NTIRE 2024 Efficient SR Challenge Factsheet -LSANet: Efficient Image Super-Resolution with Lightweight Spaced Attention Mechanism-

MViC_SR
East China Normal University
Shanghai, China
71265901062@stu.ecnu.edu.cn

1. Team details

- Team name
  MViC_SR
- Team leader name
  Jincheng Liao[1]
- Team leader address, phone number, and email
  [1]East China Normal University, China
  (+86) 17750631297
  71265901062@stu.ecnu.edu.cn
- Rest of the team members
  Bohan Jia[1], Junbo Qiao[1,2], Yunshuai Zhou[1], Yun Zhang[2,3], Wei Li[2], Shaohui Lin[1]
- Affiliations
  [1]East China Normal University, China
  [2]Huawei Noah's Ark Lab, China
  [3]The Hong Kong University of Science and Technology, China
- Affiliation of the team and/or team members with NTIRE 2024 sponsors (check the workshop website)
  The affiliation of the team is not on the list of sponsors on the website.
- User names and entries on the NTIRE 2024 Codalab competitions (development/validation and testing phases)
  All user names and entries of us are shown in Table 1.
- Best scoring entries of the team during development/validation phase
  The best scoring entries of are shown in Table 2.
- Link to the codes/executables of the solution(s) following https://github.com/cheng221/NTIRE2024_ESR

Table 1. User names and entries of our team

User name	entries	
	development	testing
BHJia	2	3
liaojc	0	3
super_zys	0	1

Table 2. Best scoring entries of our team

development		testing	
PSNR	SSIM	PSNR	SSIM
26.96	0.80	27.00	0.81

2. Method details

General method description. As shown in Figure 1, we propose a network with lightweight spaced attention (LSANet). The architecture of LSANet consists of the following parts: the shallow feature extraction, the deep feature extraction based on spaced local feature extraction module (LFEM) and reparamaterized spaced attention Block (RSAB), and the reconstruction.
Given the LR input I_{LR}, a single 3×3 convolution is applied to extract the shallow feature $F_0 \in \mathbb{R}^{C\times H\times W}$ in the first part:

$$F_0 = \mathrm{Conv}\left(I_{LR}\right), \tag{1}$$

where I_{LR}, C, H, W are the input LR image, channel dimension, height and width of the input, respectively.
In the second part, we use six alternate blocks to extract the deep feature $F_d \in \mathbb{R}^{C\times H\times W}$:

$$F_d = H_D(F_0). \tag{2}$$

Specifically, H_D is comprised of local feature extraction module (LFEM) and reparamaterized spaced attention Block (RSAB). By taking F_s and F_d as inputs, the HR im-

