IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 October 2023, accepted 10 December 2023, date of publication 19 December 2023,
date of current version 4 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3344830

RESEARCH ARTICLE

Recurrent Large Kernel Attention Network for Efficient Single Infrared Image Super-Resolution

GANGPING LIU[1,2], (Member, IEEE), SHUAIJUN ZHOU[3], XIAXU CHEN[1,2], WENJIE YUE[1,2], AND JUN KE[1,2], (Member, IEEE)

[1]School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China
[2]Key Laboratory of Photo-Electronic Imaging Technology and System, Ministry of Education of China, Beijing 100081, China
[3]Beijing Aerospace Automatic Control Institute, China Aerospace Science and Technology Corporation, Beijing 100070, China

Corresponding author: Jun Ke (jke@bit.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61675023 and Grant U2241275, and in part by the Fundamental Research Funds for the Central Universities under Grant 2022CX02002.

ABSTRACT Infrared imaging has broad and important applications. However, the infrared detector manufacture technique limits the detector resolution and the resolution of infrared images. In this work, we design a Recurrent Large Kernel Attention Neural Network (RLKA-Net) for single infrared image super-resolution(SR), and then demonstrate its superior performance. Compared to other SR networks, RLKA-Net is a lightweight network capable of extracting spatial and temporal features from infrared images. To extract spatial features, we use multiple stacked Recurrent Learning Units (RLUs) to expand the network's receptive field, while the large kernel attention mechanism in RLUs is used to obtain attention maps at various granularity. To extract temporal features, RLKA-Net uses the recurrent learning strategy to keep persistent memory of extracted features, which contribute to more precise reconstruction results. Moreover, RLKA-Net employs an Attention Gate (AG) to reduce the number of parameters and expedite the training process. We demonstrate the efficacy of the Recurrent Learning Stages (RLS), Large Kernel Attention Block (LKAB), and Attention Gate mechanisms through ablation studies. We test RLKA-Net on several infrared image datasets. The experimental results demonstrate that RLKA-Net presents state-of-the-art performance compared to existing SR models. The code and models are available at https://github.com/ZedFm/RLKA-Net.

INDEX TERMS Infrared image super-resolution, image processing, recurrent neural network, attention mechanism.

I. INTRODUCTION

Infrared imaging can provide valuable thermal information about objects, thus having broad application areas such as medical diagnostics, security surveillance, and defense. However, compared to visible images, manufacturing high-resolution infrared detectors is more challenging [1]. Hence, infrared images often present comparable low resolution and visual quality. Thus, image super-resolution (SR) is a critical technique that can reconstruct high-resolution infrared images from low-resolution (LR) measurements to recover the thermal details of an object. Moreover, image super-resolution can also enhance the performance of detection and recognition using LR infrared images [2].

Single-image super-resolution is a classical computer vision problem, which reconstruct a high-resolution image from an LR image by solving an ill-posed inverse problem [3], [4], [5]. As deep learning technology grows rapidly, a large number of convolutional neural networks (CNNs) have been designed for image SR and achieved remarkable performance. However, research on infrared image SR methods is few. Moreover, the state of the art (SOTA) CNN-based SR algorithms designed for visible images do not perform well on infrared images as shown in Fig. 1(left), even retrained on infrared images. This is due to the imaging mechanisms of infrared detectors significantly differ from visible detectors, resulting in distinct visual characteristics between infrared and visible images.

In this paper, we design a lightweight architecture named Recurrent Large Kernel Attention Neural Network

The associate editor coordinating the review of this manuscript and approving it for publication was Ángel F. García-Fernández.

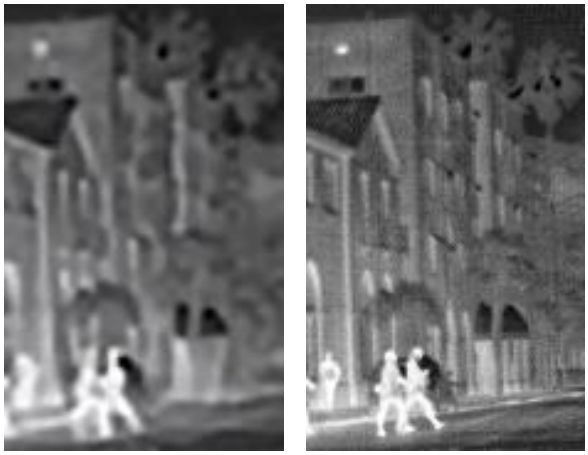

FIGURE 1. CNN-based SR model specifically designed for visible images to reconstruct high-resolution infrared images typically yields unsatisfactory performance. Scenario: image 00719 from FLIR [6] dataset. Left: RCAN [7]. Right: our RLKA-Net (×4).

(RLKA-Net) for infrared image SR, which consists of Recurrent Learning Stage (RLS), Large Kernel Attention Block (LKAB), and Attention Gate (AG). Generally, the most common way to enhance the performance of CNN is to unfold the network and deepen the structure in the temporal or the spatial dimension. In RLKA-Net, we use Recurrent Learning Stages (RLS) to unroll the network in the temporal dimension, and use multiple stacked Large Kernel Attention Blocks (LKAB) to unfold the network in the spatial dimension. With RLS and LKAB, the extracted temporal and spatial features help RLKA-Bet present superior SR performance. Furthermore, the Large Kernel Attention Block is based on the large kernel decomposition assumption, which reduces the parameters in convolutional kernels and improves the calculation efficiency. Additionally, the Attention Gate facilitates faster convergence during the training process. As shown in Fig 2, our RLKA-Net achieves state-of-the-art performance in infrared image super-resolution comparing to other SR models.

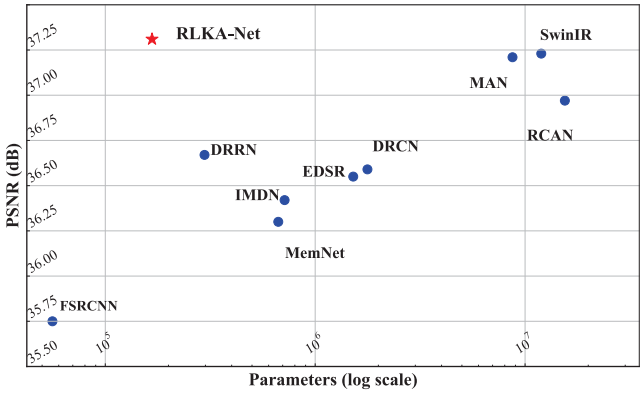

FIGURE 2. Trade-off between parameters and performance(PSNR) on CVC09 for ×4 infrared SR.

To summarize, our contributions are listed as follows:

1) We design the Recurrent Large Kernel Attention Neural Network (RLKA-Net) for single infrared image super-resolution. RLKA-Net is a lightweight model that efficiently reconstructs high-resolution infrared images, achieving state-of-the-art performance in several infrared datasets.

2) We propose the Recurrent Learning Stages (RLS) to extract temporal features for infrared SR tasks. RLS not only keeps persistent memory of the extracted features in training steps but also expands the network's receptive field.

3) We design the Large Kernel Attention Block and Attention Gate to obtain multi-scale attention maps and reduce computational complexity, which ensures RLKA-Net a lightweight model for efficient infrared image SR.

The remainder of this paper is organized as follows. In Section II, we briefly review the related works about infrared image SR method. In Section III, we discuss the architecture of our Recurrent Large Kernel Attention Network, along with RLS, LKAB, and AG. In Section IV, we present the details of the infrared image datasets and the training configurations, as well as the ablation analysis and the experimental results. Finally, we draw the conclusion and discuss the future directions of our work in Section V.

