Contourlet Residual for Prompt Learning Enhanced Infrared Image Super-Resolution

Xingyuan Li[1], Jinyuan Liu[1⋆], Zhixin Chen[2], Yang Zou[3], Long Ma[1], Xin Fan[1], and Risheng Liu[1]

[1] School of Software Technology, Dalian University of Technology
[2] Graduate School of Information, Production, and Systems, Waseda University
[3] School of Computer Science, The University of Sydney
xingyuan_lxy@163.com, atlantis918@hotmail.com

Abstract. Image super-resolution (SR) is a critical technique for enhancing image quality, playing a vital role in image enhancement. While recent advancements, notably transformer-based methods, have advanced the field, infrared image SR remains a formidable challenge. Due to the inherent characteristics of infrared sensors, such as limited resolution, temperature sensitivity, high noise levels, and environmental impacts, existing deep learning methods result in suboptimal enhancement outcomes when applied to infrared images. To address these challenges, we propose a specialized Contourlet residual framework tailored for infrared images to restore and enhance the critical details from the multi-scale and multi-directional infrared spectra decomposition. It precisely captures and amplifies the high-pass subbands of infrared images, such as edge details and texture nuances, which are vital for achieving superior reconstruction quality. Moreover, recognizing the limitations of traditional learning techniques in capturing the inherent characteristics of infrared images, we incorporate a prompt-based learning paradigm. This approach facilitates a more nuanced understanding and targeted optimization process for infrared images by leveraging the semantic comprehension offered by the visual language model. Our approach not only addresses the common pitfalls associated with infrared imaging but also sets a new paradigm for infrared image SR. Extensive experiments demonstrate that our approach obtains superior results, attaining state-of-the-art performance. Project page: https://github.com/hey-it-s-me/CoRPLE.

Keywords: Infrared Image Super-Resolution · Contourlet Residual · Prompt Learning

1 Introduction

Single image super-resolution (SR) [10, 14, 26] recovers a high-resolution image from its lower-resolution counterpart without necessitating modifications to the

⋆ Corresponding author.
This work was partially supported by China Postdoctoral Science Foundation (2023M730741), and the National Natural Science Foundation of China (No.62302078 and No.61936002).

imaging hardware. Its applications span across diverse fields, including medical imaging [20, 50], security surveillance [40], and remote sensing image processing [19, 44]. Infrared imaging, which captures grayscale images reflecting the infrared radiation energy emitted by objects, has proven a crucial role in military and civilian domains such as wilderness reconnaissance, aerospace, and home care [51]. However, the intrinsic properties of infrared cameras, including low resolution, temperature sensitivity, high noise levels, and dynamic range limitations, pose prevalent challenges that compromise the quality of infrared images [29]. These issues obstruct crucial tasks including object detection, tracking, and segmentation [13, 34, 54]. Therefore, the enhancement of infrared image resolution to bolster contrast and detail is imperative for augmenting the efficiency and accuracy of these vision tasks.

The evolution of super-resolution methodologies, from convolutional neural network (CNN)-based approaches to transformer-based innovations, has significantly advanced the field. Most CNN-based methods [10, 11, 26] prioritize intricate architectural designs, leveraging spatially invariant kernels to extract local features, which are inefficient in modeling the relations among pixels and are not adequate for the establishment of long-range dependencies [25]. Transformer-based methods [5, 25, 46] resolve those issues by the design of the self-attention mechanism that captures global interactions between contexts and has shown promising performance.

However, most existing methods are designed with visible light images and do not adequately address the unique characteristics of infrared light, thus resulting in unsatisfactory performance when applied to infrared images. Infrared images, characterized by longer wavelengths and less susceptibility to atmospheric scattering, contain fewer high spatial frequency components. Furthermore, the process of forward propagation in neural networks often results in the diminution of high-frequency details [30, 53]. Additionally, the optical components of infrared imaging systems may not focus infrared light as effectively as visible light, therefore impacting the clarity of high-frequency details in images [12].

To bridge this gap, we propose a Contourlet residual-based prompt learning approach for infrared SR. Our approach leverages the multi-directional and multi-scale analysis capabilities of the Contourlet transform, along with its efficient edge representation and noise reduction properties, to enhance the deep feature extraction of infrared images efficiently. Furthermore, by employing the visual language model, we imbue our model with a profound semantic understanding through a two-stage prompt learning strategy which learns right from wrong, guiding the optimization process through paired positive and negative textual prompts, meanwhile bolstering its learning capability and generalization. Our contributions are three-fold:

- We introduce a specialized Contourlet residual framework tailored for infrared images to restore and enhance the high-frequency details from the multi-scale infrared spectra decomposition, crucial for reconstructing high-pass subband lacked infrared images.

- We devise a prompt learning strategy that guides our super-resolution model to optimize the unique nuances of infrared image characteristics through the positive and negative prompt pairs.
- Our method surpasses existing super-resolution algorithms and achieves state-of-the-art (SOTA) performance, setting a new paradigm in the realm of infrared image super-resolution.

2 Related Work

2.1 Image Super-Resolution

GAN-based SR Method. Starting from SRCNN [9], numerous deep learning-based Super-Resolution (SR) have been proposed. Recently, BSRGAN [48] and Real-ESRGAN [42] use a wide variety of training samples with different types of degradation, leveraging GAN to enhance the quality of SR images but may introduce artifacts. Subsequent methods like LDL [24] and DeSRA [43] have mitigated these artifacts but struggle to reproduce natural details.
Transformer-based SR methods. Methods such as IPT [3] demonstrate the adaptability of the Transformer to various image processing tasks but require large datasets to reach their potential [55]. VSR-Transformer [2] enhances the resolution of video frames through temporal and spatial features but features are still extracted from CNN. Liang et al. proposed Swin IR [25], to combine the strengths of both CNNs and Transformers to enhance SR tasks. ELAN [49] is capable of computing self-attention in larger windows, and Restormer [46] can learn long-range dependencies while maintaining computational efficiency.
Diffusion-based SR methods. Methods like SR3 [39] condition a diffusion model (DM) on low-resolution (LR) images, gradually refining them to high-resolution (HR) images. Recently, SRDiff [21] employs a residual prediction strategy to accelerate training efficiency and utilizes encoded LR information for noise prediction.

