding the significance of different features.

The improvement of CASRNet (Ours) lies in the introduction of the channel splitting technique while retaining the advantages of residual learning. This design not only addresses the

feature redundancy issue in deep networks but also endows the network with the capability of feature selection through the attention mechanism, making it more efficient than VDSR and EDSR.

2.2. Infrared image super-resolution

Although general SR models can be applied to infrared images, significant differences exist in the statistical characteristics between infrared and visible-light images. Infrared images typically exhibit a lower signal-to-noise ratio (SNR) and fewer texture details; furthermore, their pixel values directly reflect temperature distributions rather than light reflections. These characteristics render the direct migration of models trained on visible-light data ineffective.

Consequently, researchers have begun to explore super-resolution networks designed for infrared images. Early work, such as the Thermal Enhancement Network (TEN) [23], attempted to apply shallow network structures similar to SRCNN [24] to infrared image enhancement. As research progressed, more complex specialized models were proposed. For instance, TherISuRNet [25], a model that performed excellently in thermal image SR challenges, adopted an asymmetric progressive learning strategy to conduct feature extraction and reconstruction in a more efficient manner [26,27]. These works demonstrate that specialized network design tailored to the characteristics of infrared images is necessary and effective. However, most infrared image SR models rely on standard convolutional operations, leaving room for improvement in feature discrimination and redundant information processing. Although specialized infrared models have made certain progress in adapting to infrared characteristics, significant differences remain in their feature processing mechanisms compared to the method proposed in this paper, as detailed in the comparison in Table 2.

Table 2. Comparison of key technologies in specialized infrared image super-resolution networks.

Method	Residual block	Channel splitting	Attention mechanism
TherISuRNet	✓	✗	✗
CASRNet (ours)	✓	✓	✓

The advantage of TherISuRNet lies in its specialized design for infrared images and its adoption of an asymmetric architecture, which yields better performance on infrared datasets compared to general-purpose models. However, its deficiency is that it adheres to traditional convolutional feature extraction methods. Since infrared images are typically characterized by low contrast and blurriness, standard convolutions struggle to effectively separate edges from the background without introducing additional noise.

The improvement of CASRNet is the adoption of channel splitting technology as its core driver. By processing feature maps through splitting streams, the network can mine more diverse and complementary feature representations under low-contrast conditions. Furthermore, the integrated attention mechanism can effectively suppress the background noise commonly found in infrared images, a capability that TherISuRNet lacks.

Furthermore, it is worth noting that beyond single-image super-resolution, the enhancement of infrared representations is also heavily investigated in multi-modal contexts. Significant parallel contributions have been made in effectively fusing infrared thermal features with other modalities to overcome inherent resolution limits. For instance, VIF-Net [28] pioneered a deep framework to fuse visible and infrared images, thereby enhancing target highlighting and background details. Moreover, advanced attention and SAM-based paradigms, such as HyPSAM [29], have demonstrated cutting-edge capabilities in processing complex spectral and spatial information. While we focus on single-image SR, these sophisticated multi-modal fusion and attention mechanisms provide valuable insights for the future evolution of infrared enhancement networks.

2.3. Attention mechanisms in deep vision models

Attention mechanisms [30] mimic the human visual system, enabling neural networks to adaptively concentrate computational resources on the most significant parts of the input data. This concept has achieved tremendous success across the field of computer vision.

Channel attention [31] aims to model the dependencies between feature channels, enabling the network to focus on "what" features are important. Its representative work is the Squeeze-and-Excitation Network (SE-Net) [32]. SE-Net utilizes a Squeeze operation to obtain the global information of each channel and then employs an Excitation operation to learn the importance weights for each channel. Finally, these weights are multiplied back into the original feature maps, thereby achieving the enhancement of critical feature channels. Residual channel attention Network (RCAN) [33] successfully applied the channel attention mechanism to the image super-resolution task, demonstrating its powerful efficacy in image reconstruction.

Spatial attention [34], on the other hand, focuses on "where" the features are important, aiming to generate a spatial weight map to represent the significance of different spatial locations. The Convolutional Block Attention Module (CBAM) [35] further combines channel and spatial attention to perform dual refinement of feature maps through serial or parallel configurations, achieving more comprehensive feature selection. To explore the applicability of attention mechanisms in infrared super-resolution tasks, a comparison between mainstream attention modules and the module proposed in this paper is presented in Table 3.

Table 3. Comparison of key technologies in attention mechanisms.

Method	Residual block	Channel splitting	Attention mechanism
SE-Net	✓	✗	✓ (Channel-only)
CBAM	✓	✗	✓ (Channel & Spatial)
CASRNet (ours)	✓	✓	✓ (Channel & Spatial w/ Splitting)

The advantage of SE-Net lies in its ability to explicitly model the dependencies between feature channels and enhance the weights of informative channels. Its deficiency is that it focuses only on the channel dimension while ignoring the spatial dimension. For infrared images with missing texture details, it fails to locate "where" the critical edges are, resulting in insufficient detail restoration.

The advantage of CBAM lies in its simultaneous integration of channel and spatial attention, which theoretically yields stronger feature extraction capabilities. Its deficiency is that, as a general-purpose plug-and-play module, it tends to misidentify high-frequency noise as texture details when directly applied to infrared images. Qualitative experiments indicate that the denoising performance of CBAM in infrared SR tasks is suboptimal.

The improvement of CASRNet (Ours) lies in the fact that it does not simply stack attention modules; instead, it embeds a dual attention mechanism within a channel splitting architecture. This synergistic design enables the network to leverage the attention mechanism for edge enhancement while utilizing the splitting branches to filter out redundant information and noise. Consequently, it achieves superior visual reconstruction results compared to SE-Net and CBAM.

2.4. Motivation of this study

Reviewing the literature, several critical observations can be synthesized:

•Although deep super-resolution networks are powerful, they universally suffer from feature redundancy. Simple increases in network depth often encounter performance bottlenecks, where the marginal gains in accuracy are offset by excessive computational costs.

•While specialized infrared image SR (IRSR) models have emerged, there remains significant room for exploration in leveraging advanced feature selection mechanisms. Most IRSR methods rely on traditional convolution, which lacks the discriminative power required for the statistical properties of thermal data.

