NTIRE 2023 Efficient SR Challenge Factsheet
-Local-Global Visual Attention Network-

Zhijian Wu, Dingjiang Huang
School of Data Science and Engineering, East China Normal University
Shanghai, China
zjwu_97@stu.ecnu.edu.cn, djhuang@dase.ecnu.edu.cn

1. Team details

- Team name

  Dase-IDEALab

- Team leader name

  Zhijian Wu

- Team leader address, phone number, and email

  Address: Shanghai, China

  Phone number: +86 15651768331

  Email: zjwu_97@stu.ecnu.edu.cn

- Rest of the team members

  Dingjiang Huang

- Team website URL (if any)

  N/A

- Affiliation

  School of Data Science and Engineering, East China Normal University

- Affiliation of the team and/or team members with NTIRE 2023 sponsors (check the workshop website)

  N/A

- User names and entries on the NTIRE 2023 Codalab competitions (development/validation and testing phases)

  User name: zjwu

  development phase entries: 7

  testing phase entries: 3

- Best scoring entries of the team during development/validation phase

PSNR	SSIM	Runtime	Params	Extra Data
29.00 (21)	0.83 (20)	0.07 (20)	118243.00 (5)	1.00 (1)

- Link to the codes/executables of the solution(s)

  https://github.com/charonf/LGVAN-NTIRE2023-ESR

2. Method details

2.1. The Overall Architecture

As shown in Fig. 1(a), the overall architecture of our LGVAN mainly consists of three parts: shallow feature extraction, deep feature extraction, and upscaling reconstruction.

Specifically, we use a 3×3 convolution for extracting shallow features f_0 from the input LR image I_{LR}:

$$f_0 = H_s(I_{LR}) \tag{1}$$

where H_s represents a 3×3 convolutional layer. Subsequently, the shallow feature is used for the deep feature extraction by a stack of local-global attention blocks (LGAB). This process can be formulated as:

$$f_k = H_k(f_{k-1}), \quad k = 1, ..., n \tag{2}$$

where H_k denotes the k-th LGAB. f_{k-1} and f_k indicate input feature and output feature of the k-th LGAB. We employ a 3×3 convolution at the tail of the deep feature extraction, and add the residual f_0 to the output feature:

$$f_r = Conv(f_k) + f_0 \tag{3}$$

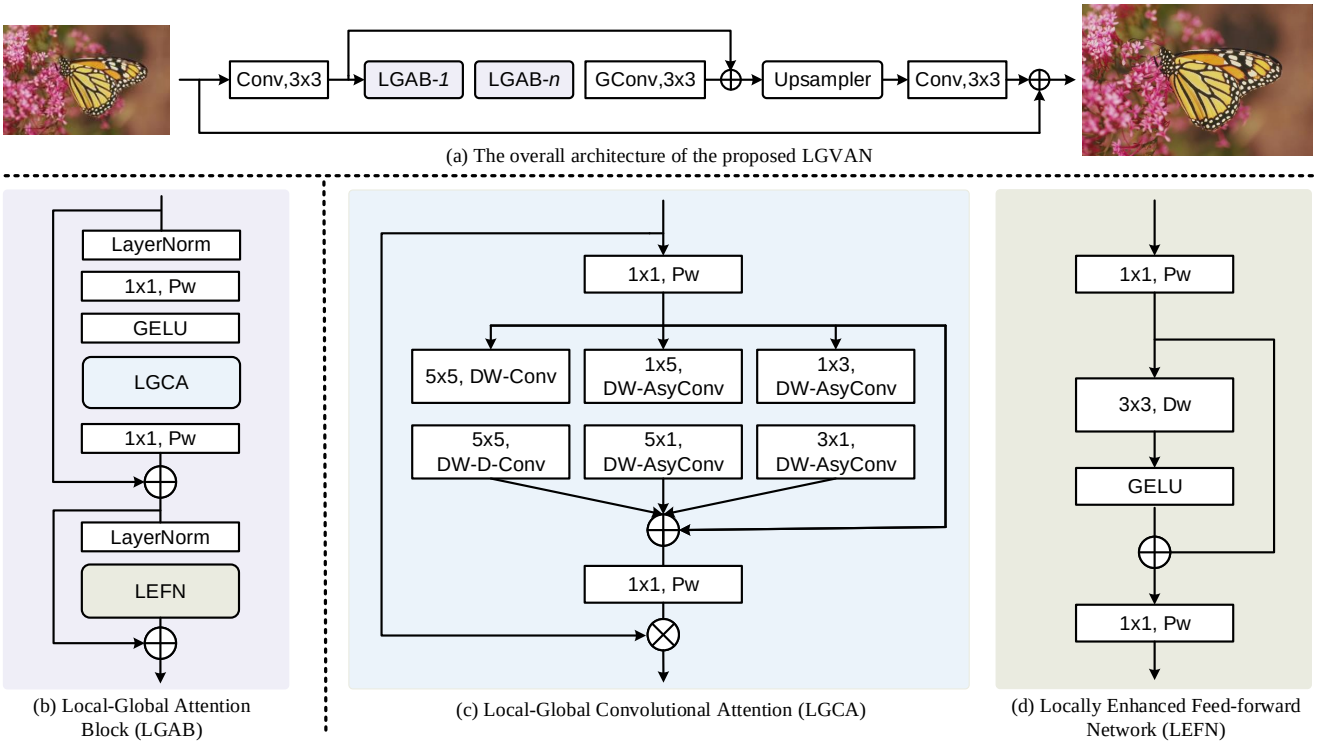
Finally, the SR images are constructed by the resulting deep features as follows:

$$I_{SR} = H_r(f_r) \tag{4}$$

where I_{SR} denotes the reconstructed SR image. H_r represents the reconstruction module composed of a pixel shuffle operator and a 3×3 convolution layer.

We optimize our LGVAN with two types of loss functions $\mathcal{L} = \mathcal{L}_1 + \lambda \mathcal{L}_f$. Pixel-wise $\mathcal{L}_1$ loss is used to ensure that the generated content is close to ground truth:

$$\mathcal{L}_1 = ||R - T||_1 \tag{5}$$



(a) The overall architecture of the proposed LGVAN

(b) Local-Global Attention Block (LGAB)

(c) Local-Global Convolutional Attention (LGCA)

(d) Locally Enhanced Feed-forward Network (LEFN)

Figure 1. The overall pipeline of our LGVAN (a), which consists of multiple Local-Global Attention Blocks (LGAB) (b). The core components of LGAB are (b) Local-Global Convolutional Attention (LGCA) that obtains local and global features while evoking attention mechanism via element-wise multiplication, and (c) Locally Enhanced Feed-forward Network (LEFN) that performs local structural enhancement to focus on fundamental detail contents.

where R and T denote the resulting SR image and the target image respectively. In addition, $\mathcal{L}_f$ is the frequency reconstruction loss [3] that enforces high-frequency details:

$$\mathcal{L}_f = ||\mathcal{F}(R) - \mathcal{F}(T)||_1 \tag{6}$$

where $\mathcal{F}(\cdot)$ represents the 2D Fast Fourier Transform. We used $\lambda = 0.5$ as the weighting factor in all experiments.

2.2. Local-Global Attention Block

As shown in Fig. 1(b), our LGAB follows the ViT architecture [2, 4, 10], but renovates the attention module and feed-forward network (FFN) for SR tasks. Next we describe the core components of the proposed LGVAN: (a) local-global convolutional attention (LGCA) and (b) locally enhanced feed-forward network (LEFN).

Local-Global Convolutional Attention The proposed LGCA is shown in Fig. 1(c). Haase et al. [5] reveal that DSC-based architectures implicitly rely on cross-kernel correlations. Inspired by this, we designed a novel multi-branch depth-wise separable structure. Specifically, the depth-wise convolutions on multi-branches share the point-wise convolution on the trunk, where the former obtains local and global information while the latter captures relationship along the channel dimension. Subsequently, we use the generated features as attention maps to recalibrate the inputs of LGCA. The computational process can be expressed as follows:

$$Attention = Conv_p \sum_{i=0}^{3} (Branch_i(Conv_p(F))) \tag{7}$$

$$Output = Attention \otimes F \tag{8}$$

where F denotes the input feature, while $Attention$ and $Output$ indicate attention map and the output feature respectively. $\otimes$ is element-wise multiplication. $Conv_p$ represents 1×1 point-wise convolution, and $Branch_i, i \in \{0, 1, 2, 3\}$ denotes i-th branch as shown in Fig. 1(c). $Branch_0$ consists of a 5×5 depth-wise convolution and a 5×5 depth-wise dilation convolution with dilation of 3 to capture long-range dependencies. This combination achieves large receptive fields of 17 while reducing a large number of parameters. $Branch_1$ and $Branch_2$ respectively consist of paired depth-wise asymmetric convolutions with kernel sizes of 3 and 5, which are used to extract local and texture information. This design is inspired by traditional edge detection operators, such as the Sobel filter, which utilizes horizontal and vertical filters to extract local structure information in different directions respectively.

Table 1. Result for NTIRE2023 ESR Challenge. FLOPs and Activation are tested on an LR image of size 256×256. The runtime is is averaged on DIV2K and LSDIR test datasets using a single NVIDIA Tesla V100 GPU.

PSNR[val]	PSNR[test]	Params[M]	FLOPs[G]	GPU Mem.[M]	Activation[M]	Average Runtime[ms]
29.00	27.07	0.1182	9.0576	1114.72	332.39	43.28

In addition, this lightweight design is more effective than the standard convolution. Besides, $Branch_3$ is an identity mapping to allow feature reuse in adjacent layers.

Locally Enhanced Feed-forward Network Similar to the standard FFN, our LEFN adopts two 1×1 convolutions as projection layers, with one expanding the feature channels and another shrinking back to the original dimension. In addition, we incorporate a 3×3 depth-wise convolution to learn the local structure, whilst introduce the self-residual strategy [8] to overcome the drawbacks associated with depth-wise convolution. Mathematically, our LEFN can be formulated as:

$$F_e = Conv_e(F) \tag{9}$$

$$Output = Conv_c(\phi((Conv_d(F_e)) + F_e)) \tag{10}$$

where F denotes the input feature. $Conv_e$ and $Conv_c$ are two 1×1 convolutions carrying out channel expansion and contraction respectively. ϕ indicates GELU activation [6]. Thus, LEFN specializes in capturing local content to preserve favorable image structure and texture. It collaborates with LGCA, leading to high-quality image reconstruction.

2.2.1 Training Details

The proposed LGVAN consists of 9 LGABs, and the number of channels was set to 32. DIV2K [1] and Filckr2K [9] datasets were used for training. Data augmentation strategies involved in our experiments included horizontal and vertical flips, and random rotations of 90, 180, and 270 degrees. In terms of the model optimization, the Adam [7] optimizer was used with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate of LGVAN was set to 5×10^{-4} and reduced by half after 2×10^5 iterations. During the training process, the mini-batch and the input patch size of LGVAN were set to 128 and 64×64. Our model was trained using a total of 1×10^6 iterations **from scratch**.

2.2.2 Experimental Results

The experimental result is shown in Tbale 1. FLOPs and Activation are tested on an LR image of size 256×256. PNSR[val] is tested on DIV2k validation dataset, while PSNR[test] is calculated on a combination of DIV2K and LSDIR test data. The runtime is is evaluated on DIV2K and LSDIR test datasets using a single NVIDIA Tesla V100 GPU.

2.3. Novelty degree of the solution

Our solution is based on the research we have submitted to other conference.

3. Other details

- Planned submission of a solution(s) description paper at NTIRE 2023 workshop.

  We are not planning to submit the solution description paper to NTIRE2023 workshop, since it has been submitted to other conference.

- General comments and impressions of the NTIRE 2023 challenge.

  The organizers provided detailed processes and instructions and we appreciate the effort they spent!

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR workshops*, pages 126–135, 2017. 3

[2] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *ECCV*, 2022. 2

[3] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *ICCV*, pages 4641–4650, 2021. 2

[4] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *IEEE TPAMI*, 2022. 2

[5] Daniel Haase and Manuel Amthor. Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets. In *CVPR*, pages 14600–14609, 2020. 2

[6] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[8] Bin Sun, Yulun Zhang, Songyao Jiang, and Yun Fu. Hybrid pixel-unshuffled network for lightweight image super-resolution. *AAAI*, 2023. 3

[9] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPR workshops*, pages 114–125, 2017. 3

[10] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022. 2