PYRAMID FUSION ATTENTION NETWORK FOR SINGLE IMAGE SUPER-RESOLUTION

Hao He[⋆]*, Zongcai Du*[⋆]*, Wenfeng Li, Jie Tang*[†]*, Gangshan Wu*

State Key Laboratory for Novel Software Technology, Nanjing University, China
{haohe, 151220022, wenfengli}@smail.nju.edu.cn, {tangjie, gswu}@nju.edu.cn

ABSTRACT

Recently, convolutional neural network (CNN) has made a mighty advance in image super-resolution (SR). Most recent models exploit attention mechanism (AM) to focus on high-frequency information. However, these methods exclusively consider interdependencies among channels or spatials, leading to equal treatment of channel-wise or spatial-wise features thus hindering the power of AM. In this paper, we propose a pyramid fusion attention network (PFAN) to tackle this problem. Specifically, a novel pyramid fusion attention (PFA) is developed where stacked residual blocks are employed to model the relationship between pixels among all channels, and pyramid fusion structure is adopted to expand receptive field. Besides, a progressive backward fusion strategy is introduced to make full use of hierarchical features, which are beneficial to obtaining more contextual representations. Comprehensive experiments demonstrate the superiority of our proposed PFAN against state-of-the-art methods.

Index Terms— single image super-resolution, convolutional neural networks, attention mechanism, feature fusion strategy

1. INTRODUCTION

Single Image Super-Resolution (SISR) is a long-standing low-level vision task which aims at recovering a high-resolution (HR) image from its low-resolution (LR) counterpart, and it has a wide range of application prospect in various areas. SISR is highly ill-posed because of one-to-many mapping from LR to HR solution space. In the literature, many algorithms have been proposed to solve this inverse problem. Recently, convolutional neural network (CNN) based methods like [1, 2, 3, 4] have improved the performance by a large margin compared with traditional methods. [5, 6, 7, 8, 9] digs deeply into the relationship between inside feature maps. Furthermore, attention mechanism (AM) has proved to be extremely powerful in increasing the quality of representations. [10, 11, 12, 9, 13] exploit different AMs, which achieved considerable progress. Meanwhile, some methods in other fields exploit pyramid attention[14, 15]. Methods above exhibit limitations in two aspects: (1) most of them focus on capturing inherent feature correlations only along channels or spatials, hindering the ability of AM; (2) most existing methods fail to sufficiently utilize intermediate features which are useful for reconstructing spatial contextual details, thereby resulting relatively-low performance.

To address above drawbacks, we propose a pyramid fusion attention network (PFAN) for constructing more powerful feature representations and enhancing the discriminative ability of the network. In detail, our building block, pyramid fusion attention block (PFAB), adopts a residual block to extract features and a pyramid fusion attention structure (PFA) to recalibrate obtained features. In the bottom pyramid level, the feature size is retained to learn pixel-wise correlations, which is more flexible in dealing with different types of information compared with channel-wise or spatial-wise attention. Higher pyramid levels can grasp more global contextual information with relatively-small overheads. Fusion between neighboring pyramid levels leads to information exchange across multi-resolution. To fully utilize the hierarchical features produced by each PFAB, we further propose a progressive backward fusion module (PBFM) to generate more discriminative features. Several PFABs and one PBFM formed a basic group (BG). We repeat BG several times and use a global skip connection to construct the final PFAN, which exhibits a superior performance using much fewer parameters (11.9M), compared with SAN [11] (15.7M) and DIN [13] (19.9M).

In summary, our main contributions are summarized as follows: (i) We propose a pyramid fusion attention network (PFAN) for highly accurate image SR. (ii) We propose a pixel attention (PA) mechanism to adaptively rescale each pixel using pyramid fusion structure, which allows the information exchange across multi-resolution and provides more global receptive field. (iii) We propose a progressive backward fusion module (PBFM) to make full use of hierarchical features to obtain more contextual information for high-frequency details recovery.

2. METHODOLOGY

In this section, we first delineate the whole architecture of our proposed Pyramid Fusion Attention Network (PFAN), then the Pyramid Fusion Attention Block (PFAB) is detailedly

[⋆] Equal contribution. [†]Corresponding author.