II. RELATED WORK

A. LIGHT-WEIGHT SINGLE IMAGE SUPER-RESOLUTION (SISR) MODELS

In SISR, SRCNN [8] is a pioneering work that introduces deep learning into the field. Early deep learning SISR works primarily focus on deeper structures to improve feature extraction abilities. In VDSR [9], a 20-layer network with 665K parameters is used. Then another dense-connected network named DRCN [10] with 1,774K parameters is designed. Later, Zhang et al. [11] design the residual dense network (RDN, 128 layers) for SISR. Group from Northeastern University [7] design the very deep residual channel attention networks (RCAN, 400 layers) to extract low-frequency information. Furthermore, based on residual learning, dense connections, and attention mechanisms, more networks with a great amount of parameters are proposed [5], [12], [13], [14], [15]. Although these networks present better results, they use a large amount of parameters, make computational cost expensive, thus limit their performance in real-time reconstruction applications.

To reduce computational cost, Dong et al. [16] redesign SRCNN by directly extracting features from LR images instead of pre-upsampling these images first. This strategy avoids processing massive data at the beginning and has been widely applied in SISR networks [17], [18], [19]. Additionally, depth-wise separable convolutions [20], [21], self-calibrated convolution [22], and group convolution [23] have been employed to make deep networks smaller. In BSRN [24], depth-wise separable convolution replaces standard convolution to reduce computational cost, while two types of attention schemes are also designed for accurate

reconstructions. In the network PAN [25], a pixel attention (PA) scheme with self-calibration for efficient SSIR is designed. In CARN-M [26], the group convolution module presents remarkable results compared to traditional time-consuming models. In addition, in network RFDN [27], the receptive field is expended while the network uses a small number of parameters. Based on adaptive cropping and information multi-distillation, the lightweight model IMDN [28] presents distinguished performance for SSIR. VapSR [29] is a vast-receptive-field pixel attention network with 342K parameters. Experimental results demonstrate this performance of the well-designed attention.

B. NETWORKS FOR INFRARED IMAGE SR(I^2SR)

Although numerous deep learning models have been designed for SISR, a few of them are for infrared images. Researchers focus more on fusion models for visible and infrared images [30], [31], [32]. On the other hand, the performance of applying visible SR models to infrared images is often unsatisfied.

In 2018, He et al. [33] present the first work of designing a deep neural network with a cascaded architecture for I^2SR. In paper [34], Rivadeneira et al. construct a comprehensive thermal image dataset composed with infrared images at various resolutions, and employed a CycleGAN architecture for I^2SR. Later, Prajapati et al. [35] design a CNN-based architecture, referred to as ChaSNet, for the PBVS-2021 Thermal SR Challenge. In ChaSNet, the concept of channel splitting is used to improve the reconstruction quality. Furthermore, Batchuluun et al. [36] split the thermal image into smaller region images for deblurring and then applied a newly proposed generative adversarial network (GAN) for infrared SR(I^2SR). Based on transfer learning, Huang et al. [37] design the progressive generative adversarial network (PSRGAN). In addition, Yang et al. [38] utilize visible images as a complementary source, and design a spatial attention residual neural network for I^2SR reconstruction. Although these I^2SR models achieve nice performance, few of them focus on a lightweight I^2SR model.

C. RECURRENT NEURAL NETWORKS

Most of the SR models mentioned earlier only unfold the network in the depth dimension. In contrast, recurrent neural networks (RNNs) unroll the structure in the time dimension, thus help capturing temporal features. In 2014, Cho et al. [39] explore RNN's potential in feature representation by applying the RNN Encoder-Decoder for statistical machine translation. Mao et al. [40] design a multi-model recurrent neural network (m-RNN) to learn visual information more effectively. Furthermore, in Mem-Net [41] and DRRN [42], recursive learning blocks with a multi-path structure are designed to improved SR performance. The network is unrolled in the temporal dimension without adding more parameters. DSRN [43] is a dual-state recurrent network for SISR. In FBRNN [44], a deep feedback architecture with multiple recurrent and residual blocks is used.

Unrolling a network in the temporal dimension enables the model to capture more detailed features from different learning periods. In addition, by incorporating suitable attention mechanism in the recurrent learning stage, the captured detailed features can lead to a better SR result [45]. Thus, in this work, we also use the recurrent learning strategy in our lightweight I^2SR network.

D. ATTENTION SCHEMES FOR IMAGE SR

The attention mechanism can be regarded as a discriminative selection process, that guides networks to focus on informative regions for a specific task. In the following, we summarize related works using three attention mechanisms: channel attention, spatial attention, and multi-attention mechanism.

1) CHANNEL ATTENTION

In 2018, Hu et al. [46] design a compact lightweight structure called Squeeze-and-Excitation (SE) network, which can adaptively recalibrate channel-wise feature maps by calculating the inter-channel dependencies. This is the first work on channel attention. Subsequently, in RCAN [7], channel attention based attention residual blocks are used for SISR to achieve superior reconstructions. Later, SAN [13] is designed to to capture channel-wise features according to second-order statistics. Generally, extracting channel-wise features using the channel attention mechanism can help a backbone network to improve its representation ability.

2) SPATIAL ATTENTION

Spatial attention focus on the relevance between specific regions which contribute to better restoration. In SeLNet, Choi and Kim [47] use spatial attention to achieve better reconstructions and low computational complexity simultaneously. Later, in HAN [48], a channel-spatial attention module is designed to learn the channel and spatial inter-feature dependencies in each layer. Further works tend to combine different attention mechanisms, or multi-attention mechanism, to learn more informative features for a better SR result.

3) MULTI-ATTENTION MECHANISM

Based on vision transformers, in IPT [49] and SwinIR [50], multi-head self-attention is used to capture long-range dependence. Furthermore, a hybrid attention scheme combining self-attention and channel-wise attention is used in HAT [51] for SISR. On the other hand, in 2022 Guo et al. [52] indicate that a large kernel convolution can be divided into three components: a spatial local convolution (depth-wise convolution), a spatial long-range convolution (depth-wise dilation convolution), and a channel convolution. Then, they propose a novel linear attention mechanism named Large Kernel Attention (LKA), which keeps the advantages of self-attention and convolution, such as preserving a long-range dependence, strong capability to extract structure information, powerful

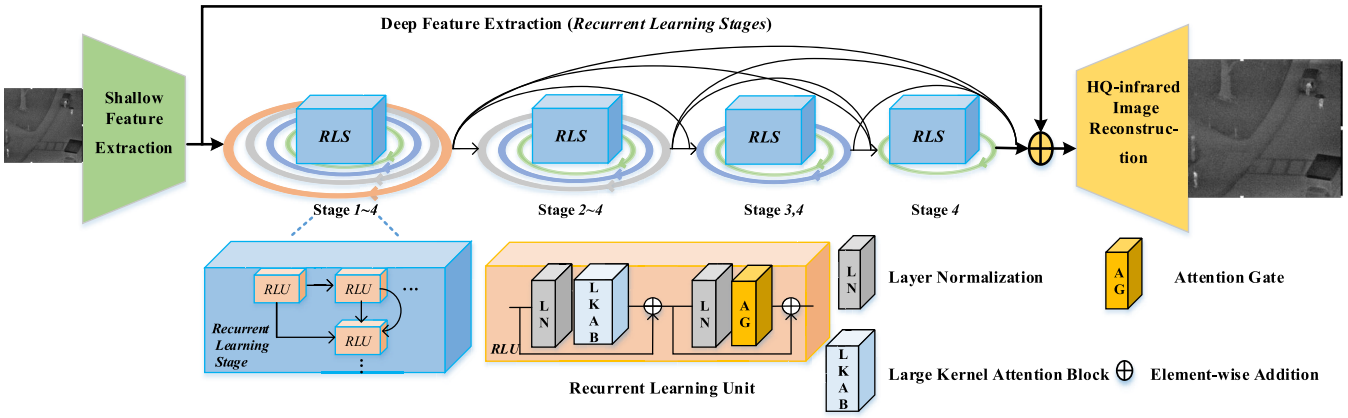

FIGURE 3. Architecture of the Recurrent Large Kernel Attention Network (RLKA-Net),number of the recurrent learning stage is set to 4.

self-adaptability, and having less parameters. Latter based on large kernel attention, MAN [53] designs Multi-Scale Large Kernel Attention (MLKA), which combines LKA with a multi-scale mechanism to build various-range correlations with low computational complexity.

