IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 6, 2022, accepted March 5, 2022, date of publication March 11, 2022, date of current version March 24, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3158936

Lightweight Single Image Super-Resolution With Similar Feature Fusion Block

ZIRUI WANG[1], YUNMENG LIU[2], RUI ZHU[3], WENMING YANG[1], (Senior Member, IEEE), AND QINGMIN LIAO[1], (Senior Member, IEEE)

[1]Shenzhen International Graduate School/Department of Electronic Engineering, Tsinghua University, ShenZhen 518055, China
[2]Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China
[3]Faculty of Actuarial Science and Insurance, City, University of London, London EC1V 0HB, U.K.

Corresponding author: Wenming Yang (yangelwm@163.com)

This work was supported in part by the Natural Science Foundation of China under Grant 61771276; in part by the Natural Science Foundation of Guangdong Province under Grant 2020A1515010711; in part by the Special Foundations for the Development of Strategic Emerging Industries of Shenzhen under Grant JCYJ20200109143010272, Grant JCYJ20200109143035495, and Grant CJGJZD20210408092804011; and in part by the Overseas Cooperative Foundations.

ABSTRACT Convolutional neural network-based image super-resolution methods have achieved great success in recent years. However, the huge memory and computational costs make most of the existing methods difficult to be applied to resource-constrained scenarios such as edge devices. To tackle this problem, we propose a generic, lightweight and efficient feature fusion block to replace the commonly used 1*1 convolution. In addition, we propose the enhanced shallow residual blocks to improve the super-resolution performance. By combining these two novel blocks, we design an efficient similar feature fusion network for single image super-resolution, based on the observation that cross-layer features of the same channel usually have high similarities. The similar feature fusion block utilizes similarity as a guide for feature clustering, enabling efficient and high-performance cross-layer feature fusion. On the other hand, the enhanced shallow residual blocks are used as the base feature extraction model for the network to improve super-resolution performance in conjunction with the feature fusion module. In the enhanced shallow residual blocks, we combine convolution with identity connection to maintain the similarity of cross-layer features that are fed into the similar feature fusion block. The spatial attention mechanism is also introduced to reinforce the useful spatial features. Experimental results on the benchmark datasets show that the proposed method can achieve comparable results to state-of-the-art methods with a small number of parameters.

INDEX TERMS Convolutional neural networks, lightweight, image super-resolution, similar feature fusion.

I. INTRODUCTION

Single-image super-resolution (SISR) is a fundamental low-level computer vision task aiming at recovering high-resolution images from low-resolution images. SISR is an ill-posed problem since a low-resolution image can theoretically correspond to infinite high-resolution images, depending on the downsampling process. In recent years, methods based on convolutional neural networks have made significant progress in super-resolution (SR) task. Dong *et al.* [1] first proposed a three-layer super-resolution convolutional neural network (SRCNN). Since then, several super-resolution methods based on convolutional neural networks have been explored. Kim *et al.* [2] increased the network depth to 20 layers and achieved better performance than SRCNN. Benefit from the residual structure [3], the depth of recent network structures generally exceeds 100 layers [4]–[7]. Deeper network structures achieve better results, but they also entail larger number of parameters and computational costs, which are unaffordable in some resource-constrained scenarios. To balance the amount of computation and the visual effect, researchers have proposed many lightweight methods. Kim *et al.* proposed the deeply-recursive convolutional network (DRNN), by employing a parameter sharing strategy to reduce the number of parameters, but the computational cost didn't decrease accordingly [8]. Ahn *et al.* [9] designed a lightweight Cascade Residual Network (CARN) and used the group convolution

The associate editor coordinating the review of this manuscript and approving it for publication was Fan Zhang.

in its variant (CARN-M) to further reduce the computational cost. Hui *et al.* introduced an information distillation network(IDN) [21].