•Attention mechanisms have demonstrated exceptional performance in feature discrimination; however, how to efficiently integrate these mechanisms into an architecture designed to mitigate redundancy remains a challenging and worthwhile research problem.

Consequently, the Convolutional and Attention-based Super-Resolution Network (CASRNet) proposed in this paper, which integrates convolution and attention mechanisms, is designed to fill the aforementioned research gaps. We employ a channel splitting strategy to proactively guide the network in learning diverse, low-redundancy features. Furthermore, we implement adaptive refinement of these features through our designed channel and spatial attention residual block (CSA_ResBlock). We believe that this synergistic design, combining feature diversification with feature focusing, can provide a significant performance breakthrough for infrared image super-resolution tasks.

3. Proposed method

Although convolutional neural networks (CNNs) have demonstrated outstanding performance in super-resolution tasks, as network depth increases, the issues of feature redundancy and insufficient discriminative power between feature channels become progressively prominent. This not only leads to a waste of computational resources but also hinders the network's ability to effectively utilize high-level information.

To address these challenges effectively, we have designed and proposed a novel network architecture named CASRNet. In this chapter, we will elaborate on its overall structure, core

innovative modules, and specific implementation details while summarizing the underlying design philosophy.

3.1. Overall network architecture

As illustrated in Figure 1, CASRNet is designed to establish a non-linear mapping between the LR thermal image ILR and its HR counterpart IHR. The network architecture primarily comprises three functional modules: the shallow feature extraction (SFE) module, the deep feature extraction (DFE) module, and the image reconstruction (IRec) module. Furthermore, a Global Residual Learning [36] strategy is employed to stabilize the training process and preserve the fundamental low-frequency information of the original image.

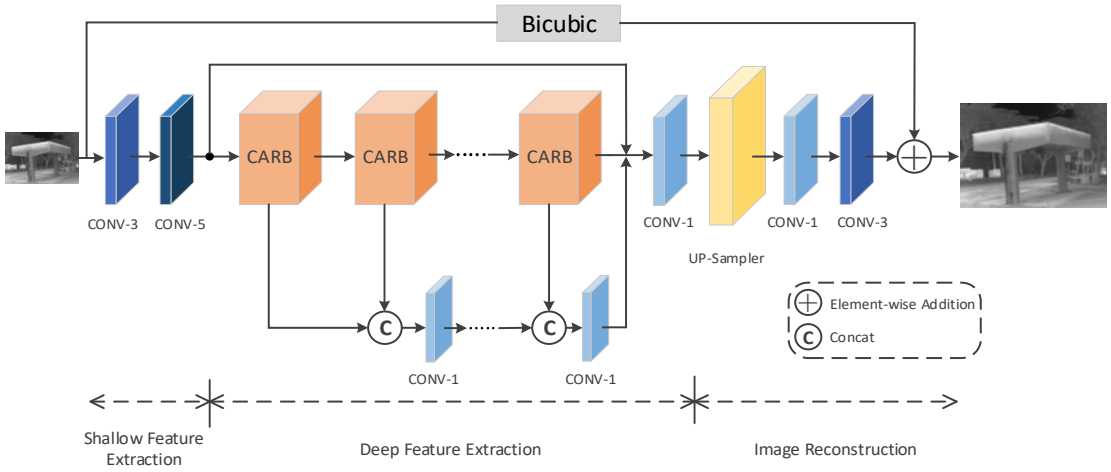

Figure 1. The overall architecture of the proposed CASRNet model. The network establishes a non-linear mapping from low-resolution (LR) to high-resolution (HR) space through three primary modules: (1) Shallow feature extraction (SFE) using multi-scale convolutions; (2) Deep feature extraction (DFE) built upon cascaded channel splitting and attention residual blocks (CARB) to mitigate redundancy; and (3) Image reconstruction utilizing sub-pixel convolution alongside a global residual connection.

3.1.1. Shallow feature extraction

We first perform shallow feature extraction on the original input LR image using two convolutional layers with kernel sizes of 3 × 3 and 5 × 5. The 3 × 3 convolutional layer is responsible for extracting fine-grained image details from the LR image, while the 5 × 5 convolutional layer handles the separation of local features with larger contours. By integrating these multi-scale responses, the initial shallow features are obtained. Mathematically, the shallow feature extraction can be expressed as:

$$F_0 = H_{SFE}(I_{LR}), \tag{1}$$

where $H_{SFE}(\cdot)$ denotes the shallow feature extraction operation.

3.1.2. Deep feature extraction

The extracted shallow features are subsequently fed into the deep feature extraction (DFE) module. As the core component of the entire network, the DFE consists of N cascaded channel splitting and attention residual blocks (CARB). Within the DFE stage, we introduce a channel splitting strategy to distinguish feature information across frequencies and to mitigate redundancy. The features processed by the deep mapping stage can be expressed as:

$$F_{DFE} = H_{DFE}(F_0), \tag{2}$$

where $H_{DFE}(\cdot)$ denotes the deep feature extraction operation. A detailed discussion regarding the internal structure of the DFE module will be provided in Section 3.2.

3.1.3. Image reconstruction

The final stage of the network is the image reconstruction module. We combine the deep features FDFE with the shallow features transmitted via the skip connection and restore the image to the target resolution through an upsampling operation. We utilize sub-pixel convolution (PixelShuffle) as the upsampler $H_{UP}(\cdot)$ due to its superior advantages in computational efficiency and reconstruction quality. The final super-resolved image is defined as:

$$I_{SR} = H_{Rec}(F_{DFE}) + H_{UP}(I_{LR}), \tag{3}$$

This global residual architecture enables the network to focus solely on learning the residual map (high-frequency details), which significantly reduces the difficulty of model convergence.

3.2. Channel splitting based DFE

In the backbone network, feature maps often contain a large amount of redundant or highly correlated information. To address this issue while enhancing the network's ability to learn feature patterns, we design the channel splitting (CS) block as the fundamental component of the CARB.