III. METHODOLOGY

In this paper, we introduce a Recurrent Large Kernel Attention Network (RLKA-Net) for infrared image super-resolution (SR). In the RLKA-Net, features are extracted from infrared images in both spatial and temporal dimensions with high efficiency. Details of RLKA-Net is presented in this section.

As depicted in Fig.3, our architecture comprises three components: the Shallow Feature Extraction Module (SFM), the Deep Feature Extraction Module (DFM) based on several Recurrent Learning Stages (RLS), and the high-quality infrared image reconstruction module.

Given a low-resolution (LR) infrared image $I_{LR} \in R^{3*H*W}$, SFM is used to extract shallow features $F_s \in R^{C*H*W}$ by using a simple 3×3 convolution function $f_{SFM}(\cdot)$, as illustrated in Eq.1:

$$F_s = f_{SFM}(I_{LR}) \tag{1}$$

Subsequently, the shallow feature F_s is fed into the DFM for further feature extraction. DFM is denoted as $f_{DFM}(\cdot)$ as shown in Eq. 2. F_d represents the extracted deep features. Utilizing multiple Recurrent Learning Stages (RLS), the DFM can efficiently extract diverse deep features with fewer parameters to be optimized. Each RLS composes several recurrent learning units (RLUs). the *l-th* RLU is represented as $RLU_l(\cdot)$. Each RLU contains a large kernel attention block (LKAB), an attention gate(AG) and two normalization layer, as shown in Fig. 3. Details about RLS, LKAB and AG are presented in the following subsections. The notation $[\![\cdot]\!]$ represents the feature refinement procedure of *l-th* RLU.

$$\begin{aligned} F_d &= f_{DFM}(F_s) \\ &= [\![RLU_l(x_1), RLU_l(x_2) \ldots RLU_l(x_n)]\!] \end{aligned} \tag{2}$$

For the reconstruction module, some SISR methods employ up-sampling operations before skip connections, which introduce additional computation. But the improvement in reconstructions is limited. To construct a lightweight SR model, we add the initial shallow feature F_s with F_d for the final reconstruction module $f_{recon}(\cdot)$ to obtain a high-resolution infrared image $I_{SR} \in R^{3*H*W}$. The final reconstruction operation is formulated as Eq. 3.

$$I_{SR} = f_{recon}(F_s + F_d) \tag{3}$$

Regarding optimization, we employ the commonly used L_1 loss for a fair comparison with several well-known SR models. Assuming the input batch contains N infrared images, i.e., I_i^{LR}, I_i^{HR} with $\{i = 1, \cdots, N\}$, the training process aims to minimize the L_1 loss as illustrated in Eq. 4,

$$L_1(\Theta) = \frac{1}{N}\sum_{i=1}^{N} ||f_{RLKA-N}(I_i^{LR}) - I_i^{HR}||_1, \tag{4}$$

where f_{RLKA-N} represents the network, and Θ denotes its optimizable parameters.

A. RECURRENT LEARNING STAGE (RLS)

The recurrent learning stage (RLS) can perform feature learning both in temporal and spatial dimensions. On one hand, RLS recursively learn temporal features. On the other hand, RLS enlarges the receptive field along the depth direction of the network architecture. Consequently, RLS can help a network having more accurate reconstructions.

Fig. 4 presents the schematic diagram of the RLS strategy. We define the symbol $O\{\cdot\}_j^i$ as the output of the *j-th* RLU in the *i-th* period in the time domain. Then the procedure of RLS can be illustrated as follows:

I. In stage 1, both the recurrent time and the network depth are set to one as shown in Fig. 4. The features extracted from previous module are sent to single one recurrent learning unit (RLU). The output of this stage is

$$O\{RLS\}_1^1 = RLU_1(F_s) \tag{5}$$

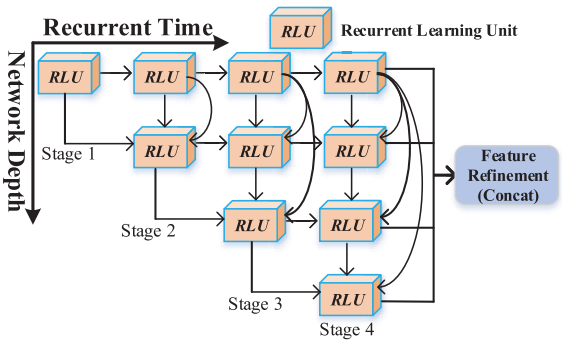

FIGURE 4. The schematic diagram of recurrent learning stage strategy.

where $O\{RLS\}_1^1$ represent the output of stage 1. RLU_1 is the operation function of the first RLU.

II. In stage 2, the network architecture is extended in both time and depth dimensions. As shown in Fig. 4, we utilize the second RLU to extend the receptive fields in the spatial dimension. Additionally, features from the previous recursive period are preserved to learn more temporal details.

$$\begin{aligned} O\{RLS\}_1^2 &= RLU_1(O\{RLS\}_1^1) \\ O\{RLS\}_2^2 &= RLU_2([O\{RLS\}_1^1, O\{RLS\}_1^2]) \end{aligned} \quad (6)$$

Eq. 6 defines that $O\{RLS\}_2^2$ integrates the feature extracted in stage 1 and stage 2, reserving more spatial and temporal information.

III. In stage 3, the network architecture is further extended in both dimensions. The procedure is similar to stage 2. Eq. 7 defines the outputs of the RLS units.

$$\begin{aligned} O\{RLS\}_1^3 &= RLU_1(O\{RLS\}_1^2) \\ O\{RLS\}_2^3 &= RLU_2([O\{RLS\}_1^3, O\{RLS\}_2^2]) \\ O\{RLS\}_3^3 &= RLU_3([O\{RLS\}_1^3, O\{RLS\}_2^3, \\ &\qquad O\{RLS\}_2^2]) \end{aligned} \quad (7)$$

IV. Stage 4 is the final recurrent stage in our network.

$$\begin{aligned} O\{RLS\}_1^4 &= RLU_1(O\{RLS\}_1^3) \\ O\{RLS\}_2^4 &= RLU_2([O\{RLS\}_1^4, O\{RLS\}_2^3]) \\ O\{RLS\}_3^4 &= RLU_3([O\{RLS\}_1^4, O\{RLS\}_2^4, \\ &\qquad O\{RLS\}_3^3]) \\ O\{RLS\}_4^4 &= RLU_3([O\{RLS\}_1^4, O\{RLS\}_2^4, \\ &\qquad O\{RLS\}_3^4, O\{RLS\}_3^3]) \end{aligned} \quad (8)$$

As shown in Eq. 8, $O\{RLS\}_4^4$ integrated all the feature from previous recurrent learning period, benefiting from the large receptive field in depth dimension and the persistent memory from recurrent learning stages.

V. In summary, the outputs of the *l-th* RLS can be summarized by the formula Eq. 9,

$$\begin{aligned} O\{RLS\}_1^l &= RLU_1(O\{RLS\}_1^l) \\ &\vdots \\ O\{RLS\}_N^l &= RLU_N([O\{RLS\}_1^l, O\{RLS\}_2^l \\ &\qquad \ldots, O\{RLS\}_{N-1}^l, O\{RLS\}_N^l]) \end{aligned} \quad (9)$$

Note that, by ablation experiments, we find that our DFE model presents the best performance in terms of reconstruction quality and speed when we set the number of RLS to 4. Thus, we use 4 RLS stages in DFE.