2.2 Text Prompt Image Processing

Methods like DALL-E-2 [36], Imagen [38] and Stable Diffusion [37] all utilize diffusion models for text-to-image generation. For image manipulation, StyleCLIP [32] combines the generative capabilities of StyleGAN [17] with the vision-language abilities of CLIP [35], and DiffusionCLIP [18] uses diffusion models alongside CLIP for image generation. Meanwhile, Prompt-to-Prompt [15] and InstructPix2Pix [1] are pre-trained and fine-tuning-free approaches, edit images within a pre-trained diffusion model by modifying prompts [52]. However, the use of text prompts in image SR has seen limited exploration. In our work, we investigate the application of text prompts in the context of infrared image SR.

2.3 Infrared Image Enhancement

Infrared image enhancement aims to improve the quality of images captured by IR sensors. However, reconstructing IR details presents a significant challenge.

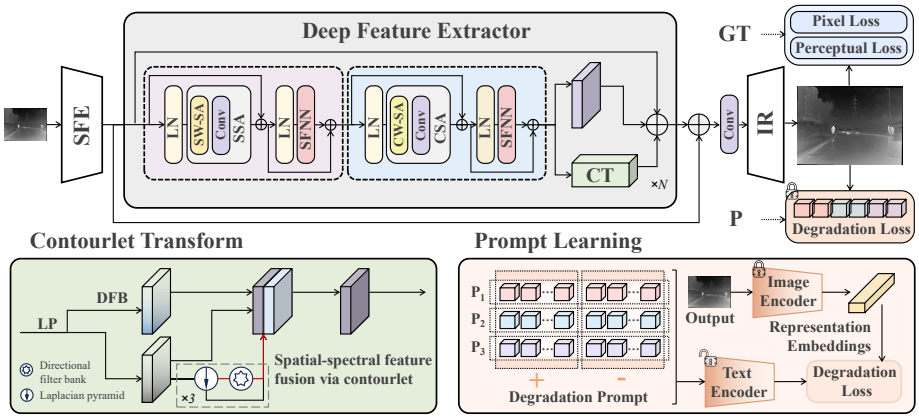


Fig. 1: Overall architecture of our network, Contourlet transform and prompt leaning optimization process.

The defects of infrared imaging devices and external environmental factors often result in images with low contrast, unclear target edges, and poor visual effects [23, 27]. These challenges are exacerbated by the longer wavelengths of infrared radiation than visible light, leading to images with reduced spatial resolution and diminished detail [28]. Many efforts have started investigating IR image enhancement [31, 45]. Marivani et al. [31] initially integrate sparse edge information from visible light images and combine it with interpretable sparse priors. Other researchers have developed modules capable of extracting high-frequency information from visible light images and using attention mechanisms to effectively introduce this pattern information into the IR feature domain [16, 33].

3 Methods

3.1 Architecture

As shown in Figure 1, the architecture of our proposed network delineates three core modules: the shallow feature extraction module, the deep feature extraction module, and the high-resolution (HR) image reconstruction module. The process begins with a low-resolution (LR) input image being processed through a shallow feature extraction phase, employing a 3×3 convolutional layer to preliminarily parse the image's basic features. Progressing deeper, the model advances into the deep feature extraction phase, which amalgamates spatial and infrared spectral features through the channel and spatial self-attention mechanisms alongside the Contourlet residuals. In the final phase, the HR image is reconstructed through the HR image reconstruction module. Here, the fused features undergo an upscaling process using the pixel shuffle method [41] with convolutional layers employed to aggregate features into the final HR image.

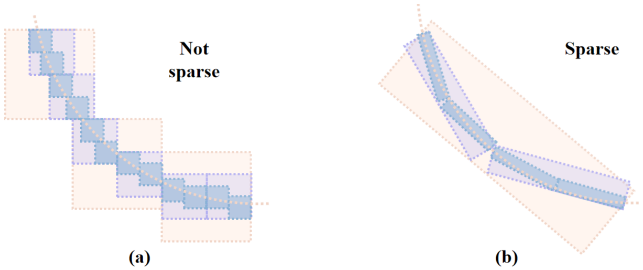


Fig. 2: Visual demonstration of the sparsity for (a). traditional wavelet transform and (b). Contourlet transform.

3.2 Contourlet Residual Deep Feature Extraction

The deep feature extraction module leverages channel and spatial self-attention mechanisms with the Contourlet transform, optimizing infrared feature fusion across spatial, channel, and infrared spectra for enhanced representation.

Spatial Window Self-Attention. This mechanism operates by segmenting feature spaces into discrete spatial windows, applying self-attention within each to capture intricate spatial relationships. For an input feature matrix $X \in \mathbb{R}^{H\times W\times C}$, we employ learnable weights to generate query (Q), key (K), and value (V) matrices. These matrices are then subdivided into non-overlapping windows, with each window processed independently to emphasize localized feature interactions. The process involves dividing these matrices into multiple heads, enabling parallel processing of diverse feature aspects. Attention scores are computed through the dot product of the queries and keys, normalized with the softmax function to obtain attention weights for each position. The attention outputs from all heads are then concatenated and subjected to a linear projection, ensuring a comprehensive integration of spatially attentive features into a unified representation.