As illustrated in Figure 2 (CARB), the CARB consists of three densely connected CS blocks followed by a 1 × 1 core convolutional layer. The structural design of the CS block is shown in Figure 2 (CS), which comprises a 1 × 1 convolutional layer and a channel splitting operation. Specifically, regarding its implementation, this splitting mechanism utilizes a fixed, non-adaptive partitioning strategy. It divides the input feature maps into exactly two parallel splits (N = 2) along the channel dimension. It is important to distinguish this from standard grouped convolutions: Our CS explicitly routes these two fixed subsets of channels into separate, parallel sequences of residual blocks (RB) for highly differentiated feature learning. As for the recombination strategy,

the outputs from these two distinct paths are integrated via a straightforward concatenation operation along the channel axis at the end of the CS unit.

In Figure 2 (RB), we present the design of the RB, which consists of two 3 × 3 convolutional layers and a ReLU activation function, optimized for extracting local spatial features.

3.3. Channel and spatial attention residual block (CSA_ResBlock)

To empower the network with the capability to adaptively determine "which" features are significant and "where" they are located, we design the channel and spatial attention (CSA) module. This module integrates a focusing operation with a dual attention mechanism, aiming to accurately screen for critical high-frequency details within infrared images characterized by significant noise. As illustrated in Figure 2 (CSA), the CSA module consists of a 1 × 1 convolutional layer, a channel attention (CA) sub-module, and a spatial attention (SA) sub-module. Structurally, the dual attention mechanism is configured in a sequential manner (channel attention followed by spatial attention). Furthermore, to ensure gradient stability and facilitate the flow of low-frequency information, the sequential attention process is tightly integrated within a local residual learning framework, adding the refined features directly back to the original input mapping.

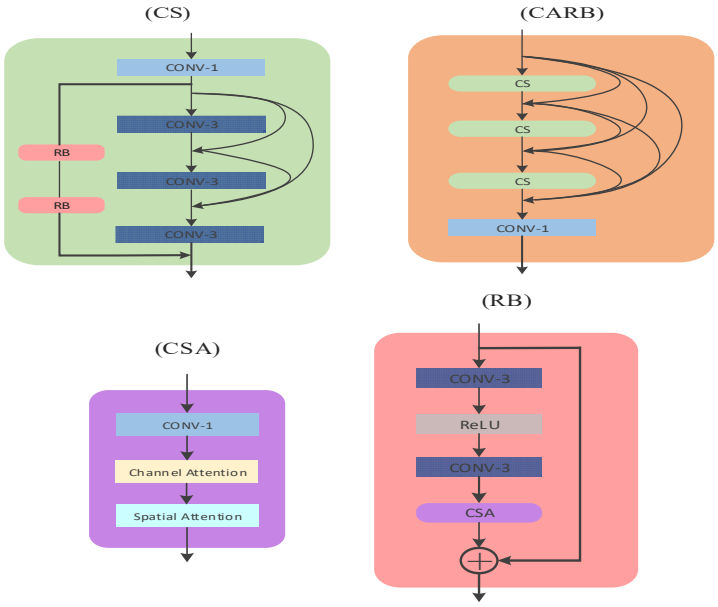

Figure 2. Architectural design of the proposed channel splitting and attention residual block (CARB). The module employs a uniform channel splitting (CS) strategy to create parallel processing streams via residual blocks (RB), followed by dense connections. The extracted features are then refined by a dual attention mechanism (CSA) that sequentially applies channel attention to select informative feature maps and spatial attention to localize critical texture boundaries.

3.3.1. Local feature extraction

Initially, the input features undergo a local feature transformation through a convolutional layer integrated with a ReLU activation function, thereby extracting the preliminary residual features F_{res}.

3.3.2. Channel attention (CA)

Infrared images typically suffer from low contrast, and channel attention facilitates the enhancement of feature channels containing critical textural information. We improve the conventional SE-Block [26] by simultaneously utilizing global average pooling (AvgPool) and global max pooling (MaxPool) to aggregate spatial information, thereby capturing richer channel statistical characteristics. The formula for calculating channel weights M_c is defined as:

$$M_c = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F_{res})) + \mathrm{MLP}(\mathrm{MaxPool}(F_{res})), \tag{4}$$

where σ denotes the sigmoid activation function and MLP represents the multi-layer perceptron. The weighted feature is obtained as follows: $F'_{res} = M_c \otimes F_{res}$.

3.3.3. Spatial attention (SA)

To further localize critical regions within the image (e.g., object boundaries and significant thermal gradient areas), we introduce spatial attention following the CA module. By compressing the feature maps along the channel dimension and employing a large-kernel convolution (7 × 7), a spatial attention map M_s is generated:

$$M_s = \sigma(\mathrm{Conv}^{7\times7}([\mathrm{AvgPool}(F^{l}_{res}); \mathrm{MaxPool}(F^{l}_{res})]), \tag{5}$$

The final refined feature is expressed as: F"res = Ms⊗F'res.

Finally, by incorporating a local residual connection [30], the refined features are added to the input features to form the module's final output. This design not only preserves the original information but also enables the network to focus on learning high-frequency residual components.

3.4. Loss function

To optimize the network parameters, we employ the L1 norm (Mean Absolute Error, MAE) as the loss function. Compared to the L2 loss, the L1 loss demonstrates superior robustness to outliers in image restoration tasks and tends to generate sharper edges, which is critically important for inherently blurred infrared images. The loss function is defined as follows:

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} \| I_{SR}^{(i)} - I_{HR}^{(i)} \|_1, \tag{6}$$

where N denotes the batch size of training samples; I_{SR} and I_{HR} represent the generated super-resolution image and the corresponding ground-truth high-resolution image, respectively; and $\|.\|$ signifies the absolute value operation.

4. Experiments

In this section, the training environment and experimental configurations of the proposed network are detailed. First, we describe the training and testing datasets utilized in our study. Subsequently, ablation studies are conducted to verify the rationality and effectiveness of the proposed components. Furthermore, we compare our model with several SOTA methods, followed by comprehensive quantitative evaluations and qualitative assessments.

4.1. Implementation details

The proposed network is optimized using the Adam optimizer [37] with its default hyperparameter settings, incorporating L2 regularization (weight decay) to penalize large weights and prevent network over-parameterization. To fine-tune the optimization process, a multi-step learning rate scheduling strategy is adopted. The initial learning rate is set to 1e-4 and is halved every 20% of the total iterations. The training process consists of 200,000 iterations in total, with a mini-batch size of 1.