In the above networks, richer information is usually obtained by fusing multi-level features. 1×1 convolution is widely adopted, which can also reduce the number of feature map channels. However, as the number of channels decreases, the feature information decreases as well. Therefore, we aim to retain useful information as much as possible during feature fusion. The 1×1 convolution does not take advantage of the surrounding pixels or allow for adaptive weighting of features according to each channel's importance, which means that the contextual and global information is completely ignored. Therefore, a large number of useful features are usually lost during the feature compression process. On the contrary, 3×3 convolution has a receptive field and can utilize surrounding features. The contextual information it extracts can contribute to better results. Unfortunately, it's too computationally intensive.

This paper proposes an efficient similar feature fusion network (SFFN) to solve the problems mentioned above. We replace the commonly used 1×1 convolution by a similar feature fusion block (SFFB) consisting of one similar feature (SE) block [11], one channel shuffle, one group convolution, and one 1×1 convolution. The SE block adaptively adjusts features by explicitly modeling the interdependence between channels. We observe that the residual block's input and output features have a substantial similarity on the same channel when the network is shallow. Based on this observation, the same channels' features with different levels can be arranged together by a channel shuffle operation, followed by a group convolution for channel compression. Using group convolution with a kernel size of 3 can increase the receptive field, fully explore the surrounding features during feature compression, and preserve more valuable features. Besides, group convolution is less computationally intensive than normal convolution. However, only using group convolution will cause difficulties in information exchange between different groups. Thus in the end, an ordinary convolution is employed to fuse the information between channels. To fully exploit the potential of SFFB, we propose an enhanced shallow residual block (ESRB). In ESRB, we combine convolution with identity connection to improve the performance without increasing the number of parameters. Also, we add a spatial attention block to make the network focus on critical spatial contents.

Our main contributions are summarized as follows:

1) A generic lightweight feature fusion module, the similar feature fusion block, is proposed, which improves the feature utilization and reduces the feature information loss during the feature channel compression process by using the feature similarity, global information, and contextual information.
2) We also propose the enhanced shallow residual block, which enhances the super-resolution performance by introducing the identity connection and spatial attention to enhance the feature expression ability.
3) Building on similar feature fusion block and enhanced shallow residual block, we develop a lightweight network SFFN, which achieves state-of-the-art SR performance on the benchmark datasets.

II. RELATED WORK

In recent years, deep neural networks have achieved great success in computer vision tasks, such as image recognition and target detection [22]. It has also been applied to the SR task. SRCNN [1] has achieved comparable performance to traditional non-deep learning-based methods using a three-layer shallow neural network. Since deeper networks tend to deliver better results, many methods based on deeper networks have been proposed. Kim *et al.* [2] first introduced the residual network for training much deeper network architectures and achieved superior performance. Enhanced deep residual networks for super-resolution(EDSR) [4] improved the residual network architecture for the SR task by removing unnecessary modules such as batch normalization. Zhang *et al.* [5] further introduced the dense connection to surpass the performance of EDSR with fewer parameters. These methods significantly improved the image fidelity, as indicated by peak signal-to-noise ratio (PSNR) or structural similarity index (SSIM) [18]. However, besides these metrics, the low-power computing devices in the real-world also need to focus on other metrics such as the number of parameters, memory consumption, FLOP, latency time, etc.

Therefore, there is growing interest to build lightweight models that are accurate as well. Some lightweight yet effective networks for the SR task have been proposed. Fast super-resolution convolutional neural networks (FSRCNN) [23] are the first lightweight networks proposed for this task. FSRCNN directly applies the SR network to the low resolution (LR) images instead of the up-sampled LR images as SRCNN does. DRRN [8] utilizes the recursive layers to reduce the number of parameters while keeping the network's depth. CARN [9] reduces the computational cost by applying several residual connections and recursive layers. Hui *et al.* [21] proposed the information distillation network (IDN) that explicitly divides the preceding extracted features into two parts.

III. PROPOSED METHOD

This section will present our proposed network structure in detail, starting with SFFN in Section III-A, followed by SFFB in Section III-B and ESRB in Section III-C.

