NTIRE 2023 Efficient SR Challenge Factsheet -Hierarchical Context Aggregation Network-

Tianle Liu, Jinpeng Shi, Shizhuang Weng
Anhui University
Anhui,China
tianle.l@outlook.com, jinpeeeng.s@gmail.com, weng_1989@126.com

1. Team details

- Team name: FRL Team 3
- Team leader name: Tianle Liu
- Team leader address, phone number, and email

  Address: Anhui,China

  Phone number: +86 15856721586

  Email: tianle.l@outlook.com
- Rest of the team members

  Jinpeng Shi, Shizhuang Weng
- Team website URL (if any)

  github.com/Fried-Rice-Lab/FriedRiceLab
- Affiliation

  School of Electronic Information Engineering, Anhui 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: TianleLiu

  development phase entries: 1

  testing phase entries: 1
- Best scoring entries of the team during development/validation phase

PSNR	SSIM	Runtime	Params	Extra Data
29.00 (15)	0.83 (12)	0.05 (15)	178736.00(13)	1.00 (1)

- Link to the codes/executables of the solution(s): github.com/TIANLE233/NTIRE2023_ESR

2. Method details

2.1. Network Architecture

As shown in Fig.1, HCAN contains three modules: (1) feature extraction, (2) nonlinear mapping, and (3) HR-image reconstruction. In the first stage, we use a 3×3 convolution layer to extract the coarse features from low resolution images. Then we implement the nonlinear mapping by cascading multiple NTBs. Next, we use one 3×3 convolution layer and one nonparametric sub-pixel operation to reconstruct high resolution images. Global shortcut connections are used. Our overall structure is neat, as complex topologies can lead to a serious reduction in inference speed [10].

2.2. Normalization-free Transformer Block

Normalization-free Transformer block (NTB) uses a Transformer architecture, and replace self-attention (SA) with hierarchical context aggregation attention (HCAA) designed to capture short-range and long-range contexts at multiple levels. To further improve effectiveness and efficiency, NTB removes Layer Normalization (LN) [2] from Transformer architecture and does not introduces other types of normalization. To the best of our knowledge, we are the first to use the Transformer architecture without any normalization in the super-resolution (SR) field. Normalization can speed up and stabilize training, but Layer Normalization (LN) increases inference time and memory usage, and Batch Normalization (BN) [4] damages performance of SR [7]. We find that Sigmoid can replace the normalization effect to some extent without causing the above losses.

2.3. Hierarchical Context Aggregation Attention

Inspired by FocalNet [9], we replace SA with HCAA in Transformer architecture. The point-wise convolution (PWconv) is placed at the head and tail of the HCAA to exchange information between channels, The body part