Regarding data preprocessing, image patches of size 128 × 128 are randomly cropped from the HR images, with corresponding patches extracted from the LR counterparts. To effectively mitigate potential overfitting and enhance the generalization capability of the model, extensive data augmentation techniques are applied during training. These include random horizontal and vertical flips, as well as random rotations (e.g., 90°, 180°, and 270°), which artificially expand the diversity of the training distributions. Furthermore, the dataset is shuffled during the loading process to mitigate potential training bias. The algorithm is implemented based on the PyTorch framework and executed on a server equipped with an NVIDIA RTX 4090 GPU, ensuring high processing speed and computational efficiency.

4.2. Datasets

The datasets employed in our experiments are the grayscale infrared datasets provided by Rivadeneira et al. [38]. Notably, only the FLIR (HR) images are utilized during the training phase, which consists of 951 and 50 high-quality grayscale infrared images for training and testing, respectively. For the training set, the 951 original images serve as the ground-truth HR dataset. Corresponding LR datasets are generated via bicubic downsampling with scaling factors of × 2 and × 4. To simulate real-world degradations, these LR images are further corrupted by additive white Gaussian noise (AWGN) with a mean of 0 and a standard deviation of 10. To evaluate the generalization capability of the proposed network, the testing phase encompasses not only the 50-image testing subset from FLIR (HR) but also the 101 ThermalTau2 dataset [39], which contains 101 grayscale infrared images captured across diverse scenarios.

4.3. Mathematical expressions

To demonstrate the effectiveness of the proposed method, we adopt two objective evaluation metrics: peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). PSNR is a widely utilized metric for assessing image or video quality, particularly in the fields of image processing and video coding. In this study, PSNR is employed to measure the degree of image distortion by quantifying the discrepancies between the original HR images and the reconstructed SR images. Its mathematical expression is defined as:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right), \tag{7}$$

where MAX denotes the maximum pixel intensity of the image, and MSE represents the Mean Squared Error, which quantifies the average degree of discrepancy between the two images. The calculation formula for MSE is expressed as follows:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - k(i,j)]^2, \tag{8}$$

where $I(i,j)$ and $k(i,j)$ represent the pixel values at position (i,j) of the HR image and the reconstructed SR image, respectively; m and n denote the height and width of the image. A higher PSNR value indicates superior image reconstruction quality.

SSIM is another pivotal metric for image quality assessment that incorporates the characteristics of the Human Visual System (HVS), thereby providing a better reflection of human perception. By comparing the similarity between two images across three dimensions, luminance, contrast, and structure, SSIM evaluates their overall resemblance. The mathematical expression of SSIM is defined as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{9}$$

where x and y denote the two images being compared; μ_x and μ_y are the average values of images x and y, respectively; σ_x^2 and σ_y^2 represent the variances of x and y; and σ_{xy} is the covariance between the two images. The terms c_1 and c_2 are constants introduced to maintain stability, typically defined as $(k_1L)^2$ and $(k_2L)^2$, where k_1 and k_2 are small constants, and L represents the maximum range of the grayscale levels. The value of the SSIM function ranges from [0, 1], where a higher value signifies lower image distortion and a greater degree of similarity between the two images.

4.4. Ablation study on proposed architectures

To rigorously validate the necessity and effectiveness of each proposed component, we perform a comprehensive ablation study comparing different internal model configurations. We investigate the network's sensitivity to the residual modules (RB), attention mechanisms (CSA), and the dense connection strategy. The SR results of these experiments are quantitatively compared in Table 4 using PSNR and SSIM metrics. We take the complete model, which incorporates all

design elements (i.e., CS, RB, and CSA modules), as the baseline. This integrated configuration achieves a PSNR of 39.73 dB and an SSIM of 0.9641.

First, to verify the significance of the residual block (RB), we train a variant of the network where the RB is removed from the CS structure. As observed in Table 4, this modification leads to a drastic decline in performance. Compared to the complete model, integrating the RB into the CS structure yields a substantial PSNR gain of + 0.75 dB, which thoroughly demonstrates the critical role of the RB as a fundamental building block of the network.

Furthermore, we investigate the contribution of the proposed CSA attention module. Upon incorporating the CSA into the RB, a performance improvement of + 0.27 dB in PSNR is observed. To further analyze the internal components of the CSA, we separately remove the channel attention (CA) and spatial attention (SA). Experimental results indicate that integrating CA and SA individually yields PSNR gains of + 0.18 dB and + 0.19 dB, respectively. This suggests that the CA and SA modules contribute positively to performance, with their combination achieving the maximum efficacy.

Finally, we evaluate the effectiveness of the dense connection strategy, which is utilized in our CARB and CS modules (as illustrated in Figure 2). As shown in Table 4, employing dense connections in the CARB and CS structures results in PSNR gains of + 0.17 dB and + 0.15 dB, respectively. These results validate the effectiveness of the design in enhancing model performance by promoting feature reuse and strengthening information flow.

Table 4: Average PSNR and SSIM values calculated for the FLIR testing dataset with a scaling factor of × 2 under identical training conditions.

Case	PSNR (dB)	SSIM
w/o dense connection in CARB	39.56	0.9593
w/o dense connection in CS	39.58	0.9616
w/o RB in CS	38.98	0.9339
w/o CSA in RB	39.46	0.9462
w/o CA in CSA	39.55	0.9587
w/o SA in CSA	39.54	0.9595
w/CS & RB	39.61	0.9514
w/CS & RB & CSA	**39.73**	**0.9641**

4.5. Quantitative comparison with baseline models

To establish the state-of-the-art status of CASRNet, we conduct extensive benchmark comparisons against a broad spectrum of baseline models. The rationale for this selection is to provide a comprehensive evaluation hierarchy: It includes traditional interpolation (Bicubic) to set the lower bound, classic deep SR models (SRCNN [24], VDSR [19], EDSR [23]) to benchmark against standard convolution-based network depths, and general attention networks (CBAM [35], SE-Net [32]) to demonstrate that simply plugging in generic attention is insufficient for thermal noise. Most importantly, we compare against the state-of-the-art infrared specialized model,

TherISuRNet [25], to directly validate the superiority of our proposed architecture in the specialized thermal domain. All benchmark models are rigorously evaluated under identical training setups on the FLIR [34] and ThermalTau2 [39] datasets, as presented in Table 5.