A. NETWORK STRUCTURE

The overall architecture of our proposed SFFN is shown in Figure 1. The network is a residual structured network consisting of the main feature extraction module ESRB, the similar feature fusion module SFFB, and the up-sampling layer. I_{lr} and I_{sr} denote the input and output of SFFN, respectively. Initially, we use a 3×3 convolution to extract shallow

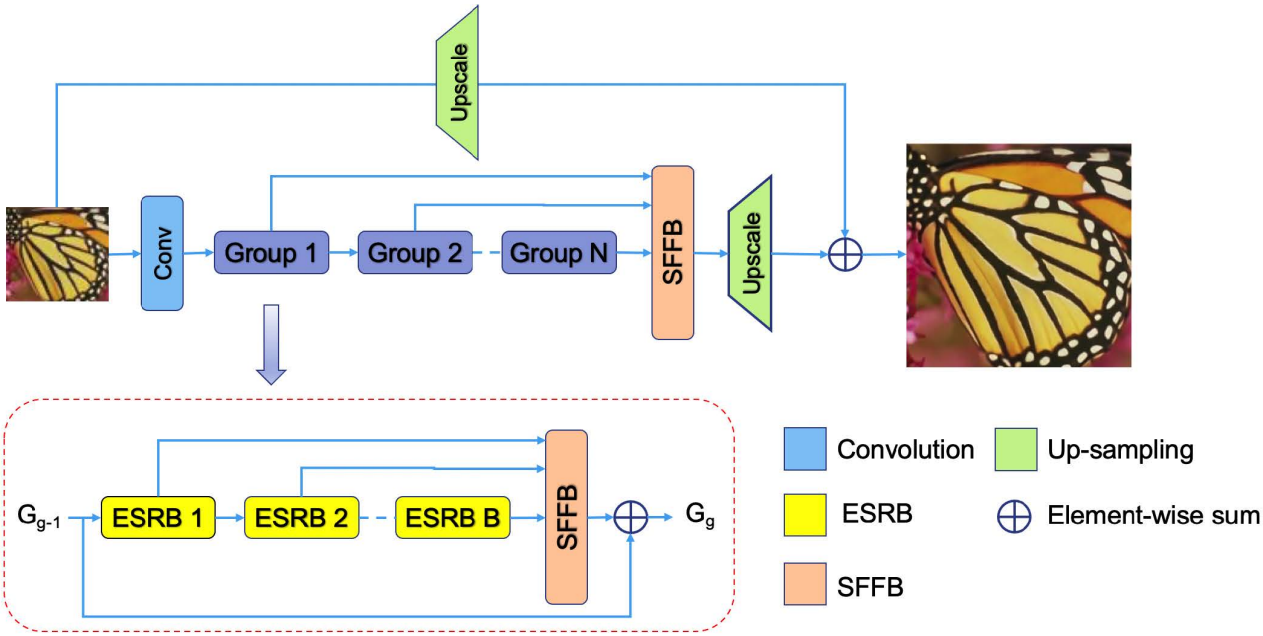


FIGURE 1. The architecture of our proposed lightweight similar feature fusion network (SFFN).

feature maps G_0 from the input image I_{lr}:

$$G_0 = F_{sf}(I_{lr}) \tag{1}$$

where F_{sf} denotes the first convolution operation, G_0 serves as the input to the followed feature fusion group, and SFFN contains G stacked feature fusion groups and a SFFB to fuse the output features of groups. Each feature fusion group is composed of B stacked ESRBs and a SFFB. The process for each feature fusion group can be represented as:

$$G_g = F_g(G_{g-1}) \tag{2}$$

$$= F_{sffb,g}([B_{g,1}, B_{g,2}, \ldots, B_{g,B}]) \tag{3}$$

where F_g denotes the function of g-th group, $F_{sffb,g}$ denotes the SFFB of the g-th group and $B_{g,k}$ is the output of the k-th ESRB inside the g-th group.

Following the feature fusion group is a global SFFB. It fuses the output features of every feature fusion group:

$$G = F_{sffb,last}([G_1, \ldots, G_N]) \tag{4}$$

where G is the output of SFFB. Finally, the output I_{sr} is given by:

$$I_{sr} = F_{up}(G) + F_{up,0}(I_{lr}) \tag{5}$$

where F_{up} and $F_{up,0}$ represent the upsampling module in the backbone network and the low-resolution image direct upsampling module, respectively. Similar structures have been used in networks such as WDSR [24] and AWSRN [25]. In this paper, we utilize pixel-shuffle layer.