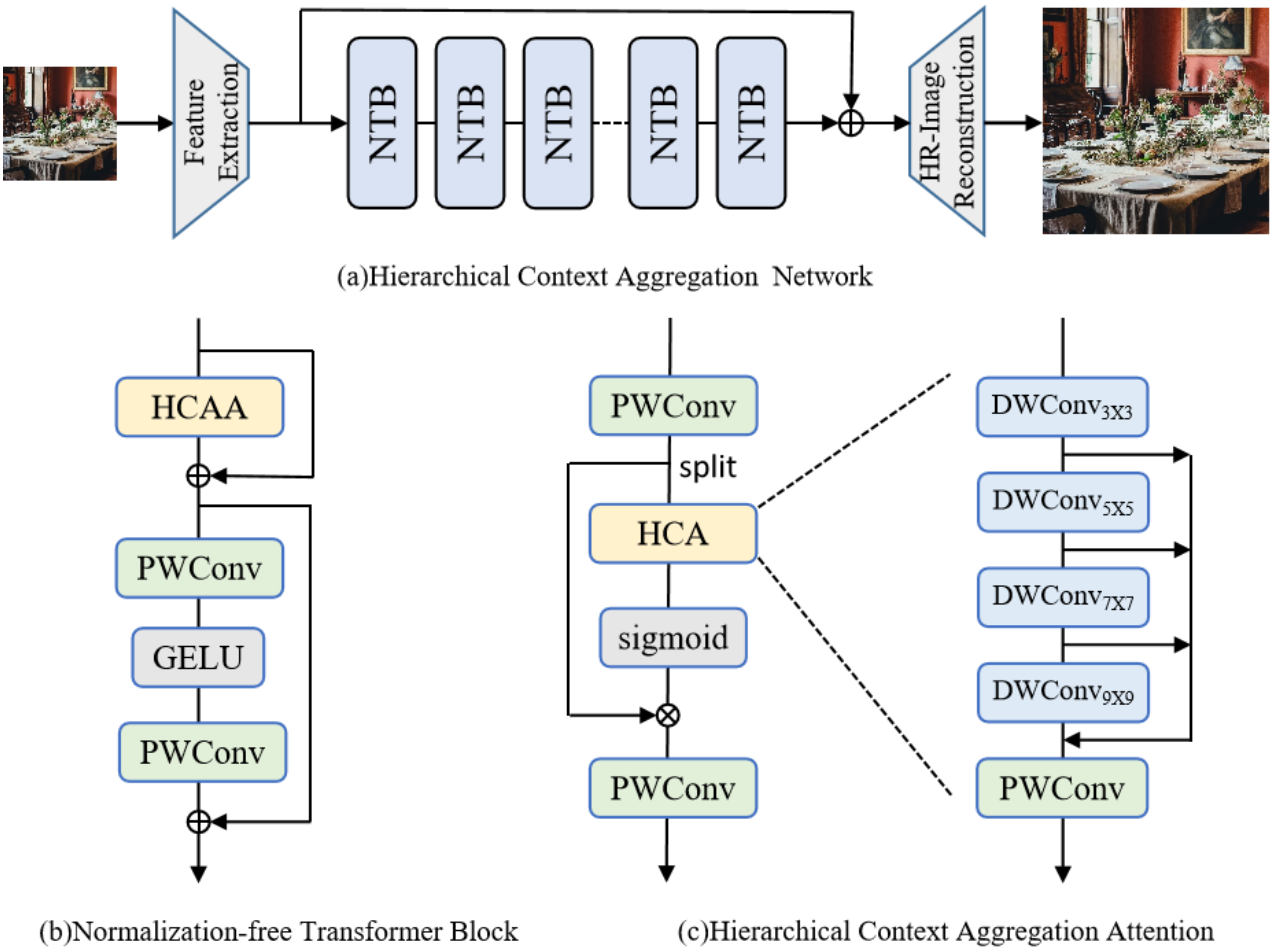

Figure 1. (a) is overall pipeline of HCAN, which consists cascading multiple NTBs (b) . NTB uses a Transformer architecture without any normalization. Its core components is hierarchical context aggregation attention (HCAA) , which can capture short-range and long-range contexts at multiple levels. Each DWconv is followed by a GELU [3].

mainly acquires spatial contextual information through hierarchical depth-wise convolution (DWconv) with increasing kernel sizes in a cascading manner. The extracted features from each level of DWconv are added together to generate features with short-range and long-range contexts. The generated features are channel blended by PWconv and then attention maps are generated using Sigmoid. Finally, it is multiplied by elements with the unprocessed feature map. Compared to FocalNet, we remove the global aggregation and gated aggregation. The former impair performance [5], while the latter impair efficiency and lead to unstable training.

2.4. Training strategy

The proposed HCAN consists of 12 NTBs and the number of channels is set to 32. We train our models on DIV2K [1], Flickr2K [7] and LSDIR dataset. Data augmentation methods of random rotation by $90°$, $180°$, $270°$ and flipping horizontally are utilized. The minibatch size is set to 128 and the patch size of each LR input is set to 64×64. The model is trained by L1 loss [8] with Adam optimizer [6] ($\beta_1 = 0.9$, $\beta_2 = 0.999$) for 1×10^6 total iterations. The learning rate is initialized as 5×10^{-4} and scheduled by cosine annealing learning.

Table 1. Result of DIV2K and LSDIR test sets

PSNR	SSIM	Average Runtime[ms]	Params[M]
27.09(4)	0.81 (2)	0.04 (10)	178736.00(13)
GPU Mem.[M]	FLOPs[G]	Activation[M]	Conv2d
262.6	11.5	285.4	112

2.5. Experimental results

We test our model on the DIV2K and LSDIR test sets, and the experiments are performed on an RTX 3090, using the official code. The results is shown in Table 1.

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. 2

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

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

[4] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 1

[5] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1645, 2016. 2

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[7] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 1, 2

[8] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2

[9] Jianwei Yang, Chunyuan Li, and Jianfeng Gao. Focal modulation networks. *arXiv preprint arXiv:2203.11926*, 2022. 1

[10] 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. 1