Table 5. Average PSNR (dB) and SSIM results obtained for the FLIR (50 images) and ThermalTau2 (101 images) test sets. The bold values indicate the best performance.

Dataset	Method	X2		X4	
		PSNR	SSIM	PSNR	SSIM
FLIR	Bicubic	35.54	0.9145	32.95	0.7988
	CBAM	38.87	0.9363	34.91	0.8997
	SE-Net	38.46	0.9436	34.80	0.8953
	THerISuRNet	39.25	0.9604	35.13	0.9028
	SRCNN	39.16	0.9587	34.89	0.9025
	EDSR	39.34	0.9618	35.38	0.9145
	VDSR	39.58	0.9596	35.34	0.9128
	Ours	**39.73**	**0.9639**	**35.45**	**0.9156**
Thermal	Bicubic	37.89	0.9059	35.20	0.8245
	CBAM	39.24	0.9342	35.67	0.8957
	SE-Net	39.13	0.9252	35.35	0.8910
	THerISuRNet	39.60	0.9438	36.49	0.9076
	SRCNN	39.56	0.9310	36.44	0.9018
	EDSR	40.18	0.9422	36.58	0.9118
	VDSR	40.27	0.9447	36.54	0.9125
	Ours	**40.45**	**0.9628**	**36.67**	**0.9134**

Table 5 provides a detailed breakdown of the test results for each method in terms of peak signal-to-noise ratio (PSNR) and Structural Similarity (SSIM). Based on the data presented in the table, several key observations and analyses can be derived:

(1) Overall performance superiority

For the FLIR dataset (× 2), our method achieves a PSNR of 39.73 dB, representing a substantial improvement of 4.19 dB over traditional Bicubic interpolation (35.54 dB). This significant margin demonstrates the overwhelming advantage of deep learning-based approaches in recovering fine details in infrared imagery. More importantly, CASRNet maintains a clear lead over the second-best performing models. Specifically, in the FLIR (× 2) test, CASRNet outperforms VDSR and EDSR by 0.15 dB and 0.39 dB, respectively. These results indicate that increasing network depth is not the sole path to performance enhancement. Instead, the introduced CS strategy more efficiently utilizes network capacity by effectively reducing feature redundancy.

(2) Comparison with attention-based networks

In the FLIR (× 4) task, CASRNet (35.45 dB) achieves a performance gain of 0.54 dB and 0.65 dB over CBAM (34.91 dB) and SE-Net (34.80 dB), respectively. These results suggest that simply integrating generic attention modules as "plug-in" components does not fully unleash their potential for infrared image super-resolution.

Through the designed CSA_ResBlock, CASRNet establishes a deep synergy between channel/spatial attention and the feature extraction process. This architectural integration enables the network to more precisely focus on the subtle edges and texture information inherent in infrared images, thereby achieving superior reconstruction performance.

(3) Comparison with infrared-specific networks

For the Thermal dataset (× 2), CASRNet (40.45 dB) outperforms TherISuRNet (39.60 dB) by a significant margin of 0.85 dB. This substantial gap indicates that although TherISuRNet accounts for infrared characteristics, the integration of channel splitting techniques with the dual attention mechanism in CASRNet provides superior feature discrimination and reconstruction robustness. This architectural synergy is particularly effective in addressing the inherent challenges of infrared imagery, such as blurring and low contrast, enabling the network to better distinguish fine thermal details and maintain stable performance across scenarios.

(4) Cross-dataset and cross-scale generalization capability

CASRNet consistently maintains its top-tier ranking for the FLIR dataset and the more scene-complex Thermal dataset. Notably, for the Thermal dataset, the PSNR exceeds the 40 dB threshold, which underscores the model's robust generalization capability across infrared sensors and environmental scenarios.

(5) Statistical stability and variance across runs

To justify the significance of our improvements and rule out stochastic variations during optimization, we monitor the statistical variance of our model across 5 independent training runs with different random seeds. For the FLIR (× 2) dataset, the network consistently converges with a mean PSNR of 39.73 dB and a standard deviation of only ± 0.03 dB. Similarly, the SSIM remains highly stable with a minimal variance of ± 0.0004. This statistically significant stability confirms that the performance gains over baseline models are robust and directly attributable to the proposed architectural innovations rather than random training noise.

In the × 4 super-resolution task, all models exhibit a performance decline due to the dramatic increase in information loss. Nevertheless, CASRNet sustains high PSNR scores of 35.45 dB and 36.67 dB for the FLIR (× 4) and Thermal (× 4) datasets, respectively. Furthermore, its SSIM metrics (0.9156 and 0.9134) remain significantly higher than those of competing models. These results indicate that even with extremely low-resolution inputs, our network effectively infers and recovers plausible high-frequency details through global residual learning and deep feature refinement.

4.6. Qualitative analysis

To more intuitively demonstrate the superiority of our network in terms of visual perceptual quality, we select representative scenes from the FLIR and Thermal datasets for visualization comparison. Figures 3 and 4 illustrate the visual effects of super-resolution reconstruction at × 2 and × 4 scales, respectively. To facilitate a clear observation of the detail recovery, key regions within the images (indicated by the red boxes) are displayed as magnified local views.

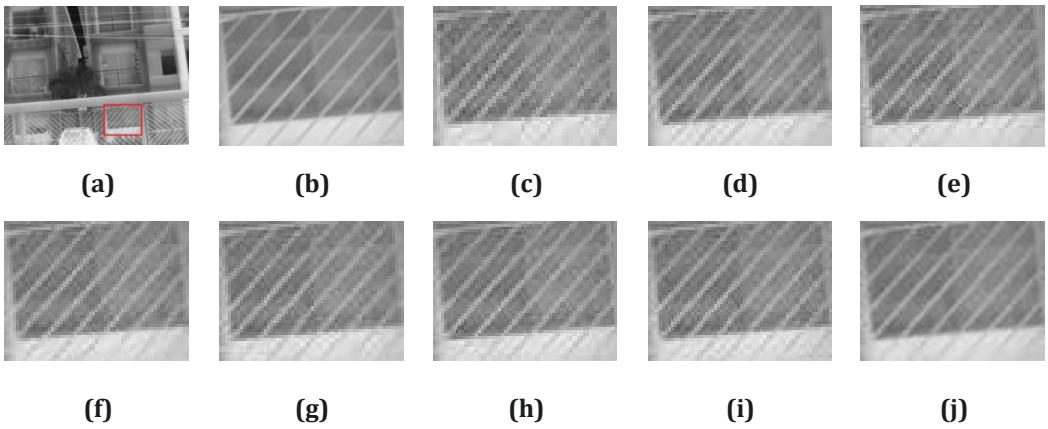

Figure 3. Qualitative results for the FLIR testing dataset with a scaling factor of × 2: (a) 0036.jpg from FLIR; (b) HR PSNR(dB)/SSIM; (c) Bicubic 30.78/0.9330; (d) CBAM 36.18/0.9704; (e) SE-Net 35.76/0.9798; (f) THerlSuRNet 37.13/0.9841; (g) SRCNN 36.98/0.9614; (h) EDSR 37.89/0.9874; (i) VDSR 37.45/0.9763; and (j) Ours 38.27/0.9885.