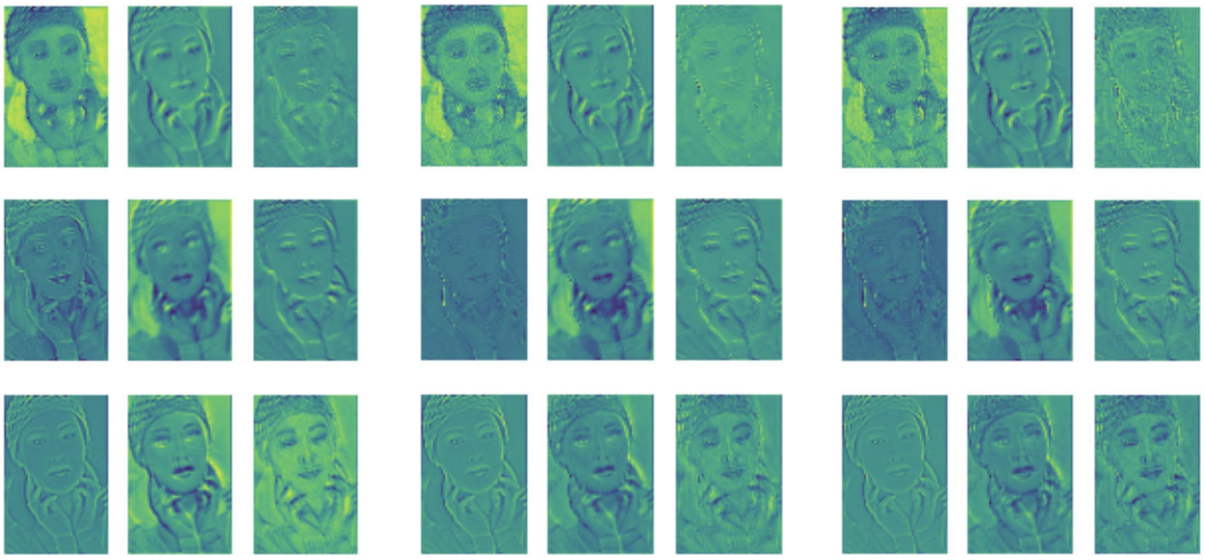
B. SIMILAR FEATURE FUSION BLOCK

This section elaborates on the proposed SFFB, designed to fuse multi-level features. The drawback of the commonly used 1×1 convolution is the small receptive field, which does not allow exploring the contextual and global information. Therefore, lots of useful features are lost during channel compression. In SFFB, we manage to incorporate more contextual and global information without increasing the computational cost.

The input and output feature maps of the residual structure blocks have high similarity in corresponding channels due to the identity connection. Figure 2 shows the visualized output feature maps of different feature fusion groups. Each channel's feature is arranged in order and the same order presents the same channel. It can be observed that the features of the same channel are more similar than other channels visually. To further demonstrate this similarity quantitatively, we calculate the PSNRs between features in Figure 2. Table 1 shows the PSNR between the i-th channel output feature of group N and the j-th channel output feature of the other groups. The table presents the i-th channel of output features of group N horizontally, while the j-th channel of the other groups vertically. Bolded font indicates the highest PSNR per column.

The results in Table 1 demonstrate that the cross-layer feature similarity in the same channel in our network structure is indeed relatively higher than those in different channels. Based on this observation, we propose to use the 3×3 group convolution to fuse multi-level features of the same channel, which reduces the computation, increases the receptive field, effectively utilizes similar features and alleviates the information loss caused by channel compression.

As shown in Figure 3(b), SFFB comprises the SE Block, channel shuffle, group convolution and 1×1 convolution. SE block adaptively adjusts features by explicitly modeling the interdependence between channels. Important channel



(a) Feature map of group 1 (b) Feature map of group 2 (c) Feature map of group 3

FIGURE 2. The output feature maps of different groups.