B. RLU WITH LARGE KERNEL ATTENTION BLOCK (LKAB)

As illustrated in Fig. 3, given an input feature map $X \in R^{H\times W\times C}$, the procedure of a RLU can be formulated as Eq. 10,

$$\begin{aligned} N &= LN(X) \\ X &= X + \lambda_1 f_{LKAB}(N) \\ N &= LN(X) \\ X &= X + \lambda_2 f_{AT}(N) \end{aligned} \quad (10)$$

where $LN(\cdot)$ is layer normalization, λ_1 and λ_2 are the learnable weight factors, $f_{LKAB}(\cdot)$ and $f_{AT}(\cdot)$ are the operation of large kernel attention block (LKAB) and attention gate (AG) modules. The point-wise convolution utilized for preserving dimensions is omitted in Eq. 10. We present the details of LKAB and AT in following.

1) LARGE KERNEL DECOMPOSITION

As well-known, large kernel convolutions typically result in a significant amount of computational overhead and parameters, making optimization challenging. However, it has been demonstrated [52] that the standard kernel convolution can be decomposed into three parts: depth-wise convolution (spatial local convolution), depth-wise dilation convolution (spatial long-range convolution), and channel convolution (1×1 convolution). For instance, given a $K \times K$ large kernel convolution, it can be decomposed into a $\lceil \frac{K}{d} \rceil \times \lceil \frac{K}{d} \rceil$ depth-wise dilation convolution (DWD Conv) $f_{DW}(\cdot)$ with the dilation d, a $(2d-1)\times(2d-1)$ depth-wise convolution (DW Conv) $f_{DWD}(\cdot)$, and a 1×1 point-wise convolution (PW-Conv) $f_{PW}(\cdot)$. Thus, this large kernel convolution can be illustrated as Eq. 11,

$$LKA(X) = f_{PW}(f_{DWD}(f_{DW}(X))), \quad (11)$$

where X is the input feature map $X \in R^{H\times W\times C}$, and the output is denoted as $LKA(\cdot)$. Note that, a large kernel convolution is also referred to as a large kernel attention.

In terms of computational complexity, assuming the input and output features having the same size $H \times W \times C$, we denote the number of parameters as $P(K, d)$ and the floating-point operations (FLOPs) as $F(K, d)$, where d

represents the dilation rate, and K is the size of the large kernel. Then, we have [53]

$$P(K,d) = C(\left[\frac{K}{d}\right]^2 \times C + (2d-1)^2) + C^2 \tag{12}$$

and

$$F(K,d) = P(K,d) \times H \times W. \tag{13}$$

From Eq.12 and Eq.13, we can find that by reducing the parameter d, $P(K,d)$ and $F(K,d)$ decrease quickly. Thus, the large kernel decomposition can save a substantial amount of computing costs by selecting an appropriate dilation rate d.

2) LARGE KERNEL ATTENTION BLOCK (LKAB)

Large kernel attention (LKA) has been applied to visible visual tasks [52], [53] to achiev remarkable performance. We employ multi-scale LKA to form a large kernel attention block for feature extraction in I^{2}SR, as shown in Fig. 5:

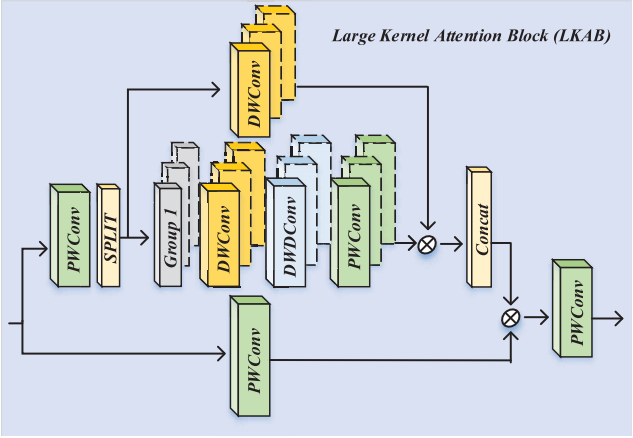

FIGURE 5. Schematic diagram of the multi-scale large attention block (LKAB).

I. Given the input feature map $X \in R^{H\times W\times C}$, LKAB initially splits X into n groups, X_1, X_2, $\cdots$, and $X_n \in R^{H\times W\times C/n}$.
II. For each group X_i, an independent LKA module is used generate the homogeneous scale attention map LKA_i.
III. A spatial attention (SA_i) module is used for feature aggregation as show in Eq. 14,

$$LKAB(X) = SA(X_i) \otimes LKA_i(X_i). \tag{14}$$

where $SA(X_i)$ represents the i-th spatial attention module, and LKA_i denotes the corresponding LKA module. This is due to that excessive dilation and partition operations have been found to make visible SR results [53] having block artifacts.

3) ATTENTION GATE (AG)

Inspired by the gate unit in transformer blocks used to improve feature representation [51], we construct the attention gate (AG) by integrating the SA module and depth-wise convolution, as illustrated in Fig. 6. The Attention Gate (AG) serves a dual function: it incorporates an adaptive gating mechanism into the recurrent learning unit, enabling the model to selectively attend to related spatial information. At the same time, AG can help to reduce the number of parameters thus the computational cost, thereby augmenting the network's efficiency in reconstruction, which is critical for conducting a lightweight I^{2}SR model.

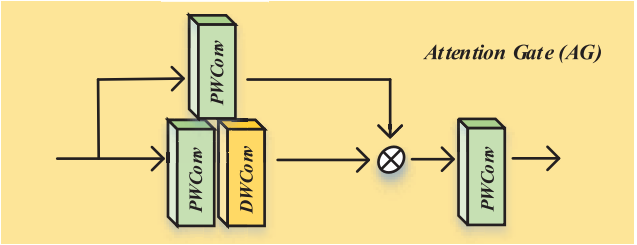

FIGURE 6. Detail about the attention gate (AT).

IV. EXPERIMENTS

In this section, we demonstrate our network with a series of experiments.

A. DATASETS AND IMPLEMENTATION DETAILS

Datasets are very important for learning-based reconstruction methods. However, most public datasets for infrared images are on tasks such as fusion, classification, and detection. Open-source datasets for I^{2}SR are rare. To obtain HR and LR image pairs for training and evaluation, we utilize the images from datasets FLIR [6], OSU [54], CVC09 [55], and LLVIP [56] as HR samples. Fig. 7 presents typical scenes from these datasets. We then apply bicubic or Lanczos [57] downsampling with different levels of noise to obtain the LR samples. Furthermore, we use peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) to evaluate the infrared reconstruction quality.

In the training phase, we train the network for each dataset. The OSU, FLIR, CVC09, and LLVIP datasets contain 270, 2000, 4000, and 6,000 images, respectively. We also do dataset augmentation by horizontal flipping and rotating each pair of IR images by 90, 180, and 270 degrees.

We set the batch size and patch size to 16 and 64 $\times$ 64, respectively. The Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\varepsilon = 10^{-8}$, is used for model optimization. The initial learning rate is set at 2×10^{-4} and follows a cosine annealing schedule over 200,000 iterations. All experiments utilize the PyTorch framework and are executed on two Nvidia RTX 3090 GPUs.

In the testing phase, we utilize totally 1227 images, 27 from OSU, 200 from FLIR, 400 from CVC09, and 600 from LLVIP. The Y channel data of the reconstructed images are used for quality evaluation.

Additionally, we retrain the SR models designed for visible band on infrared datasets to generate the test results presented later in this manuscript.