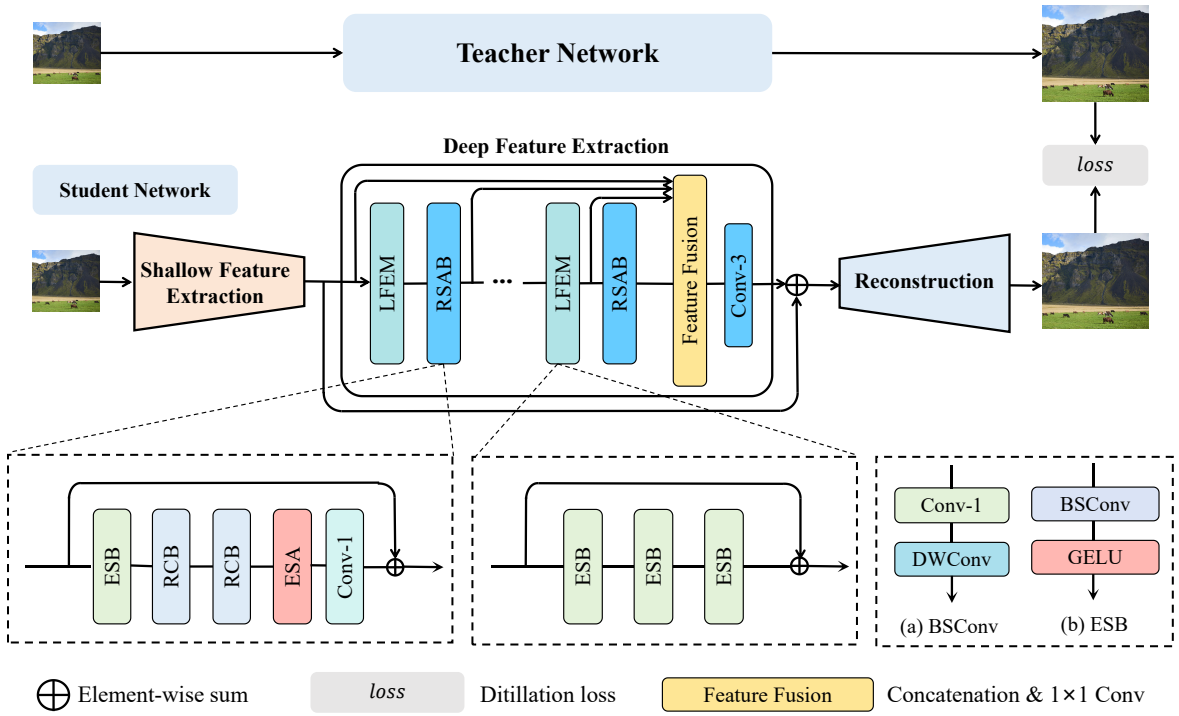

Figure 1. The pipeline of LSANet. (a) Structure of BSConv. (b) Structure of ESB.

age I_{HR} is reconstructed with an upsampler as:

$$I_{HR} = H_{RC}(F_0 + F_d), \tag{3}$$

where H_{RC} is the reconstruction module involves a single 3×3 convolution followed by a pixel shuffle operation [9]. Previous methods mostly use several plain convolutions to extract features, which is inefficient with soaring computational complexity. Inspired by the blueprint shallow residual block [8], we design an efficient yet effective local feature extraction module (LFEM) to alleviate computing burden while largely maintaining model performance. As illustrated in Figure 1(a), the LFEM contains three efficient shallow blocks (ESB) which include a 1×1 point-wise convolution with a 3×3 depth-wise convolution followed by GELU activation [4].

The attention mechanism guides the network to adaptively focus on the most relevant and informative part of the input. Previous works typically apply attention modules in consecutive blocks. However, we find that it is not compulsory to continuously add an attention module in every block. For two adjacent blocks, we can remove the attention module from one of them, ensuring that there is only one attention module between any two consecutive blocks.This lies in the intuition that human visual perception has a lasting effect. Similarly, attention block can preserve its representation ability to ajacent blocks. This simple change reduces model complexity without hindering performance.

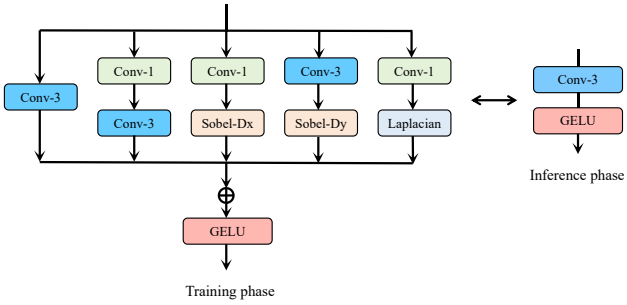

Figure 2. The structure of RCB.

Consequently, we propose a reparamaterized spaced attention Block (RSAB), which is composed of an efficient shallow block and two reparameterized convolution blocks (RCB) followed by an enhanced spatial attention block (ESA) [6] and a convolution. The spaced ESA blocks only used in RSAB are employed for comprehensive extraction and modulation of deep features. We use a 1×1 convolution after the ESA block to further refine the weighted feature and capture local patterns. The detailed structure is shown in Figure 2. Intermittently configurating attention blocks substantially lower model complexity without noticeable performance drop. At the end of this part, features extracted from all blocks are concatenated and aggregated using two convolutions.