Channel-Wise Self-Attention. Unlike Spatial Window Self-Attention, Channel-Wise Self-Attention applies attention across channels for each spatial window, capturing global channel relationships. Given input $X \in \mathbb{R}$, we linearly project it to form query (Q), key (K), and value (V) matrices. These are then reshaped to Q_c, K_c, and V_c with dimensions $\mathbb{R}^{HW\times C}$, and divided into multiple heads (h). Attention is computed for each head, using a learnable scaling factor (α) for normalization:

$$Y_c^i = \text{softmax}\left(\frac{(Q_c^i)^T K_c^i}{\alpha}\right) \cdot V_c^i, \tag{1}$$

yielding the final channel-wise attention output by concatenating and reshaping head outputs, akin to the spatial attention process.

Spatial Feed-forward Neural Network. The Spatial Feed-Forward Neural Network (SFNN) is a neural network architecture that enhances traditional feed-forward networks by incorporating a spatial gating mechanism for improved spatial information processing and channel redundancy reduction via depth-wise

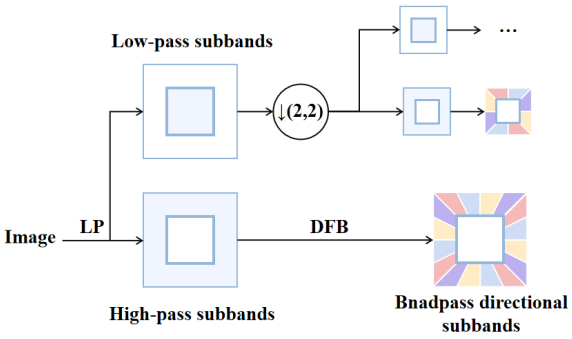


Fig. 3: Architecture of the Contourlet transform, The input feature is first decomposed by an LP filter to low- and high-pass subbands. Then, the high-pass subbands are decomposed into 2^i directional subspaces through the DFB.

convolution and element-wise multiplication. It makes the architecture more efficient by using depth-wise convolution. SFNN splits the input feature map ($\hat{X}$) along the channel axis, processes each part through convolutional and multiplicative operations, and then merges them:

$$\begin{aligned} \hat{X}' &= \sigma(W_p^1 \hat{X}), \quad [\hat{X}_1', \hat{X}_2'] = \hat{X}', \\ \text{SFNN}(\hat{X}) &= W_p^2(\hat{X}_1' \odot (W_d \hat{X}_2')), \end{aligned} \tag{2}$$

where $\odot$ denotes the element-wise multiplication, W_p^1 and W_p^2 is the weight matrix used for linear projection. σ represents the GELU function, and W_d represents the depth-wise convolutional parameters.

Contourlet Residual. Our preference for the Contourlet transform over the conventional wavelet transform stems from its superior ability to handle the multidimensional singularities typical of infrared images, such as lines, edges, and contours. Although wavelet transform exhibits satisfactory time-frequency localization, its square supports are suboptimal for effectively capturing high-dimensional features and fail to provide a sparse representation. Additionally, the convolutional nature of wavelet transform is computationally intensive and does not possess translation invariance, which can result in the Gibbs phenomenon [8]. In contrast, the Contourlet transform leverages the strengths of wavelets into a higher-dimensional space, and achieves a sparse representation where smooth image contours can be efficiently captured with fewer coefficients, thereby enhancing robustness across scales and orientations as shown in Figure 2.

Upon extracting global deep features through the SFNN block, we employ a Contourlet-based residual network, as depicted in Figure 3. This network initiates with a Laplacian Pyramid decomposition of the deep features, denoted as $\mathbf{X}$, segregating them into low and high-pass components. The low-frequency subband $\mathbf{X_{low}}$, encapsulating the core structure and broad contours of the image, is procured through Gaussian filter down-sampling expressed as

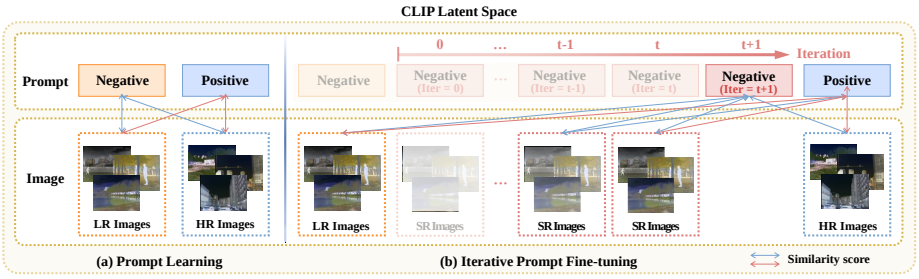


Fig. 4: Overview of the prompt learning process. (a). The unlocked text encoder optimizes the learnable prompts to maximize the distance between negative and positive semantics in the latent space. (b). The degradation loss guides the SR to align with positive prompts while distancing from the negative ones with the locked text encoder.

$G_i(x, y) = (X_{i-1} * h)(2x, 2y)$, where G_i represents the image at the i^{th} level of the Gaussian pyramid, X_{i-1} is the feature from the previous level, h denotes the Gaussian filter, $*$ is the convolution operation, and $(2x, 2y)$ signifies the downsampling process, capturing the general features across scales. The Laplacian layer L_i, embodying the high-frequency subband $\mathbf{X_{high}}$, is derived by:

$$L_i(x, y) = X_{i-1}(x, y) - (G_i * h^T)(x, y), \tag{3}$$

with L_i forming the i^{th} level of the Laplacian pyramid and h^T being the transposed Gaussian filter for reconstructing the preceding layer's features. The high-frequency subband $\mathbf{X_{high}}$, ensuing from the LP decomposition, undergoes a subsequent refinement via the Directional Filter Bank decomposition:

$$B_{l,k}(x, y) = (X_l * f_k)(x, y), \tag{4}$$

where $B_{l,k}$ denotes the subband for the l^{th} level and k^{th} direction, and f_k is the directional filter. The DFB decomposition dissects the image into directionally-sensitive subbands, allowing the model to discern textural and edge details within the high-frequency domains of the infrared spectrum. This decomposition is iteratively applied to the low-pass subband at each LP level, recursively utilizing both LP and DFB to extract and refine features across the infrared spectrum from coarse to fine.