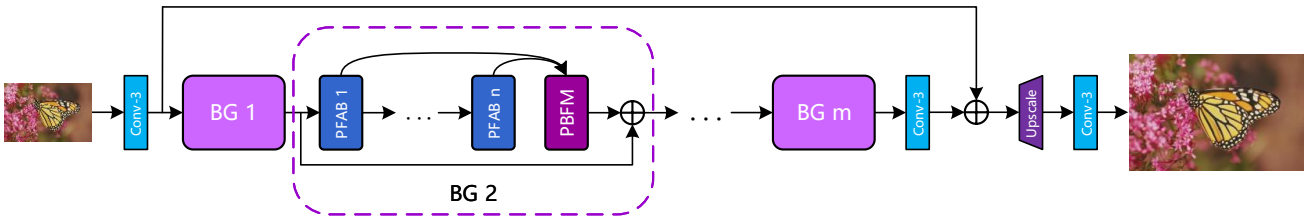


Fig. 1. The whole architecture of our proposed Pyramid Fusion Attention Network (PFAN).

described, combining with our new attention mechanism. Finally, we illustrate our novel progressive backward fusion module (PBFM).

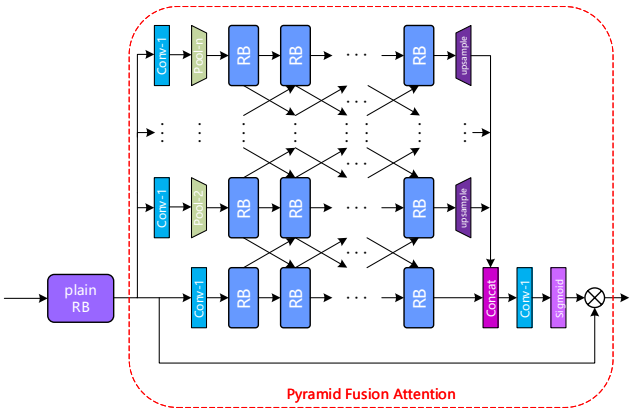


Fig. 2. The architecture of our proposed Pyramid Fusion Attention Block (PFAB), the red dotted rectangle represents the Pyramid Fusion Attention Mechanism.

2.1. Pyramid Fusion Attention Network

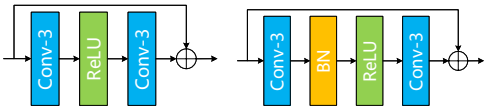
Our proposed PFAN, as is illustrated in figure 1, mainly consists of four crucial parts: shallow feature extraction, stacked basic groups, upscale module and reconstruction part. We keep pace with previous works such as [5, 10, 13] to ensure that our improvement comes from our network design, which means the $L1$ loss is adopted to optimize the proposed network. Given a training set with N LR patch images with their HR counterparts, the loss function can be denoted as

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N} ||H_{PFAN}(I^i_{LR}) - I^i_{HR}||_1 \qquad (1)$$

where Θ stands for the parameter set of PFAN.

2.2. Pyramid Fusion Attention Block

We now show our Pyramid Fusion Attention Block (PFAB) (Figure 2), which consists of one plain residual block and a proposed Pyramid Fusion Attention (PFA).



(a) plain RB (b) RB

Fig. 3. (a) plain residual block. (b) residual block with BN.

We first adopt one plain Residual Block (plain RB) to extract features. Followed by EDSR [5], we remove the batch normalization (BN) layers in our plain RB structure (figure 3(a)), which is vital for low-level tasks like SR. As is delineated in figure 2, stacked residual blocks (RBs) are employed to form one level to model the relationship between pixels. Different with previous works, we reconsider the BN layer in these RBs (figure 3(b)). Ever since EDSR [5] explores the effects of BN layer in low-level tasks, almost all subsequent methods drop this mechanism away. Nevertheless, attention module aims at guiding the network to focus more on salient areas, rather than directly participating in the calculation of SR results, leading this part to be high-level. Accordingly, reintroducing BN layer into attention module could not result in performance degradation, but is contrarily advantageous to accelerate the convergence of network training and prevent gradient exploring or disappearing.

To further enlarge the receptive field and lift the power of attention, the pyramid structure is adopted (figure 2). The architecture of pyramid levels are analogous, with a max pooling layer after the 1×1 convolutional layer (pooling-n stands for a max pooling layer which downscale the input by the factor of n). Each pyramid level adopts different value of n. Owing to the design above, the bottom pyramid level maintains the feature size to learn pixel-wise correlations, which is more flexible in dealing with different types of information, while higher levels can get larger receptive field and grasp more global contextual information. Furthermore, middle pyramid levels take the output from relative upper and lower RBs as input, using one 1×1 convolutional layer to fuse, providing information exchange across multi-resolution. Outputs from all pyramid levels are ultimately concatenated and then sent into a convolutional layer and a sigmoid layer to get the final attention map.