TABLE 1. Experiments result of different numbers of recurrent learning stage.

#Stage	Params	Multi-Adds	Dataset							
			OSU [54]		FLIR [6]		CVC09 [55]		LLVIP [56]	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
2	50.1 K	2.91 G	26.67	0.5743	26.34	0.5723	31.75	0.7844	34.38	0.8271
3	100.3 K	5.82 G	28.94	0.5996	29.07	0.6016	34.31	0.8216	37.92	0.9345
4	167.1 K	9.70 G	31.06	0.6504	31.12	0.6506	37.31	0.9294	42.26	0.9890
5	514.5 K	19.9 G	31.15	0.6541	31.23	0.6549	37.67	0.9306	42.31	0.9895
6	872.6 K	31.6 G	**31.21**	**0.6547**	**31.32.00**	**0.6573**	**37.79**	**0.9311**	**42.33**	**0.9897**

* The scale of infrared image super-resolution is 4.

TABLE 2. Experiments result of different unfolding method.

Unfolding Method	Dataset							
	OSU [54]		FLIR [6]		CVC09 [55]		LLVIP [56]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RLKA-Net-T	27.76	0.5881	27.94	0.5946	31.63	0.7906	35.56	0.8574
RLKA-Net-S	29.24	0.6044	29.87	0.6117	33.79	0.8071	39.49	0.9461
RLKA-Net-TS	**31.06**	**0.6504**	**31.12**	**0.6506**	**37.31**	**0.9294**	**42.26**	**0.9890**

* The number of Recurrent Learning Stage is set to 4, the scale of SR is 4.
* RLKA-Net-TS denotes the SR model that is unfolded in both temporal and spatial dimensions.
* RLKA-Net-TS is the standard model we propose for infrared image super-resolution.

FIGURE 7. Typical scenes from the datasets, top-left, CVC09 [55]; bottom-left, LLVIP [56]; top-right, FLIR [6]; and bottom-right, OSU [54].

B. ABLATION ANALYSIS

Here, we present the ablation analysis of our Super-Resolution (SR) model. Several experiments have been designed to demonstrate the efficacy of the recurrent learning strategy (RLS), the large kernel attention block (LKAB), and the attention gate (AG) components. These experiments aim to systematically evaluate the individual contributions of these modules to the performance of the overall model. Furthermore, we will investigate the potential synergistic effects that may arise from the interaction between these components. This analysis will enable a deeper understanding of our designed SR model, paving the way for future improvements and potential applications.

1) ABLATION EXPERIMENT ON RLS

As previously discussed, we use RLSs for efficient infrared image SR. We set the large kernel size as (13 × 13). A large kernel convolution is decomposed into a (5 × 5) depth-wise convolution (DW Conv), a 5 × 5 depth-wise dilation convolution (DWD Conv) with dilation rate 3, and a point-wise convolution (PW Conv).

a: THE OPTIMAL NUMBER OF RLS

First, we need to determine the optimal number of RLS. The experimental results are presented in Table 1. When we set the number of RLS to 2, the modal has fewer parameters and relatively low computational complexity. However, the SR reconstruction quality is not high. The PSNR and SSIM values are pretty low. As we increase the number of RLS, the reconstruction quality, the PSNR and SSIM values are improved. However, this also leads to a significant rise in the number of the model's parameters and the computational complexity. Notably, When the number of RLS is in increased from 4 to 6, the PSNR only improves 0.15dB, while the number of the parameters increase 800k. Consequently, to balance computational complexity and reconstruction performance, we set the number of RLS to 4.

b: DISCUSS ON UNFOLDING METHOD

We evaluate three unfolding models on multiple datasets for I^2SR, unfolding in the temporal, spatial, and in both temporal and spatial dimensions. The results are presented in Table 2. We observe that, based on the PSNR and SSIM values, unfolding in the temporal domain alone presents relatively low performance. The PSNR and SSIM values for unfolding in the spatial domain are larger. But the best performance is observed when using unfolding in both dimensions.

2) ABLATION EXPERIMENT ON LKAB

In order to demonstrate the effectiveness of LKAB, we compare the LKAB module with other typical modules used in SR, including the Enhanced Residual Block (ERB) in EDSR [5], the pixel attention block (PAB) in PAN [25],

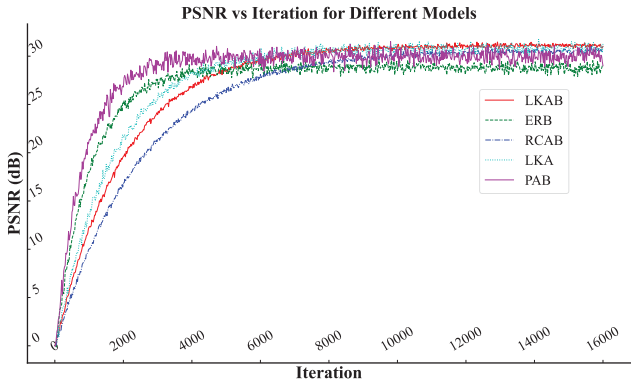

FIGURE 8. PSNR vs Iteration of different SR module.

TABLE 3. Experiments result of different super-resolution module.

SR Module	Params	FLIR [6]	
		PSNR	SSIM
ERB [5]	1518 K	28.82	0.6278
PAN [25]	98 K	29.98	0.6392
RCAB [7]	5200 K	30.62	0.6448
LKA [52]	226 K	30.82	0.6473
LKAB (Ours)	167 K	**31.06**	**0.6504**

* Dataset is FLIR ×4.

TABLE 4. Experiments result of with or without attention gate.

Dataset	SR model			
	RLKA-NeT-wo		RLKA-Net	
	PSNR	SSIM	PSNR	SSIM
OSU [54]	30.11	0.6451	**31.06**	**0.6504**
FLIR [6]	30.42	0.6477	**31.12**	**0.6506**
CVC09 [55]	36.97	0.9037	**37.31**	**0.9294**
LLVIP [56]	40.14	0.9641	**42.26**	**0.9890**

* RLKA-Net-wo represents RLKA-Net without Attention Gate.

TABLE 5. Discussion on computational complexity.

Mechanism	Params	Multi-adds	PSNR	SSIM
LKA	167.1 K	9.7 G	**31.06**	**0.6504**
LKA-wo	488.3 K	28.2 G	30.57	0.6437

* LKA-wo represents without large kernel decomposition, dataset is FLIR ×4.
*

the residual channel Attention block (RCAB) in RCAN [7], and the large kernel attention (LKA) in VAN [52]. For a fair comparison, we use the same architecture that unfolds the network in both temporal and spatial dimensions, only replacing LKAB with the aforementioned SR modules. The experimental results are illustrated in Fig. 8 and Table. 3.

We observe that compared to other attention-based modules, the ERB performs worst in terms of PSNR and SSIM. Although the PAB has the lowest number of parameters, the reconstruction performance is not good. The RCAB module has the largest number of parameters. But it does not achieve the best reconstruction results. Our LKAB module presents the highest PSNR and SSIM values, while using the second smallest number of parameters.

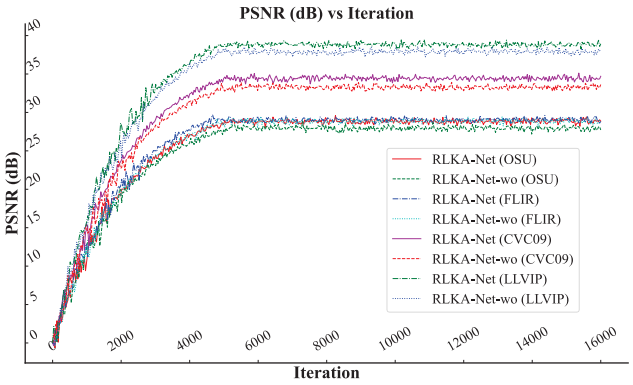

FIGURE 9. PSNR vs Iteration of RLKA-Net and RLKA-Net-wo in different datasets.