The resultant coefficients $\mathbf{X}_{spectral}$ are a confluence of the multi-scale, multi-directional features, expressed as $\mathbf{X}_{spectral} = \{L_i(x, y)\} \cup \{B_{l,k}(x, y)\}$, where $\{L_i(x, y)\}$ is the set of all Laplacian pyramid layers, and $\{B_{l,k}(x, y)\}$ is the set of all DFB decomposed subbands. These coefficients encompass not only the sparse high-frequency nuances but also the fundamental low-frequency aspects, delivering a comprehensive suite of features characteristic of infrared images. After employing the Contourlet transform to enrich the spectral features, we fuse these with the spatial features extracted from the SFNN block through residual to enhance the detail representation in infrared images.

3.3 Prompt Learning Based Optimization

To refine the quality of super-resolved infrared images, we introduce a two-stage prompt learning strategy leveraging the capabilities of the CLIP model. This strategy augments the capacity to adaptively interpret and improve upon the quality of SR and detail within reconstructed scenes.

The foundation of this prompt learning strategy is the prompt-based degradation loss, which utilizes the CLIP model's proficiency in semantic parsing to align the generated images with textual descriptors. We apply the three pairs of prompts to represent degradation specifically common in infrared images and formulate the degradation loss using CLIP to guide the super-resolution process toward generating images that semantically align with positive textual descriptors while distancing from the negative ones.

As shown in Figure 4, our method initializes the prompt pairs from high-resolution (HR) and low-resolution (LR) image counterparts. The HR image undergoes encoding via the locked image encoder $\mathbf{\Phi}_{\text{image}}$ of the CLIP model, producing its latent representation. Concurrently, the latent codes for the dichotomous prompts are derived through the unlocked text encoder $\mathbf{\Phi}_{\text{text}}$. Leveraging the latent space similarity metric $\text{SIM}(\mathbf{I}, \mathbf{T}) = e^{\cos(\mathbf{\Phi}_{\text{image}}(\mathbf{I}), \mathbf{\Phi}_{\text{text}}(\mathbf{T}))}$, we apply binary cross-entropy loss to fine-tune the initial prompt pair, distinguishing HR from LR images:

$$\mathcal{L} = -(y * \log(\hat{y}) + (1 - y) * \log(1 - \hat{y})), \tag{5}$$

$$\hat{y} = \frac{\text{SIM}(\mathbf{\Phi}_{\text{image}}(\mathbf{I}), \mathbf{\Phi}_{\text{text}}(\mathbf{T}_{\text{pos}}))}{\sum_{i \in \{\text{neg,pos}\}} \text{SIM}(\mathbf{\Phi}_{\text{image}}(\mathbf{I}_i), \mathbf{\Phi}_{\text{text}}(\mathbf{T}_i))}, \tag{6}$$

where $\mathbf{I}$ signifies the paired HR and LR images, and y is their corresponding label, designated 0 for LR and 1 for HR. $\mathbf{T}_{\text{pos}}$ and $\mathbf{T}_{\text{neg}}$ encapsulate the encoded features of the positive and negative prompts, respectively. After the initial stage of prompt optimization, we then lock the text encoder $\mathbf{\Phi}_{\text{text}}$, and advance to refine our network with the degradation loss. The degradation loss $\mathcal{L}_{degrad}$ for a batch of SR images $\mathbf{I}$ is computed as:

$$\mathcal{L}_{degrad} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{SIM}(\mathbf{\Phi}_{\text{image}}(\mathbf{I}_i), \mathbf{\Phi}_{\text{text}}(\mathbf{T}_{\text{neg}}))}{\text{SIM}(\mathbf{\Phi}_{\text{image}}(\mathbf{I}_i), \mathbf{\Phi}_{\text{text}}(\mathbf{T}_{\text{pos}}))}, \tag{7}$$

where N indicates the batch size. This loss function motivates the network to yield visually aligned images with high-quality descriptors while diverging from the qualities of the low-quality ones, thus ensuring a visual alignment. The training alternates between refining the prompts and fine-tuning the enhancement network until the outputs achieve visual excellence. The total loss, $\mathcal{L}_{total}$, integrates the degradation loss with pixel and perceptual losses to holistically optimize the infrared images for visual fidelity, perceptual quality, and semantic congruence with high-quality image descriptions:

$$\mathcal{L}_{total} = \mathcal{L}_{degrad} + \mathcal{L}_{pixel} + \mathcal{L}_{perceptual}, \tag{8}$$

Table 1: Quantitative comparison with the SOTA SR methods. The best result is in red whereas the second best one is in blue.