2.3. Progressive Backward Fusion

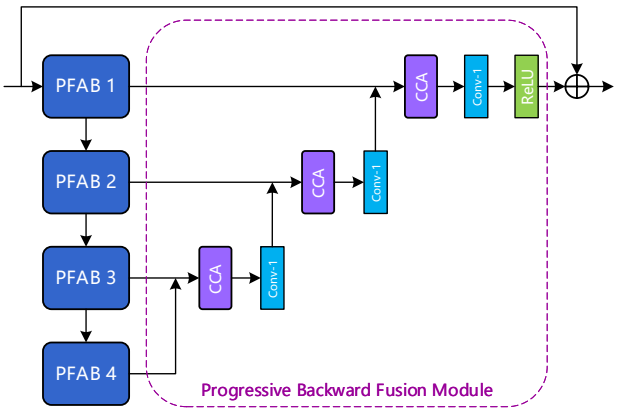


Fig. 4. The architecture of our Basic Group (BG), the dotted rectangle represents the Progressive Backward Fusion Module.

Figure 4 displays the framework of the Basic Group (BG), which consists of 4 PFABs and one progressive backward fusion module (PBFM). Based on that features from deeper blocks could guide shallow features for better restoration, we concatenate the outputs of two adjacent PFABs and feed them into a well-chosen attention module to further enhance crucial information. Let's denote the output of the $i-th$ block as B_i and the fused output from the $(i+1)-th$ block as $B_i^{'}$, we have

$$B_{i-1}^{'} = H_F(concat(B_{i-1} + B_i^{'})) \tag{2}$$

where $H_F(\cdot)$ stands for the combination of attention module and the convolutional layer and $concat(\cdot)$ represents the concatenate operation. To fully fuse features and ensure the module to be lightweight, we adopt contrast channel attention (CCA) [12] as the attention module, which enhance details like information about structures, textures and edges [12]. After the above progressive fusion steps, one activate function layer is also adopted to maintain nonlinearity. We use a skip-connection to forward shallow features after the fusion operation to produce the ultimate feature of this module.

3. EXPERIMENTAL EVALUATION

3.1. Setup

Following [10, 11, 13], we utilize 800 HR images from DIV2K as training dataset. For testing, 5 standard benchmarks are adopted: Set5 [16], Set14 [17], B100 [18], Urban100 [19], Manga109 [20]. For training process, data augment is conducted via randomly horizontally flipping and rotating 90°, 180°, 270°. Each min-batch takes 16 LR RGB patches with size 48×48 as input. Our model is trained by Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$. The initial learning rate is set to 10^{-4} and then decreases to half every 200 epochs. We implement our PFAN with Pytorch framework on an Nvidia 1080Ti GPU.

3.2. Ablation Study

Methods	baseline	baseline+PBFM	baseline+PFA	PFAN
PBFM	✗	✓	✗	✓
PFA	✗	✗	✓	✓
PSNR	37.70	37.76	37.88	37.92

Table 1. Effectiveness of our proposed PFA and PBFM. We report the best PSNR(db) values on Set5($\times 2$) in 5×10^3 iterations.

To verify the effectiveness of PFA and PBFM, we compare the networks without adopting these two modules. Studies are conducted with the magnification factor of $\times 2$ on Set5 (Table 1). The baseline is designed by removing PFA and PBFM from our original model. To ensure fairness, we duplicate the basic group (BG) by 30 times to get roughly the same size model. Compared with the baseline, PBFM promotes 0.06dB with regard to PSNR and PFA boosts 0.18dB. When combining these two modules together, our proposed algorithm achieves better results by up to 0.22dB. These observations authenticate the superiority of the two modules.

3.3. Results with Bicubic Degradation (BI)

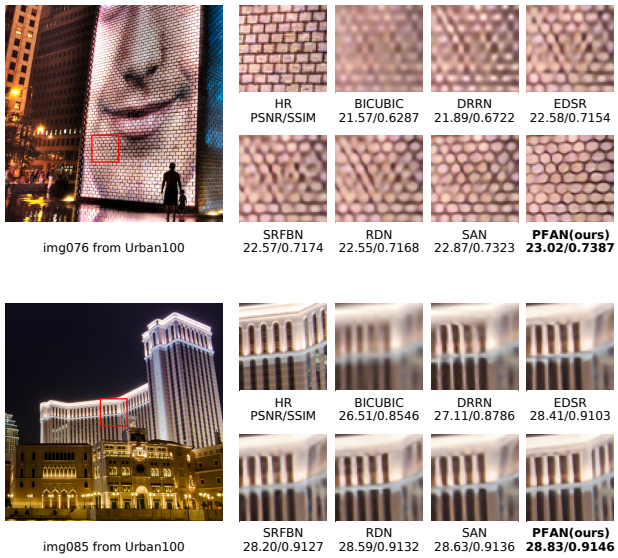


Fig. 5. Visual comparisons for ×4 SR with BI degradation model

Simulating LR images with BI degradation model is widely applied in SISR setting. To manifest the superiority