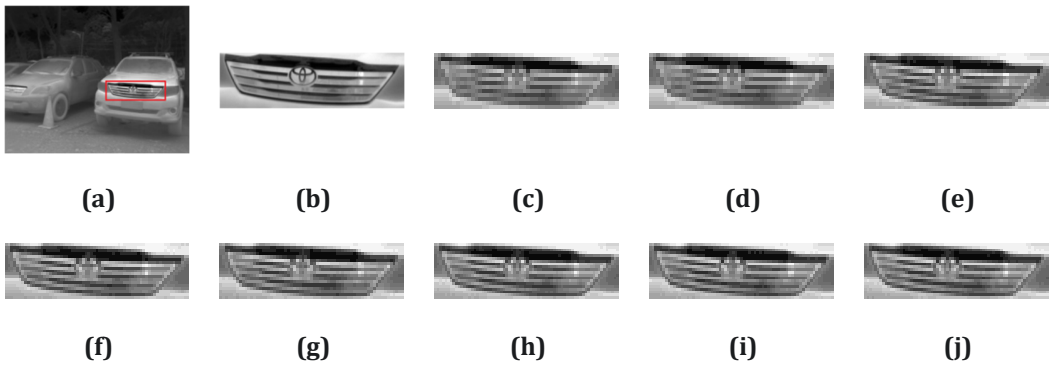

Figure 4. Qualitative results on the ThermalTau2 testing dataset with a scaling factor of × 4: (a) thermal_054.png from Thermal101; (b) HR PSNR(dB)/SSIM; (c) Bicubic 28.79/0.9218; (d) CBAM 33.41/0.9589; (e) SE-Net 31.80/0.9546; (f) THerlSuRNet 35.22/0.9627; (g) SRCNN 34.97/0.9608; (h) EDSR 35.43/0.9763; (i) VDSR 35.38/0.9771; and (j) Ours 35.67/0.9786.

Based on the comparative analysis in Figures 3 and 4, the following observations can be derived:

(1) Comparison of overall visual effects

Images generated by Bicubic interpolation (Figures 3c and 4c) exhibit severe blurring accompanied by significant blocking artifacts, failing to recover the high-frequency textures of the original images. Although SRCNN (Figures 3g and 4g) and VDSR (Figures 3i and 4i) improve

sharpness compared to the interpolation algorithm, they tend to produce over-smoothed effects when processing infrared image edges, leading to a loss of fine texture details.

While CBAM (Figures 3d and 4d) incorporates an attention mechanism, it lacks specialized processing for infrared noise; consequently, its reconstruction results often contain noticeable residual background noise, preventing the image from appearing clean and sharp. Furthermore, while SE-Net (Figures 3e and 4e) attempts to enhance channel features, it occasionally introduces unnatural artifacts at the edges, which compromises the geometric structure of the image.

(2) Precise reconstruction of complex textures

Figure 3 illustrates a scene containing dense striped textures. At the × 2 scaling factor, many competing methods, such as SRCNN and EDSR, struggle to distinguish the dense lines, resulting in merging or blurring of the textures. In contrast, CASRNet successfully reconstructs the gaps between the stripes with sharp edges, providing the result most consistent with the Ground Truth (Figure 3b). This superior performance is attributed to the CS strategy, which effectively preserves feature diversity and prevents the loss of fine structural information during reconstruction.

(3) Contour preservation of fine structures

Figure 4 displays a close-up of a vehicle's front grille, a detail that is notoriously difficult to recover in low-resolution infrared imagery. Under the × 4 scaling factor, Bicubic and SRCNN essentially lose the geometric structure of the grille, leaving only a blurred shadow. While SE-Net recovers partial contours, it introduces significant geometric distortion.

In contrast, CASRNet restores the horizontal line structure of the grille with sharp black-and-white contrast. This success is primarily attributed to our designed CSA_ResBlock, where the spatial attention mechanism precisely localizes object edges, while the channel attention mechanism effectively suppresses the surrounding infrared thermal noise.

Summary of qualitative evaluation. The qualitative experimental results demonstrate that CASRNet not only leads in numerical metrics but also possesses a significant advantage in human visual perception. It consistently generates high-quality infrared images with sharper edges, reduced noise, and fewer artifacts. These findings further validate the effectiveness of our collaborative design, combining channel splitting with dual attention, in addressing the critical issues of feature redundancy and insufficient discriminative power in infrared image reconstruction.

4.7. Model complexity analysis

To comprehensively evaluate the practicality of the proposed CASRNet for real-world deployment, we quantitatively analyze its computational overhead. Specifically, when evaluated at a standard low-resolution input patch size of 128 × 128, CASRNet requires approximately 1.85 M parameters and 28.4 G FLOPs.

While adding attention mechanisms typically increases computational burden, our proposed CS strategy efficiently mitigates this. By decomposing the feature maps into parallel branches, the number of input channels for standard convolutions within the residual blocks is effectively halved. This structural design ensures that CASRNet maintains a highly competitive parameter and FLOPs count compared to extremely deep networks like EDSR (which typically requires over 40 M parameters in its full configuration). The inclusion of the CSA_ResBlock introduces a negligible

increase in parameters (primarily from 1 × 1 convolutions and MLP layers) but yields disproportionately high gains in texture reconstruction accuracy, proving that CASRNet achieves a highly favorable trade-off between model complexity and super-resolution performance.