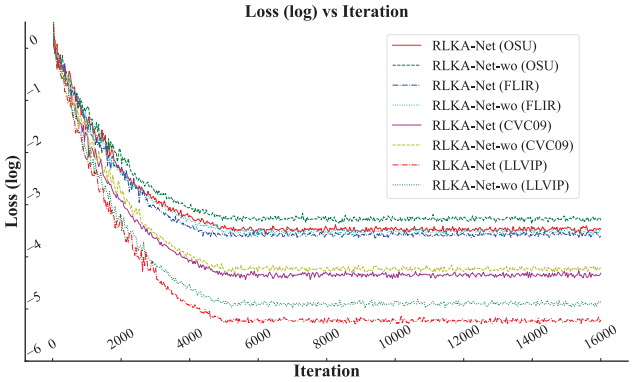

FIGURE 10. Loss vs Iterlation of RLKA-Net and RLKA-Net-wo in different datasets.

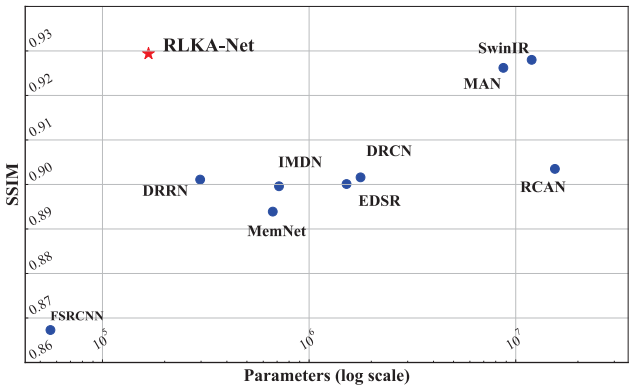

FIGURE 11. Trade-off between parameters and performance(SSIM) on CVC09 for ×4 infrared SR.

3) ABLATION EXPERIMENT ON AG

We also conduct experiments on the attention gate (AG) in the model using the aforementioned datasets. For clarity, the model using AG is referred to as RLKA-Net, while the model without AG is named as RLKA-Net-wo. We set the number of RLS as 4. The PSNR results are summarized in Table. 4. The PSNR and Loss vs. iteration curves are displayed in Fig. 9 and Fig. 10. It is clear that RLKA-Net with AG consistently outperforms RLKA-Net-wo in terms of PSNR.

TABLE 6. Quantitative comparison (average PSNR/SSIM) with other SR models in different infrared datasets.

Method	Scale	Params	Dataset							
			OSU [54]		FLIR [6]		CVC09 [55]		LLVIP [56]	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	× 2	\	29.35	0.6717	29.72	0.6981	38.77	0.9116	39.09	0.9778
FSRCNN [16]	× 2	56 K	30.98	0.6974	31.93	0.7461	39.65	0.9341	39.78	0.9806
EDSR [5]	× 2	1518 K	32.34	0.7445	32.57	0.7803	41.70	0.9472	40.51	0.9821
DRCN [10]	× 2	1774 K	32.23	0.7387	32.31	0.7714	41.58	0.9401	40.49	0.9816
DRRN [42]	× 2	297 K	32.29	0.7397	32.53	0.7726	41.63	0.9415	40.53	0.9827
MemNet [41]	× 2	667 K	32.21	0.7382	32.46	0.7765	41.67	0.9459	40.41	0.9814
RCAN [7]	× 2	15472 K	32.46	0.7457	33.97	0.8163	42.06	0.9509	41.56	0.9843
IMDN [28]	× 2	715 K	32.58	0.7513	33.67	0.8066	41.89	0.9458	40.64	0.9824
MAN [53]	× 2	8712 K	32.80	0.7606	34.89	0.8419	42.23	0.9541	38.92	0.9773
SwinIR [50]	× 2	11942 K	32.93	0.7618	34.82	0.8417	42.21	0.9537	41.87	0.9849
RLKA-Net (Ours)	× 2	167 K	**33.11**	**0.7621**	**34.94**	**0.8457**	**42.45**	**0.9665**	**42.19**	**0.9895**
Bicubic	× 4	\	27.85	0.5744	28.02	0.5871	34.04	0.8429	37.98	0.9539
FSRCNN [16]	× 4	56 K	29.16	0.5998	28.89	0.6016	34.62	0.8673	38.53	0.9783
EDSR [5]	× 4	1518 K	30.57	0.6398	29.78	0.6207	36.55	0.9001	40.83	0.9839
DRCN [10]	× 4	1774 K	30.49	0.6346	29.63	0.6192	36.59	0.9016	39.98	0.9802
DRRN [42]	× 4	297 K	30.52	0.6353	29.74	0.6201	36.67	0.9011	40.64	0.9827
MemNet [41]	× 4	667 K	30.39	0.6298	29.57	0.6169	36.30	0.8939	40.59	0.9819
RCAN [7]	× 4	15472 K	30.62	0.6407	30.49	0.6385	36.97	0.9035	41.63	0.9845
IMDN [28]	× 4	882 K	30.49	0.6371	30.53	0.6419	36.42	0.8996	41.45	0.9839
MAN [53]	× 4	9431 K	30.98	0.6478	31.07	0.6489	37.21	0.9262	38.82	0.9765
SwinIR [50]	× 4	14367 K	30.70	0.6439	30.86	0.6472	37.26	0.9287	41.83	0.9867
RLKA-Net (Ours)	× 4	193 K	**31.06**	**0.6504**	**31.12**	**0.6506**	**37.31**	**0.9294**	**42.06**	**0.9890**