Methods	Scale	Set5			Set15			Set20		
		PSNR↑	MSE ↓	SSIM ↑	PSNR ↑	MSE ↓	SSIM ↑	PSNR ↑	MSE ↓	SSIM ↑
SwinIR-Light	x2	46.670	5.365	0.9896	47.120	5.158	0.9900	47.680	4.151	0.9900
SwinIR	x2	46.880	5.121	0.9899	47.240	5.001	0.9902	47.820	4.023	0.9901
DAT-Light	x2	48.188	5.196	0.9920	48.555	5.175	0.9922	49.093	4.196	0.9924
DAT-S	x2	48.454	4.868	0.9919	48.825	4.831	0.9917	49.366	3.902	0.9918
DAT	x2	48.434	4.818	0.9922	48.779	4.800	0.9924	49.348	3.848	0.9925
CAT-R-2	x2	48.467	4.784	0.9911	48.817	4.721	0.9923	49.422	3.763	0.9914
CAT-R	x2	48.490	4.710	0.9915	48.842	4.652	0.9923	49.453	3.707	0.9918
ART-S	x2	48.368	5.858	0.9913	48.747	5.214	0.9923	49.361	5.705	0.9916
ART	x2	48.470	5.096	0.9917	48.834	4.479	0.9924	49.443	4.061	0.9925
HAT-S	x2	48.456	4.805	0.9919	48.841	4.740	0.9914	49.416	3.811	0.9918
HAT	x2	48.430	4.878	0.9921	48.790	4.805	0.9923	49.368	3.867	0.9924
HAT-L	x2	48.584	4.474	0.9921	48.974	4.418	0.9924	49.518	3.561	0.9925
EDT-T	x2	48.115	4.837	0.9917	48.601	4.598	0.9920	49.104	3.797	0.9921
EDT-B	x2	48.188	4.497	0.9919	48.709	4.392	0.9921	49.243	3.550	0.9922
EDT-B†	x2	48.570	4.527	0.9921	49.008	4.354	0.9924	49.558	3.503	0.9925
Ours	x2	48.581	4.476	0.9930	49.233	4.389	0.9932	49.614	3.525	0.9930
SwinIR-Light	x4	38.720	35.478	0.9502	38.780	38.190	0.9473	40.070	24.694	0.9620
SwinIR	x4	39.010	33.155	0.9524	40.120	36.662	0.9598	41.680	22.877	0.9700
DAT-Light	x4	40.345	32.555	0.9606	40.399	35.299	0.9574	41.637	23.078	0.9701
DAT-S	x4	40.679	30.101	0.9622	40.702	32.518	0.9593	41.919	21.232	0.9709
DAT	x4	40.764	29.454	0.9625	40.812	31.854	0.9597	42.046	20.783	0.9712
CAT-R-2	x4	40.580	30.675	0.9617	40.621	32.879	0.9588	41.940	21.139	0.9708
CAT-R	x4	40.604	30.786	0.9619	40.612	33.615	0.9585	41.879	21.390	0.9706
ART-S	x4	40.597	30.604	0.9615	40.626	30.250	0.9586	41.917	33.101	0.9708
ART	x4	40.696	32.469	0.9619	40.754	31.241	0.9593	42.045	20.702	0.9712
HAT-S	x4	40.677	30.536	0.9620	40.760	32.389	0.9594	41.994	21.191	0.9710
HAT	x4	40.736	30.118	0.9622	40.740	32.487	0.9594	41.996	20.991	0.9711
HAT-L	x4	40.754	30.467	0.9624	40.741	31.985	0.9595	41.997	20.753	0.9713
EDT-T	x4	40.136	34.249	0.9532	40.383	34.953	0.9523	41.259	23.516	0.9621
EDT-B	x4	40.520	31.501	0.9610	40.641	33.549	0.9584	41.877	21.817	0.9706
EDT-B†	x4	40.618	30.606	0.9616	40.715	32.497	0.9590	41.948	21.041	0.9708
Ours	x4	40.779	29.387	0.9626	40.765	30.196	0.9600	41.949	20.691	0.9717

where the pixel loss $\mathcal{L}_{pixel}$, fundamentally an MSE calculation, assesses the pixel-level discrepancies between the super-resolved images and their high-resolution ground truth counterparts. Conversely, the perceptual loss $\mathcal{L}_{perceptual}$ leverages a VGG network to extract and compare feature representations of the generated and ground truth images, focusing on minimizing differences in high-level feature representations. This comprehensive loss function ensures that our SR network produces high-quality images that not only closely resemble the ground truth but also align with the semantic expectations.

4 Experiments

4.1 Experimental Settings

Dataset and evaluation metrics. We assessed our model against several benchmarks using publicly accessible infrared datasets M^3FD, TNO, and Road-

Table 2: Quantitative analysis of model complexity comparisons (×4). PSNR (dB) on TNO and RoadScene, FLOPs, and Params are reported with the best result in red whereas the second best one in blue.

Methods	SwinIR	DAT	CAT-R	ART	HAT	EDT	Ours
Params(M)	11.90	14.90	16.20	16.55	20.80	11.6	6.28
FLOPS(G)	215.3	275.8	292.7	573.2	102.4	37.6	52.5
TNO	35.32	36.65	36.87	36.86	36.88	36.73	36.91
RoadScene	27.86	29.28	29.24	29.27	29.25	28.84	29.31

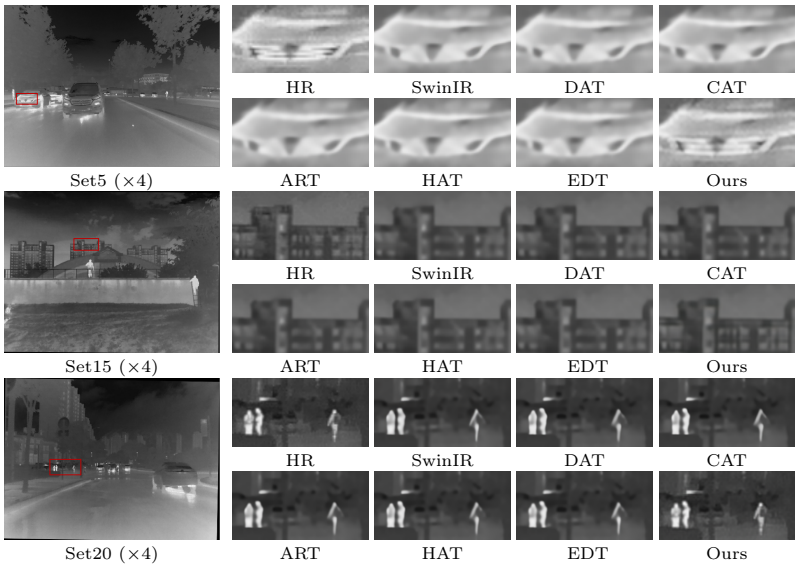
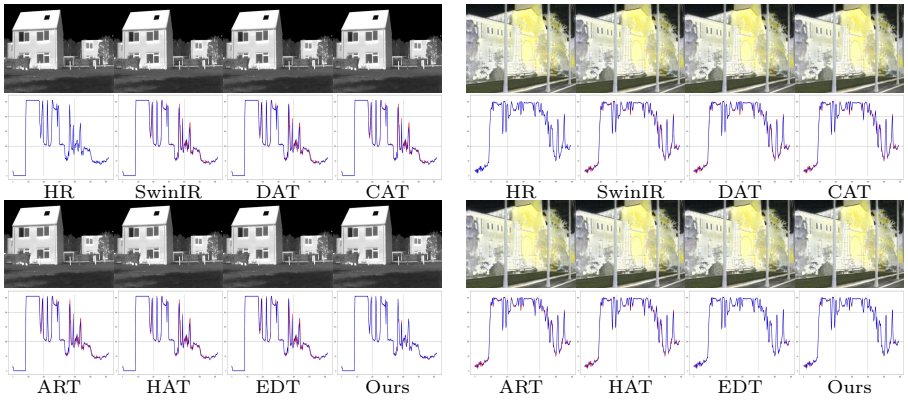