5. Conclusions

In this paper, we proposed CASRNet, a novel collaborative network architecture designed to address the pervasive challenges of feature redundancy and insufficient discriminative power in infrared image super-resolution. By structurally guiding the network through a CS strategy, we effectively reduced redundant information and promoted feature diversity. Furthermore, the introduction of the channel and spatial attention residual block (CSA_ResBlock) empowered the network to adaptively focus on critical high-frequency edges while suppressing thermal noise. Extensive benchmark evaluations demonstrated that CASRNet quantitatively and qualitatively outperforms current state-of-the-art specialized and general SR models, achieving superior PSNR/SSIM scores and recovering highly complex geometric textures.

Limitations and future work:

Despite achieving encouraging results, we acknowledge certain limitations in this model that present opportunities for future improvement:

(1) Computational overhead in edge devices: While the CS strategy balances parameters, the large-kernel spatial attention (7 × 7) introduces computational latency. In the future, exploring lightweight attention mechanisms (e.g., depth-wise separable convolutions) will be necessary to deploy the model on low-power edge devices (like drone-mounted thermal cameras).

(2) Reliance on synthetic paired data: This model is trained on synthetically downsampled and noise-added images, which may not perfectly model the complex, non-Gaussian degradation of real-world infrared sensors. Exploring unsupervised or self-supervised learning paradigms using unpaired real-world data remains a critical next step.

(3) Temporal inconsistency: The model processes single images. Extending this architecture to Video Super-Resolution (VSR) by incorporating recurrent neural networks (RNNs) to leverage inter-frame temporal correlations could significantly improve performance in continuous monitoring scenarios.

(4) Generalization on extreme degradations: While CASRNet performs excellently on standard testing scales (× 2, × 4), its performance may be challenged when reconstructing extremely low-resolution inputs where critical thermal cues are corrupted. Furthermore, the model's zero-shot generalization across inherently different infrared sensor types (e.g., Uncooled Microbolometers vs. Photon Detectors) with vastly different noise distributions remains an area requiring further investigation and robust training strategies.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was funded in part by the Henan Province Key R&D Program Project, "Research and Application Demonstration of Class II Superlattice Medium Wave High Temperature Infrared Detector Technology" under Grant No. 2311111210400.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. K. I. Danaci, E. Akagunduz, A survey on infrared image & video sets, *Multimedia Tools Appl.*, **83** (2024), 16485–16523. https://doi.org/10.1007/s11042-023-15327-8
2. J. Wang, J. Ou, Y. Fan, L. Cai, M. Zhou, Online monitoring of electrical equipment condition based on infrared image temperature data visualization, *IEEJ Trans. Electr. Electron. Eng.*, **17** (2022), 583–591. https://doi.org/10.1002/tee.23545
3. F. Hou, Y. Zhang, Y. Zhou, M. Zhang, B. Lv, J. Wu, Review on infrared imaging technology, *Sustainability*, **14** (2022), 11161. https://doi.org/10.3390/su141811161
4. M. Alhameed, F. Jeribi, B. M. E. Elnaim, M. A. Hossain, M. E. Abdelhag, Pandemic disease detection through wireless communication using infrared image based on deep learning, *Math. Biosci. Eng.*, **20** (2023), 1083–1105. https://doi.org/10.3934/mbe.2023050
5. B. Jiang, S. Chen, B. Wang, B. Luo, MGLNN: Semi-supervised learning via multiple graph cooperative learning neural networks, *Neural Networks*, **153** (2022), 204–214. https://doi.org/10.1016/j.neunet.2022.05.024
6. A. M. Roy, J. Bhaduri, DenseSPH-YOLOv5: An automated damage detection model based on DenseNet and Swin-Transformer prediction head-enabled YOLOv5 with attention mechanism, *Adv. Eng. Inf.*, **56** (2023), 102007. https://doi.org/10.1016/j.aei.2023.102007
7. S. Jamil, A. M. Roy, An efficient and robust phonocardiography (PCG)-based valvular heart diseases (VHD) detection framework using vision transformer (VIT), *Comput. Biol. Med.*, **158** (2023), 106734. https://doi.org/10.1016/j.compbiomed.2023.106734
8. D. C. Lepcha, B. Goyal, A. Dogra, V. Goyal, Image super-resolution: A comprehensive review, recent trends, challenges and applications, *Inf. Fusion*, **91** (2023), 230–260. https://doi.org/10.1016/j.inffus.2022.10.007
9. D. Qiu, Y. Cheng, X. Wang, Medical image super-resolution reconstruction algorithms based on deep learning: A survey, *Comput. Methods Programs Biomed.*, **238** (2023), 107590. https://doi.org/10.1016/j.cmpb.2023.107590