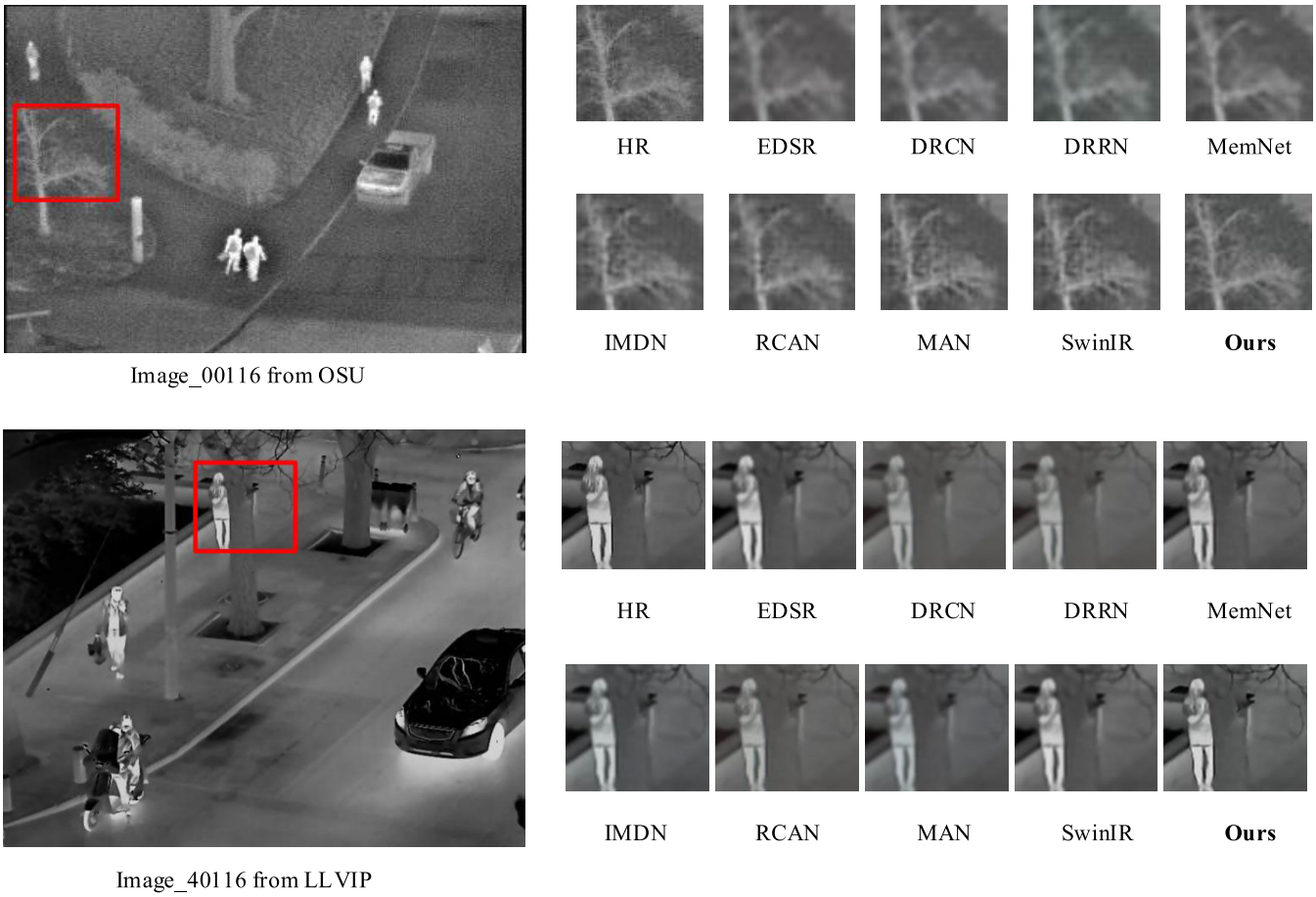

FIGURE 12. Visual comparisons of our RLKAN with other SR methods on OSU [54] and LLVIP [56] dataset. (×4).

4) DISCUSSION ON COMPUTATIONAL COMPLEXITY

To demonstrate the reduced computational complexity of using large kernel attention block (LKAB), we replace the block with conventional spatial attention (SA) module, and replace the large kernel decomposition with the directly calculated large-scale convolution. The results are shown

FIGURE 13. Visual comparisons of our RLKAN with other SR methods on FLIR [6] and CVC09 [55] dataset. (×4).

in Table.5 We observed that, Large Kernel Decomposition not only reduced the computational cost of the network, but also its PSNR and SSIM performance slightly outperformed those obtained through direct large kernel computation. That's because large kernel attention effectively captures complex spatial relationships and enhances the final reconstruction result.

C. EXPERIMENTAL RESULTS

1) QUANTITATIVE EVALUATION

To demonstrate the network RLKA-Net, we conducted comparative experiments with the other 9 super-resolution models, FSRCNN [16], EDSR [5], DRCN [10], DRRN [42], MemNet [41], RCAN [7], IMDN [28], MAN [53], and SwinIR [50]. Four datasets based on OSU [54], FLIR [6], CVC09 [55], and LLVIP [56] are used. Details on the four datasets are discussed in substion IV-A. The number of parameters in each model, as well as the PSNR and SSIM values are presented in Table. 6, Fig. 11 and Fig. 2.

From Table. 6, we can observe that the number of parameters in RLKA-Net is only 100k more than FSRCNN, much fewer than the other models. In terms of PSNR and SSIM, deep learning-based methods present better results than bicubic interpolation. Different models adopt various feature extraction strategies, such as deepening the network structure, constructing recurrent modules, and introducing multiple attention mechanisms, which result in diverse final reconstruction outcomes. Overall, our RLKA-Net achieves the highest average PSNR and SSIM values for both scale factors, $\times 2$ and $\times 4$, across all four datasets. This indicates that our model is more effective in capturing and preserving the essential features of infrared images for the SR problem.

Additionally, it is worth to notice that, the number of parameters for RLKA-Net is the second smallest one among the 10 models, and much lower than the numbers for EDSR, DRCN, RCAN, IMDN, MAN, and SwinIR. This makes it more efficient and potentially more suitable for real-time applications.

2) SUBJECTIVE VISUAL EVALUATION

In Fig. 12 and Fig. 13, we present reconstruction samples using the 9 deep learning models which have the higher PSNR and SSIM values. Although FSRCNN has the fewest perameters, its performance is the worst. Thus the reconstructions are not present here. Fig. 12 presents results using samples from OSU and LLVIP, while Fig. 13 is for results using FLIR and CVC09.

Infrared images in OSU are more about campus scenes captured from a long distance. Image_00116 as shown in Fig. 12 is a typical one. By comparing the enlarged HR image and the reconstructions obtained using the 9 models, it can be observed that, IMDN, MAN, SwinIR and our RLKA-Net present much more details, while the reconstruction using EDSR, DRCN, DRRN, and MemNet have blurry and blocky artifacts. However, RLKA-Net uses much less parameters.

The LLVIP [56] dataset also focuses on campus scenes. But the images are captured from a closer distance and with a higher resolution. Using image_40116 from LLVIP as the HR sample, we can find that, the reconstruction using RLKA-Net has higher contrast compared with the other 8 SR models. Moreover, using the other models, the pedestrians in the image are blurred, and the details of the tree branches are severely lost, while the reconstruction using RLKA-Net has better resolution and more details.

Comparing the reconstructions of the sample image_004087 in the FLIR [6] dataset (as shown in Fig. 13), we can observe the followings. For the lattice-like structures of distant buildings, our RLKA-NET is the only one which can reconstruct the structure clearly. Additionally, for the nearby pedestrians and vehicles, other SR methods lose details of the pedestrian's head and backpack. The vehicle contours are blurred. In contrast, our RLKA-NET restores the details with a better visual quality.

In the last set of reconstructions using image_00253 from the CVC09 [55] dataset, we can see that, for the windows of the building, our RLKA-NET has clearer reconstructions with more details of the edges than the other models. We can get the same observation for the enlarged person's head part.

Generally, the reconstructions using RLKA-Net have much better visual quality.

V. CONCLUSION

In this paper, we design a Recurrent Large Kernel Attention Network (RLKA-Net) for infrared image super-resolution (I^2SR). The RLKA-Net is based on recurrent learning strategy and extracts a diverse set of features in both temporal and spatial dimensions, yielding better I^2SR reconstructions. In RLKA-Net, we design large kernel attention block (LKAB) and attention gate (AG) specific for I^2SR. LKAB enables efficient multi-scale feature extraction with a few number of model parameters, thus improving the performance of the network. Furthermore, the application of AG accelerates the model's training process, allowing the training loss to reach a low level quickly. Experiment results show that our model (RLKA-Net) achieves the state-of-the-art performance for the I^2SR problem.

For future work, we expect the following directions:

1) Our infrared super-resolution model (RLKA-Net) can be integrated with target recognition and detection tasks. Under low-resolution and low-contrast conditions, RLKA-Net can produce higher-quality images, enabling more accurate target recognition and detection results.

2) The RLKA-Net model is also applicable to the fusion of infrared and visible light image super-resolution tasks. The lightweight RLKA-Net can meet real-time processing requirements in various scenarios.

3) The RLKA-Net network can be further investigated for larger-scale SR, or SR for extremely low signal-to-noise ratios cases.

ACKNOWLEDGMENT

The authors would like to thank the National Natural Science Foundation of China (NSFC) (61675023, U2241275) and the Fundamental Research Funds for the Central Universities (2022CX02002).

REFERENCES

[1] A. Rogalski, P. Martyniuk, and M. Kopytko, "Challenges of small-pixel infrared detectors: A review," *Rep. Prog. Phys.*, vol. 79, no. 4, Apr. 2016, Art. no. 046501.