Fig. 5: Visual comparison of infrared image SR (×4) with SOTA methods.

Scene. For training, we randomly selected 182 images from M^3FD, while a set of 78 images from the same dataset was reserved for validation. To ensure a thorough evaluation of fusion performance, we employed M^3FD **Set5**, **Set15**, and **Set20** along with RoadScene (60 images), and TNO (37 images) as test datasets. We conduct experiments with upscaling factors of ×2 and ×4, where low-resolution (LR) images are derived from high-resolution (HR) counterparts through bicubic degradation. To quantitatively assess the performance of our model, we utilize three metrics, PSNR, MSE, and SSIM, offering a comprehensive evaluation of our model's effectiveness in image super-resolution.

Implementation Details. Our network was trained on a GeForce RTX 4090 GPU, utilizing the Adam optimizer for parameter updates. We set the initial learning rate to $1e^{-4}$, employing an exponential decay strategy to refine the learning process over time. The training was executed with patch size 64×64 and batch size 32. To enhance the robustness and generalizability of our model, we employed data augmentation techniques including random rotations (90°, 180°, 270°) and horizontal flips.



(a) TNO **(b)** RoadScene

Fig. 6: Visual comparison with SOTA methods on TNO and RoadScene datasets (×4) with pixel intensity variations visualized as line charts. The colors are added solely for a better view.

4.2 Comparison with State-of-the-Art Methods

We conduct a comprehensive comparison of our approach against six state-of-the-art methods, including SwinIR [25], DAT [6], CAT [7], ART [47], HAT [4], and EDT [22] with their large, classic, and light versions.

Quantitative Results. Our method's quantitative super-resolution results, as summarized in Table 1, demonstrate its superior performance across different scales when benchmarked against 15 state-of-the-art models on three representative sets from the M^3FD. Notably, our approach consistently outperforms existing methods in terms of PSNR and SSIM metrics, which are critical indicators of image quality and structural integrity. At a scale of ×2, our technique not only surpasses all classic and light models but also outmatches the majority of the larger versions of six cutting-edge methods. It trails only marginally behind HAT-Large and EDT-$B^\dagger$, marking a substantial advancement over its contemporaries. At a scale of ×4, our model secures the leading position in **Set5**, and closely contends with DAT, which achieves a marginally higher PSNR in **Set15** and **Set20**, underscoring our model's efficacy in reconstructing finer details and achieving higher fidelity in super-resolved images. Crucially, our model obtains this superior performance with notably less computational complexity and model size over its contemporaries as quantized in Table 2.

Qualitative comparison. We further conduct a qualitative analysis to compare the super-resolution quality of our model against various baseline models. The ×4 SR outcomes depicted in Figure 5 highlight our method's proficiency in diminishing artifacts while retaining more structural integrity and finer high-frequency details. For example, in the **Set5** results, competing methods often yielded over-smoothed reconstructions, failing to preserve infrared high-frequency details. This tendency is similarly observed in the **Set20** results, where other models struggled with the maintenance of high-frequency details in complex infrared

Fig. 7: Visual ablation results (×4) of different attention choices.

Table 3: Ablation study of Contourlet structure.

DFB	LP	Params(M)	FLOPS(G)	PSNR	SSIM
✓		6.27	52.2	40.462	0.9598
	✓	6.26	52.1	40.517	0.9601
✓	✓	6.28	52.5	40.779	0.9626

Table 4: Ablation study of prompt choice.

Baseline	Positive	Negative	PSNR	SSIM
✓			40.342	0.9606
✓	✓		40.731	0.9622
✓	✓	✓	40.779	0.9626

scenes. In contrast, our approach demonstrates an efficient preservation of infrared detail and texture, as evidenced by the sharper and more defined reconstructions. The experiment illustrated in Figure 6 further intuitively corroborates the nuanced multi-scale and multi-directional spectral comprehension of infrared images by our model. We draw line charts that trace pixel intensity variations along the image diagonals, the HR image is delineated in blue, while the reconstructed SR output is drawn in red. Notably, the red lines in our charts closely overlay the blue in regions where high-frequency details are ignored by other methods, underscoring our method's efficient edge representation and spectral comprehension properties.

4.3 Ablation Studies

Experiments on Attention Strategy. Figure 7 delineates the individual and combined contributions of Spatial Window Self-Attention (SSA) and Channel-Wise Self-Attention (CSA) mechanisms in our framework. The experimental results highlight that utilizing both SSA and CSA synergistically yields a significant enhancement in restoring the spatial features in the reconstruction of infrared super-resolved images. Models devoid of SSA exhibit a noticeable decline in maintaining structural details, while those lacking CSA demonstrate an

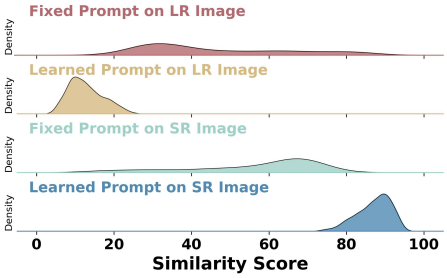


Fig. 8: Similarity comparison of ablation study on the prompt learning strategy.