Reparameterization. Reparameterization [2, 3, 11] has

proven effective to enhance feature representation without introducing additional computational overhead. Different from the reparameterization module design of high-level tasks, we design an isotropic edge-oriented convolutional block in our model. As shown in Figure 2(a), the Sobel-Dx and Sobel-Dy employ the isotropic Sobel function to improve the representation capabilities of our model. During the inference phase, all branches are combined to a simplified 3×3 convolution, which significantly reduces computation cost.

Training Strategy. We apply two stages to train our network on DIV2K [1] and the first 10K data of LSDIR [7]. We randomly augment the input with a flip and 90-degree rotation to enhance the robustness of the network. We design a teacher-student distillation strategy for training, which makes the large version of SAFMN [10] as teacher and the proposed network as student. Three LFEM and RSAB blocks are stacked alternately in our student network, and the feature channel is 36. The mini-batch size is fixed to 64 in all stages. The details of the two training stages are as follows:

- **Stage 1.** We minimizing the L1 loss (student prediction and groun-truth) and distillation loss (student prediction and teacher preditction) to optimize the student network by Adam optimizer [5] for $1000k$ iterations. The initial learning rate is set to 2×10^{-3}, which will be halved at $100k$, $500k$, $800k$, $900k$, $950k$. The HR patch size is 256 during this stage.
- **Stage 2.** We initialize the weight of the student network with the pre-trained student in Stage 1. We enlarge the HR patch size to 640 and minimize MSE loss instead of L1 loss.

Experimental Results. As shown in Table 3, our method achieves 26.90 dB, 27.00 dB on the validation and testing phase data, respectively. We measure the inference time on validation and testing data by executing the script five times and average these results to obtain mean inference time. We maintain the average inference time and reduce the FLOPs and the number of parameters for 11.51G and 0.154M compared with the baseline RLFN. This significantly reduces the amount of computation and memory overhead.

Table 3. The complexity comparison that tests on an RTX 3090 GPU

Method	Time[ms]			PSNR[dB]		#Params	Flops	GPU Mem
	Ave.	Val.	Test	Val.	Test	[M]	[G]	[M]
RLFN [6]	15.54	15.58	15.50	26.96	27.07	0.317	19.67	774.28
LSANet	15.68	16.17	15.19	26.90	27.00	0.139	8.16	634.586

3. Other details

- We may submit a solution(s) description paper at NTIRE 2024 workshop.
- General comments and impressions of the NTIRE 2024 challenge.
  This challenge is a great event for the field of image restoration, promoting its development.
- What do you expect from a new challenge in image restoration, enhancement and manipulation?
  Promote the development of efficient super-resolution networks
  Learning to design a fast and effective image restoration network has yielded efficient methods.
- Other comments: encountered difficulties, fairness of the challenge, proposed subcategories, proposed evaluation method(s), etc.
  There could be a more standardized testing platform that can provide accurate time measurements with minimal deviation.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 3

[2] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Diverse branch block: Building a convolution as an inception-like unit. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10886–10895, 2021. 2

[3] Zongcai Du, Ding Liu, Jie Liu, Jie Tang, Gangshan Wu, and Lean Fu. Fast and memory-efficient network towards efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 853–862, 2022. 2

[4] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 2

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 3

[6] Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 766–776, 2022. 2, 3

[7] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1787, 2023. 3

[8] Zheyuan Li, Yingqi Liu, Xiangyu Chen, Haoming Cai, Jinjin Gu, Yu Qiao, and Chao Dong. Blueprint separable residual network for efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 833–843, June 2022. 2

[9] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan

Wang. 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*, pages 1874–1883, 2016. 2

[10] Long Sun, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Spatially-adaptive feature modulation for efficient image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[11] Xindong Zhang, Hui Zeng, and Lei Zhang. Edge-oriented convolution block for real-time super resolution on mobile devices. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4034–4043, 2021. 2