10. K. Chauhan, S. N. Patel, M. Kumhar, J. Bhatia, S. Tanwar, I. E. Davidson, Deep learning-based single-image super-resolution: A comprehensive review, *IEEE Access*, **11** (2023), 21811–21830. https://doi.org/10.1109/ACCESS.2023.3251396
11. X. Wang, J. Yi, J. Guo, Y. Song, J. Lyu, J. Xu, et al., A review of image super-resolution approaches based on deep learning and applications in remote sensing, *Remote Sens.*, **14** (2022), 5423. https://doi.org/10.3390/rs14215423
12. M. Wei, X. Zhang, Super-resolution neural operator, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 18247–18256. https://doi.org/10.3390/rs14215423
13. Y. Zhang, Q. Fan, F. Bao, Y. Liu, C. Zhang, Single-image super-resolution based on rational fractal interpolation, *IEEE Trans. Image Process.*, **27** (2018), 3782–3797. https://doi.org/10.1109/TIP.2018.2826139
14. S. D. Sims, Frequency domain-based perceptual loss for super resolution, in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, (2020), 1–6. https://doi.org/10.1109/MLSP49062.2020.9231718
15. P. Singh, S. S. Bose, A quantum-clustering optimization method for COVID-19 CT scan image segmentation, *Expert Syst. Appl.*, **185** (2021), 115637. https://doi.org/10.1016/j.eswa.2021.115637
16. P. Singh, S. S. Bose, Ambiguous D-means fusion clustering algorithm based on ambiguous set theory: Special application in clustering of CT scan images of COVID-19, *Knowledge-Based Syst.*, **231** (2021), 107432. https://doi.org/10.1016/j.knosys.2021.107432
17. P. Singh, Y. P. Huang, AKDC: Ambiguous kernel distance clustering algorithm for COVID-19 CT scans analysis, *IEEE Trans. Syst. Man Cybern.: Syst.*, **54** (2024), 6218–6229. https://doi.org/10.1109/TSMC.2024.3418411
18. P. Singh, Y. P. Huang, An ambiguous edge detection method for computed tomography scans of coronavirus disease 2019 cases, *IEEE Trans. Syst. Man Cybern.: Syst.*, **54** (2023), 352–364. https://doi.org/10.1109/TSMC.2023.3307393
19. M. C. Catalbas, Modified VDSR-based single image super-resolution using naturalness image quality evaluator, *Signal, Image Video Process.*, **16** (2022), 661–668. https://doi.org/10.1007/s11760-021-02005-1
20. F. Kong, M. Li, S. Liu, D. Liu, J. He, Y. Bai, et al., Residual local feature network for efficient super-resolution, preprint, arXiv:2205.07514.
21. B. M. Kuriakose, J. Archpaul, V. E. Naveen, A. Lincy, EDSR: Empowering super-resolution algorithms with high-quality DIV2K images, *Intell. Decis. Technol.*, **17** (2023), 1249–1263. https://doi.org/10.3233/IDT-230043
22. X. Wang, L. Sun, A. Chehri, Y. Song, A review of GAN-based super-resolution reconstruction for optical remote sensing images, *Remote Sens.*, **15** (2023), 5062. https://doi.org/10.3390/rs15205062

23. V. Chudasama, H. Patel, K. Prajapati, K. Upla, R. Ramachandra, K. Raja, et al., Therisurnet-a computationally efficient thermal image super-resolution network, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, (2020), 388–397. https://doi.org/ 10.1109/CVPRW50498.2020.00051
24. K. Umehara, J. Ota, T. Ishida, Application of super-resolution convolutional neural network for enhancing image resolution in chest CT, *J. Digital Imaging*, **31** (2018), 441–450. https://doi.org/10.1007/s10278-017-0033-z
25. K. Prajapati, V. Chudasama, H. Patel, K. Upla, K. Raja, R. Ramachandra, Direct unsupervised super-resolution using generative adversarial network (DUS-GAN) for real-world data, *IEEE Trans. Image Process.*, **30** (2021), 8251–8264. https://doi.org/10.1109/TIP.2021.3113783
26. Z. Wang, B. Du, Y. Guo, Domain adaptation with neural embedding matching, *IEEE Trans. Neural Networks Learn. Syst.*, **31** (2019), 2387–2397. https://doi.org/10.1109/TNNLS.2019.2935608
27. Y. Ma, X. Wang, W. Gao, Y. Du, J. Huang, F. Fan, Progressive fusion network based on infrared light field equipment for infrared image enhancement, *IEEE/CAA J. Autom. Sin.*, **9** (2022), 1687–1690. https://doi.org/10.1109/JAS.2022.105812
28. R. Hou, D. Zhou, R. Nie, D. Liu, L. Xiong, Y. Guo, et al., VIF-Net: An unsupervised framework for infrared and visible image fusion, *IEEE Trans. Comput. Imaging*, **6** (2020), 640–651. https://doi.org/10.1109/TCI.2020.2965304
29. R. Hou, X. Li, T. Ren, D. Zhou, G. Wu, J. Cao, et al., HyPSAM: Hybrid prompt-driven segment anything model for RGB-thermal salient object detection, *IEEE Trans. Circuits Syst. Video Technol.*, **36** (2026), 2697–2712. https://doi.org/10.1109/TCSVT.2025.3613770
30. Y. Liu, Y. Wang, N. Li, X. Cheng, Y. Zhang, Y. Huang, et al., An attention-based approach for single image super resolution, in *2018 24Th International Conference on Pattern Recognition (ICPR)*, (2018), 2777–2784. https://doi.org/10.1109/ICPR.2018.8545760
31. J. Cai, Z. Meng, C. M. Ho, Residual channel attention generative adversarial network for image super-resolution and noise reduction, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, (2020), 454–455. https://doi.org/10.1109/CVPRW50498.2020.00234
32. B. Cui, H. Zhang, W. Jing, H. Liu, J. Cui, SRSe-net: Super-resolution-based semantic segmentation network for green tide extraction, *Remote Sens.*, **14** (2022), 710. https://doi.org/10.3390/rs14030710
33. Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in *Computer Vision–ECCV 2018*, (2018), 294–310. https://doi.org/10.1007/978-3-030-01234-2_18
34. C. Chen, D. Gong, H. Wang, Z. Li, K. Wong, Learning spatial attention for face super-resolution, *IEEE Trans. Image Process.*, **30** (2020), 1219–1231. https://doi.org/10.1109/TIP.2020.3043093

35. M. Yin, Z. Chen, C. Zhang, A CNN-transformer network combining CBAM for change detection in high-resolution remote sensing images, *Remote Sens.*, **15** (2023), 2406. https://doi.org/10.3390/rs15092406
36. H. Fang, M. Xia, G. Zhou, Y. Chang, L. Yan, Infrared small UAV target detection based on residual image prediction via global and local dilated residual networks, *IEEE Geosci. Remote Sens. Lett.*, **19** (2021), 1–5. https://doi.org/10.1109/LGRS.2021.3085495
37. Z. Wu, J. Chen, L. Tan, H. Gong, Y. Zhou, G. Shi, A lightweight GAN-based image fusion algorithm for visible and infrared images, preprint, arXiv:2409.15332.
38. R. E. Rivadeneira, A. D. Sappa, C. Wang, J. Jiang, Z. Zhong, P. Chen, et al., Thermal image super-resolution challenge results-pbvs 2024, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2024), 3113–3122. https://doi.org/10.1109/CVPRW63382.2024.00317
39. R. E. Rivadeneira, P. L. Suárez, A. D. Sappa, B. X. Vintimilla, Thermal image superresolution through deep convolutional neural network, in *Proceedings of the Image Analysis and Recognition: 16th International Conference (ICIAR2019)*, (2019), 417–426. https://doi.org/10.1007/978-3-030-27272-2_37

AIMS Press

©2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)