Table 5: Ablation study of prompt learning strategy.

Methods	PSNR	SSIM
Fixed prompts(Positive and Negative)	40.655	0.9611
Learned prompts(Negative)	40.718	0.9621
Learned prompts(Positive)	40.702	0.9619
Ours	40.779	0.9626

impaired ability to reconstruct textural nuances. In contrast, our integrated approach, which leverages both SSA and CSA, closely mirrors the high-resolution (HR) target, capturing the intricate high-frequency details and preserving the fidelity of the HR image.

Examine Contourlet Residual mechanisms. The results illustrated in Table 3 provide an insight into the influence of the Laplacian Pyramid (LP) and Directional Filter Bank (DFB) within our Contourlet Residual block. Three scenarios were tested: the combined use of DFB and LP, the exclusive use of DFB, and the sole employment of LP. The result indicates the efficacy of DFB and LP, using both components achieves the highest PSNR, highlighting their complementary roles in capturing the multi-directional and multi-scale infrared high-pass details. The omission of either component reduces the model's effectiveness, emphasizing the importance of these components in our model.

Analyzing the Prompt Choice. The impact of positive and negative prompts on the super-resolution process is detailed in Table 4. The baseline model, devoid of any prompts, set the foundation for our comparison. Subsequent incorporation of positive prompts guided the model closer to the positive description and resulted in a marked improvement in both PSNR and SSIM metrics, signifying the utility of infrared characteristic relevant textual guidance. The addition of negative prompts further guided the model closer to the positive cues while distancing from the negative ones, refining the results as evidenced by the highest recorded PSNR and SSIM scores. This progression underscores the nuanced role that prompts play, effectively steering the model towards a more precise reconstruction of high-fidelity infrared images.

Impact of Learning Strategy. Figure 8 delves into the effects of the prompt learning strategy, presenting the kernel density estimates of similarity scores between prompts and images across the M^3FD test dataset. The curves exhibit how learned prompts, as opposed to fixed ones, yield a distinct peak in similarity scores when compared with SR images, suggesting a closer alignment with the SR outputs. In contrast, the flatter distribution associated with the fixed prompts indicates a less accurate reflection of the SR image attributes. Notably, the shift towards higher similarity scores underscores the effectiveness of our two-stage learning strategy in fine-tuning the prompts to embody the high-resolution

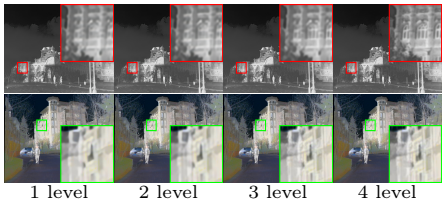


Fig. 9: Colored visual comparison of different contourlet decomposition levels.

Table 6: Ablation study of different Contourlet levels.

Methods	Params(M)	FLOPS(G)	PSNR	SSIM
1-level	5.35	49.7	40.735	0.9621
2-level	5.66	50.8	40.752	0.9625
3-level	5.97	51.2	40.768	0.9626
4-level	6.28	52.5	40.779	0.9626

characteristics more faithfully. As the quantitative results indicate in Table 5, learned prompts maximize the distance between the representation of negative and positive samples in the CLIP latent space. Consequently, this precision in representation ensures that the learned prompts more effectively guide the model toward emulating positive attributes while avoiding negative ones.

Evaluating the Contourlet Level. To further validate the influence of varying levels in the Contourlet decomposition process on image super-resolution, we provide the quantitative comparison in Table 6. An increase in the number of Contourlet levels augments the layers in the Laplacian Pyramid and the corresponding directional filter banks (DFBs) at each layer. This multi-layered approach empowers our model to capture high-frequency details across a broader spectrum of scales and orientations, thereby enriching the interpretation of infrared spectral characteristics. As presented, the slight uptick in PSNR and SSIM metrics with each added level substantiates the model's enhanced proficiency in delineating intricate infrared image textures. While a corresponding increase in Parameters and FLOPS is observed, the marginal improvement in capturing the nuanced details of infrared images justifies the additional computational overhead. The visual ablation results in Figure 9 also prove the efficacy of the decomposition level.

5 Conclusion

In this paper, we propose an efficient paradigm for infrared image SR through the development of a Contourlet residual-based prompt learning framework. Different from existing SR techniques predominantly optimized for visible light imaging, which neglect the distinctive characteristics of infrared light, we analyze the unique spectral signatures of infrared images and efficiently restore high-frequency details. Specifically, the Contourlet residual amalgamates spatial and infrared spectral features through the Contourlet transform, capturing multi-directional and multi-scale high-pass subband effectively. Our two-stage prompt learning strategy incrementally optimizes the model, facilitating the alignment with infrared image characteristics. Extensive experiments demonstrate that our method gains a deeper understanding of infrared imaging with relatively fewer parameters and FLOPS, achieves state-of-the-art performance, and establishes a new paradigm in the field of infrared image SR.