[2] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal Process.*, vol. 128, pp. 389–408, Nov. 2016.

[3] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 349–356.

[4] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[5] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.

[6] F. ADAS. (2022). *Free Teledyne Flir Thermal Dataset for Algorithm Training*. [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-form/

[7] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[8] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 57–65.

[9] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[10] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.

[11] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[12] W. Zhang, Y. Liu, C. Dong, and Y. Qiao, "RankSRGAN: Generative adversarial networks with ranker for image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3096–3105.
[13] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11057–11066.
[14] Y. Hu, J. Li, Y. Huang, and X. Gao, "Channel-wise and spatial feature modulation network for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 3911–3927, Nov. 2020.
[15] J. Zhang, C. Long, Y. Wang, H. Piao, H. Mei, X. Yang, and B. Yin, "A two-stage attentive network for single image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1020–1033, Mar. 2022.
[16] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 391–407.
[17] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[18] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1905–1914.
[19] Y. Huang, J. Li, X. Gao, Y. Hu, and W. Lu, "Interpretable detail-fidelity attention network for single image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 2325–2339, 2021.
[20] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang, "NTIRE 2017 challenge on single image super-resolution: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1110–1121.
[21] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.
[22] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10093–10102.
[23] J. He, C. Dong, and Y. Qiao, "Modulating image restoration with continual levels via adaptive feature modification layers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11048–11056.
[24] Z. Li, Y. Liu, X. Chen, H. Cai, J. Gu, Y. Qiao, and C. Dong, "Blueprint separable residual network for efficient image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 832–842.
[25] H. Zhao, X. Kong, J. He, Y. Qiao, and C. Dong, "Efficient image super-resolution using pixel attention," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 56–72.
[26] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 252–268.
[27] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, "Residual feature aggregation network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2356–2365.
[28] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2024–2032.
[29] L. Zhou, H. Cai, J. Gu, Z. Li, Y. Liu, X. Chen, Y. Qiao, and C. Dong, "Efficient image super-resolution using vast-receptive-field attention," 2022, *arXiv:2210.05960*.
[30] B. Wang, Y. Zou, L. Zhang, Y. Li, Q. Chen, and C. Zuo, "Multimodal super-resolution reconstruction of infrared and visible images via deep learning," *Opt. Lasers Eng.*, vol. 156, Sep. 2022, Art. no. 107078.
[31] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.
[32] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.
[33] Z. He, S. Tang, J. Yang, Y. Cao, M. Ying Yang, and Y. Cao, "Cascaded deep networks with multiple receptive fields for infrared image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2310–2322, Aug. 2019.
[34] R. Rivadeneira, A. Sappa, and B. Vintimilla, "Thermal image super-resolution: A novel architecture and dataset," in *Proc. 15th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2020, pp. 111–119.
[35] K. Prajapati, V. Chudasama, H. Patel, A. Sarvaiya, K. Upla, K. Raja, R. Ramachandra, and C. Busch, "Channel split convolutional neural network (ChaSNet) for thermal image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4363–4372.
[36] G. Batchuluun, Y. W. Lee, D. T. Nguyen, T. D. Pham, and K. R. Park, "Thermal image reconstruction using deep learning," *IEEE Access*, vol. 8, pp. 126839–126858, 2020.
[37] Y. Huang, Z. Jiang, R. Lan, S. Zhang, and K. Pi, "Infrared image super-resolution via transfer learning and PSRGAN," *IEEE Signal Process. Lett.*, vol. 28, pp. 982–986, 2021.
[38] X. Yang, M. Zhang, W. Li, and R. Tao, "Visible-assisted infrared image super-resolution based on spatial attention residual network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[39] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.
[40] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," 2014, *arXiv:1412.6632*.
[41] Y. Tai, J. Yang, X. Liu, and C. Xu, "MemNet: A persistent memory network for image restoration," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4539–4547.
[42] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2790–2798.
[43] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, "Image super-resolution via dual-state recurrent networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1654–1663.
[44] J. Lee, J. Park, K. Lee, J. Min, G. Kim, B. Lee, B. Ku, D. K. Han, and H. Ko, "FBRNN: Feedback recurrent neural network for extreme image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2021–2028.
[45] J. Qin, L. Chen, K. Liu, G. Jeon, and X. Yang, "Spatial-temporal feature refine network for single image super-resolution," *Int. J. Speech Technol.*, vol. 53, no. 8, pp. 9668–9688, Apr. 2023.
[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[47] J.-S. Choi and M. Kim, "A deep convolutional neural network with selection units for super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1150–1156.
[48] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 191–207.
[49] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12294–12305.
[50] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.
[51] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," 2022, *arXiv:2205.04437*.
[52] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," 2022, *arXiv:2202.09741*.
[53] Y. Wang, Y. Li, G. Wang, and X. Liu, "Multi-scale attention network for single image super-resolution," 2022, *arXiv:2209.14145*.
[54] J. W. Davis and M. A. Keck, "A two-stage template approach to person detection in thermal imagery," in *Proc. 7th IEEE Workshops Appl. Comput. Vis.*, Jan. 2005, pp. 364–369.
[55] K. Piniarski and P. Pawłowski, "Multi-branch classifiers for pedestrian detection from infrared night and day images," in *Proc. Signal Process., Algorithms, Archit., Arrangements, Appl. (SPA)*, Sep. 2016, pp. 248–253.
[56] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "LLVIP: A visible-infrared paired dataset for low-light vision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3489–3497.
[57] C. E. Duchon, "Lanczos filtering in one and two dimensions," *J. Appl. Meteorol.*, vol. 18, no. 8, pp. 1016–1022, Aug. 1979.

GANGPING LIU (Member, IEEE) received the B.S. degree in optical engineering from the Beijing Institute of Technology, China, in 2018, where he is currently pursuing the Ph.D. degree with the School of Optics and Photonics.

His current research interests include full-waveform LiDAR signal processing, infrared image super-resolution, single image super-resolution, and non-line of sight imaging. He is a member of the Society of Photo-Optical Instrumentation Engineers (SPIE) and the Chinese Optical Society (COS). He is a Reviewer of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.

SHUAIJUN ZHOU received the M.S. degree in optical engineering from the Beijing University of Technology, China, in 2017. He was the Deputy Chief Designer of the Aerospace Control Research Group, primarily focusing on seeker technology, guidance and detection technologies, and cutting-edge research in the applications of optical detection and guidance. He is currently with the Beijing Aerospace Automatic Control Institute, China Aerospace Science and Technology Corporation.

XIAXU CHEN received the B.S. degree in electronic science and technology from Xidian University, China, in 2023. He is currently pursuing the M.S. degree with the School of Optics and Photonics, Beijing Institute of Technology. His current research interests include single image super-resolution, video super-resolution, compressed sensing, and non-line of sight imaging.

WENJIE YUE received the B.S. degree in optical engineering from the Ocean University of China, China, in 2023. He is currently pursuing the M.S. degree with the School of Optics and Photonics, Beijing Institute of Technology. His current research interests include infrared image super-resolution, compressed sensing, and non-line of sight imaging.

JUN KE (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the University of Electronic and Science Technology of China, in 1996 and 1999, respectively, the M.S. degree in mathematics from Purdue University, in 2002, and the Ph.D. degree from the Department of Electrical and Computer Engineering (ECE), The University of Arizona (UA), in 2010.

She is currently an Associate Professor with the School of Optics and Photonics, Beijing Institute of Technology, China. Her research interests include optical science and computational imaging. She is a Senior Member of The Optical Society (OSA) and a member of the Society of Photo-Optical Instrumentation Engineers (SPIE). She is a Reviewer of *Journal of the Optical Society of America* (JOSA), *Applied Optics*, *Optics Letters*, and *Optics Express*.