Dataset	Bicubic	SRCNN	LapSRN	DRRN	EDSR	SRFBN	RDN	SAN	DIN	PFAN(Ours)	PFAN+(Ours)
Set5	30.39/0.8682	32.75/0.9090	33.82/0.9227	34.03/0.9244	34.65/0.9280	34.70/0.9292	34.71/0.9296	34.75/0.9300	34.76/0.9298	34.79/0.9302	**34.86/0.9307**
Set14	27.55/0.7742	29.30/0.8215	29.79/0.8320	29.96/0.8349	30.52/0.8462	30.51/0.8461	30.57/0.8468	30.59/0.0.8476	30.65/0.8480	30.65/0.8487	**30.77/0.8500**
B100	27.21/0.7385	28.41/0.7863	28.82/0.7973	28.95/0.8004	29.25/0.8093	29.24/0.8084	29.26/0.8093	29.33/0.8112	29.29/0.8098	29.31/0.8110	**29.37/0.8119**
Urban100	24.46/0.7349	26.24/0.7989	27.07/0.8272	27.56/0.8376	28.80/0.8653	28.73/0.8641	28.80/0.8653	28.93/0.8671	28.94/0.8682	29.00/0.8696	**29.23/0.8728**
Manga109	26.95/0.8556	30.48/0.9117	32.19/0.9334	32.42/0.9359	34.17/0.9476	34.18/0.9481	34.13/0.9484	34.30/0.9494	34.46/0.9496	34.40/0.9498	**34.70/0.9511**

Table 2. Quantitative results with BI degradation model. Average PSNR/SSIM values for scaling factor $\times 3$. The best performance is shown **highlighted** and the second best underlined.

Dataset	Bicubic	SPMSR	SRCNN	VDSR	IRCNN	SRMD	RDN	RCAN	SAN	PFAN(Ours)	PFAN+(Ours)
Set5	28.78/0.8308	32.21/0.9001	32.05/0.8944	33.25/0.9150	33.38/0.9182	34.01/0.9242	34.58/0.9280	34.70/0.9288	34.75/0.9290	34.81/0.9296	**34.89/0.9300**
Set14	26.38/0.7271	28.89/0.8105	28.13/0.7736	29.46/0.8244	29.63/0.8281	30.11/0.8364	30.53/0.8447	30.63/0.8462	30.68/0.8466	30.69/0.8473	**30.79/0.8487**
B100	26.33/0.6918	28.13/0.7740	28.13/0.7736	28.57/0.7893	28.65/0.7922	28.98/0.8009	29.23/0.8079	29.32/0.8093	29.33/0.8101	29.33/0.8104	**29.38/0.8111**
Urban100	23.52/0.6862	25.84/0.7856	25.70/0.7770	26.61/0.8136	26.77/0.8154	27.50/0.8370	28.46/0.8582	28.81/0.8647	28.83/0.8646	28.87/0.8656	**29.08/0.8688**
Manga109	25.46/0.8149	29.64/0.9003	29.47/0.8924	31.06/0.9234	31.15/0.9245	32.97/0.9391	33.97/0.9465	34.38/0.9483	34.46/0.9487	34.51/0.9493	**34.79/0.9505**

Table 3. Quantitative results with BD degradation model. Average PSNR/SSIM values for scaling factor $\times 3$. The best performance is shown **highlighted** and the second best underlined.

of our proposed PFAN, we compare our model with 8 state-of-the-art models: SRCNN [1], LapSRN [21], DRRN [3], EDSR [5], SRFBN[7], RDN [6], SAN [11] and DIN [13]. Self-ensemble strategy is also employed, denoted as PFAN+, to further improve our PFAN. Table 2 manifest all quantitative results with scaling factor $\times 3$. In general, our PFAN achieves comparable or superior results compared with all the other methods. For Set5, Set14, Urban, our PFAN achieves best results on both PSNR and SSIM, and best SSIM on Manga109, even without self-ensemble. Taking self-ensemble into consideration, our model gets all best results.

Zoomed visual comparisons of the same area are compared in Figure 5. Take "img076" from Urban100 of scale $\times 4$ for instance, heavy blur artifacts are produced from most methods. For early methods, even the basic structure cannot be recovered. Recent methods restores the basic outline while fail to get sharp details. Compared with high resolution (HR) ground-truth, our PFAN achieves faithful results.