References

1. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
2. Cao, J., Li, Y., Zhang, K., Van Gool, L.: Video super-resolution transformer. arXiv preprint arXiv:2106.06847 **2**(3), 7 (2021)
3. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12299–12310 (2021)
4. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22367–22377 (2023)
5. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X.: Recursive generalization transformer for image super-resolution. In: ICLR (2024)
6. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., Yu, F.: Dual aggregation transformer for image super-resolution. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12312–12321 (2023)
7. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yuan, X., et al.: Cross aggregation transformer for image restoration. Advances in Neural Information Processing Systems **35**, 25478–25490 (2022)
8. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. IEEE Transactions on image processing **14**(12), 2091–2106 (2005)
9. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13. pp. 184–199. Springer (2014)
10. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015)
11. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 391–407. Springer (2016)
12. Driggers, R.G., Friedman, M.H., Nichols, J.: Introduction to infrared and electro-optical systems. Artech House (2012)
13. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3146–3154 (2019)
14. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: 2009 IEEE 12th international conference on computer vision. pp. 349–356. IEEE (2009)
15. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
16. Huang, Y., Jiang, Z., Lan, R., Zhang, S., Pi, K.: Infrared image super-resolution via transfer learning and psrgan. IEEE Signal Processing Letters **28**, 982–986 (2021)
17. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)

18. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2426–2435 (2022)
19. Lei, S., Shi, Z., Zou, Z.: Super-resolution for remote sensing images via local–global combined network. IEEE Geoscience and Remote Sensing Letters **14**(8), 1243–1247 (2017)
20. Li, G., Lv, J., Tian, Y., Dou, Q., Wang, C., Xu, C., Qin, J.: Transformer-empowered multi-scale contextual matching and aggregation for multi-contrast mri super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20636–20645 (2022)
21. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)
22. Li, W., Lu, X., Qian, S., Lu, J., Zhang, X., Jia, J.: On efficient transformer-based image pre-training for low-level vision. arXiv preprint arXiv:2112.10175 (2021)
23. Li, X., Zou, Y., Liu, J., Jiang, Z., Ma, L., Fan, X., Liu, R.: From text to pixels: A context-aware semantic synergy solution for infrared and visible image fusion. arXiv preprint arXiv:2401.00421 (2023)
24. Liang, J., Zeng, H., Zhang, L.: Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5657–5666 (2022)
25. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)
26. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
27. Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., Luo, Z.: Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5802–5811 (2022)
28. Liu, J., Liu, Z., Wu, G., Ma, L., Liu, R., Zhong, W., Luo, Z., Fan, X.: Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8115–8124 (2023)
29. Ma, J., Ma, Y., Li, C.: Infrared and visible image fusion methods and applications: A survey. Information fusion **45**, 153–178 (2019)
30. Ma, J., Yu, W., Liang, P., Li, C., Jiang, J.: Fusiongan: A generative adversarial network for infrared and visible image fusion. Information fusion **48**, 11–26 (2019)
31. Marivani, I., Tsiligianni, E., Cornelis, B., Deligiannis, N.: Joint image super-resolution via recurrent convolutional neural networks with coupled sparse priors. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 868–872. IEEE (2020)
32. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2085–2094 (2021)
33. Patel, H., Chudasama, V., Prajapati, K., Upla, K.P., Raja, K., Ramachandra, R., Busch, C.: Thermisrnet: an efficient thermal image super-resolution network. Optical Engineering **60**(7), 073101–073101 (2021)

34. Pu, M., Huang, Y., Guan, Q., Zou, Q.: Graphnet: Learning image pseudo annotations for weakly-supervised semantic segmentation. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 483–491 (2018)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
36. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
38. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
39. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022)
40. Shamsolmoali, P., Zareapoor, M., Jain, D.K., Jain, V.K., Yang, J.: Deep convolution network for surveillance records super-resolution. Multimedia Tools and Applications **78**, 23815–23829 (2019)
41. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
42. Wang, X., Xie, L., Dong, C., Shan, Y.: Realesrgan: Training real-world blind super-resolution with pure synthetic data supplementary material. Computer Vision Foundation open access **1**, 2 (2022)
43. Xie, L., Wang, X., Chen, X., Li, G., Shan, Y., Zhou, J., Dong, C.: Desra: detect and delete the artifacts of gan-based real-world super-resolution models. arXiv preprint arXiv:2307.02457 (2023)
44. Yang, D., Li, Z., Xia, Y., Chen, Z.: Remote sensing image super-resolution: Challenges and approaches. In: 2015 IEEE international conference on digital signal processing (DSP). pp. 196–200. IEEE (2015)
45. Yang, Y., Li, Q., Yang, C., Fu, Y., Feng, H., Xu, Z., Chen, Y.: Deep networks with detail enhancement for infrared image super-resolution. IEEE Access **8**, 158690–158701 (2020)
46. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)
47. Zhang, J., Zhang, Y., Gu, J., Zhang, Y., Kong, L., Yuan, X.: Accurate image restoration with attention retractable transformer. arXiv preprint arXiv:2210.01427 (2022)
48. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4791–4800 (2021)

49. Zhang, X., Zeng, H., Guo, S., Zhang, L.: Efficient long-range attention network for image super-resolution. In: European Conference on Computer Vision. pp. 649–667. Springer (2022)
50. Zhang, Y., Li, K., Li, K., Fu, Y.: Mr image super-resolution with squeeze and excitation reasoning attention network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13425–13434 (2021)
51. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., Van Gool, L.: Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5906–5916 (2023)
52. Zhao, Z., Deng, L., Bai, H., Cui, Y., Zhang, Z., Zhang, Y., Qin, H., Chen, D., Zhang, J., Wang, P., et al.: Image fusion via vision-language model. arXiv preprint arXiv:2402.02235 (2024)
53. Zhao, Z., Xu, S., Zhang, J., Liang, C., Zhang, C., Liu, J.: Efficient and model-based infrared and visible image fusion via algorithm unrolling. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1186–1196 (2021)
54. Zhao, Z., Zhang, J., Gu, X., Tan, C., Xu, S., Zhang, Y., Timofte, R., Van Gool, L.: Spherical space feature decomposition for guided depth map super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12547–12558 (2023)
55. Zhao, Z., Zhang, J., Xu, S., Lin, Z., Pfister, H.: Discrete cosine transform network for guided depth map super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5697–5707 (2022)