3.4. Results with Blur-downscale Degradation (BD)

Followed by [10, 11], blur-downscale (BD) degradation model is also applied to our method. We compare our PFAN with 8 recent state-of-the-art SR models: SPMSR [22], SRCNN [1], VDSR [2], IRCNN [23], SRMD [24], RDN [6], RCAN [10] and SAN [11]. As is shown in Table 3, our PFAN achieves entirely better performance comparing with other methods even without self-ensemble. Specifically, we achieve 0.06dB PSNR gain over SAN on Set5. Considering B100, our PFAN gets the same result with SAN on PSNR, while achieves better SSIM.

3.5. Model Complexity Analyses

Table 4 illustrates the comparisons about model size and performance between 7 recent state-of-the-art methods:

Methods	DRRN	MemNet	EDSR	RDN	SAN	DIN	**PFAN**
Param.	297K	677K	43M	22M	15.7M	19.9M	11.9M
PSNR	33.23	33.28	33.92	34.01	34.07	34.03	**34.16**

Table 4. Parameter number (Param.), and PSNR (dB) comparisons. The PSNR values are based on Set14 with scaling factor $\times 2$.

DRRN [3], MemNet [4], EDSR [5], RDN [6], SAN [11] and DIN [13]. As is shown, DRRN and MemNet make use of fewer parameters, with a large sacrifice of model performance. Compared with RDN, SAN and DIN, our PFAN employs much fewer parameters while achieving superior performance, meaning that our model has a good trade-off between model performance and complexity, which signifies the effectiveness of our method.

4. CONCLUSION

In this paper, we propose a pyramid fusion attention network (PFAN) to recover HR images from given LR images by introducing a novel pyramid fusion attention module (PFA), which models the relationship between pixels for better enhancing the discriminative ability of SR network. Specifically, our PFA adopts pyramid fusion structure, stacked with modified residual blocks, using downscale operation and multi-scale fusion to expand receptive field and grasp more contextual information. Furthermore, we propose a progressive backward fusion strategy to make full use of hierarchical features produced by intermediate blocks. Plenty experiments on SR with BI and BD degradation models show the effectiveness of our PFAN over other state-of-the-art methods in terms of quantitative and visual results.

5. REFERENCES

[1] Chao Dong and Loy, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[2] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[3] Ying Tai, Jian Yang, and Xiaoming Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3147–3155.

[4] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu, "Memnet: A persistent memory network for image restoration," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4539–4547.

[5] Bee Lim and Son, "Enhanced deep residual networks for single image super-resolution," in *CVPRW*, 2017, pp. 136–144.

[6] Yulun Zhang and Tian, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[7] Zhen Li and Yang, "Feedback network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3867–3876.

[8] Xiaotong Luo and Xie, *LatticeNet: Towards Lightweight Image Super-Resolution with Lattice Block*, Springer, Cham, 2020.

[9] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu, "Residual feature aggregation network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2359–2368.

[10] Yulun Zhang and Li, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.

[11] Tao Dai and Cai, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11065–11074.

[12] Zheng Hui and Gao, "Lightweight image super-resolution with information multi-distillation network," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2024–2032.

[13] Feng Li, Runming Cong, Huihui Bai, and Yifan He, "Deep interleaved network for image super-resolution with asymmetric co-attention," *arXiv preprint arXiv:2004.11814*, 2020.

[14] Hanchao Li and Xiong, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[15] Yiqun Mei and Fan, "Pyramid attention networks for image restoration," *arXiv preprint arXiv:2004.13824*, 2020.

[16] Marco Bevilacqua and Roumy, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[17] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[18] David Martin and Fowlkes, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. IEEE, 2001, vol. 2, pp. 416–423.

[19] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja, "Single image super-resolution from transformed self-exemplars," in *CVPR*, 2015, pp. 5197–5206.

[20] Yusuke Matsui and Ito, "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools and Applications*, vol. 76, no. 20, pp. 21811–21838, 2017.

[21] Wei-Sheng Lai and Huang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *CVPR*, 2017, pp. 624–632.

[22] Tomer Peleg and Michael Elad, "A statistical prediction model based on sparse representations for single image super-resolution," *IEEE transactions on image processing*, vol. 23, no. 6, pp. 2569–2582, 2014.

[23] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang, "Learning deep cnn denoiser prior for image restoration," in *CVPR*, 2017, pp. 3929–3938.

[24] Kai Zhang, Wangmeng Zuo, and Lei Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *CVPR*, 2018, pp. 3262–3271.