TABLE 1. The PSNR between different features.

Group 1

	1	2	3	4	5	6
1	**30.5728**	28.8340	28.853	28.6862	30.1447	28.7037
2	29. 1392	**33.3892**	31.7067	30.7315	29.912	31.5146
3	29.3430	30.0034	**32.4780**	29.2847	29.1197	29.5102
4	29.0515	29.9932	30.2182	**31.6626**	30.0232	30.7189
5	28.9376	29.7230	29.9853	30.4797	**32.1482**	30.2307
6	29.0916	31.3240	30.1447	30.8161	29.8687	**33.3938**

Group 2

	1	2	3	4	5	6
1	**30.5247**	29.0459	29.2760	28.9101	28.8809	28.9695
2	28.7221	**33.8213**	29.7756	30.3149	29.8110	31.4342
3	28.8582	30.9392	**30.7639**	29.8315	29.5916	31.3508
4	28.7027	30.3773	29.1645	**32.9472**	30.3179	30.5242
5	28.6124	29.9694	29.0071	30.1667	**32.3775**	29.9673
6	28.6824	31.4975	29.4396	30.6420	29.9830	**34.0008**

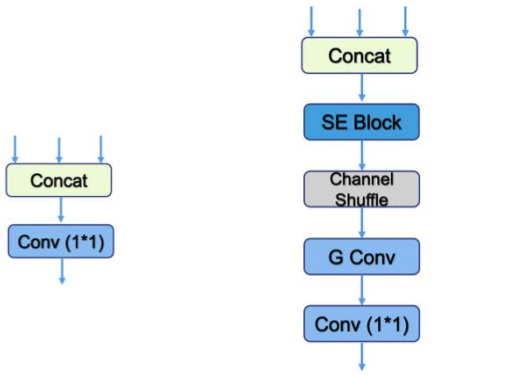
Group 3

	1	2	3	4	5	6
1	**30.3954**	28.9036	29.1192	28.9137	28.7689	28.8939
2	28.9223	**32.7142**	30.0044	29.9346	29.8978	31.36703
3	29.0469	30.5441	**30.9151**	29.5975	29.5236	30.8597
4	28.7559	30.6105	29.3310	**32.5701**	30.3535	30.5647
5	28.8999	29.9284	29.2425	30.0481	**31.9986**	29.8225
6	28.7931	31.4944	29.8032	30.3498	30.1688	**33.4795**

feature maps are given more attention and vice versa. In Figure 3, G Conv is the shorthand for group convolution and SE block is the shorthand for the squeeze and extraction block. The output of the SE block f_{se} is calculated as

$$f_{se} = F_{SE}([B_{g,1}, \ldots, B_{g,B}]) \tag{6}$$

where F_{SE} denotes the function of the SE Block. Channel shuffle is used for feature channel rearrangement. The same channel of multi-level features is placed together to facilitate subsequent group convolution. We use group convolution as our channel compression strategy. Compared to 1×1 convolution, 3×3 group convolution allows more exploration of the feature map, provides better feature retention for every



(a) 1×1 convolution (b) similar feature fusion block

FIGURE 3. A comparison between the 1*1 convolution and similar feature fusion block.

channel and larger receptive field. The compressed feature map $f_{compress}$ is calculated as

$$f_{compress} = F_{group}(CS(f_{se})) \tag{7}$$

where F_{group} and CS refer to group convolution and channel shuffle, respectively. Group convolution reduces the amount of computation at the expense of losing the exchange of information between groups. Despite the fact that inter-channel information has been mixed during channel shuffle operation, it is not sufficient. We further employ an ordinary 1×1 convolution convolution to perform feature fusion between channels on the compressed features to compensate to the loss caused by group convolution. The output feature map of SFFB G_g is calculated as

$$G_g = F_{conv1}(f_{compress}) \tag{8}$$

where F_{conv1} denotes 1×1 convolution.

Standard convolutions have the computational cost of: $C_{in} \times C_{out} \times H \times W \times K \times K$. C_{in} and C_{out} denote the number of channels in the input and output feature map, respectively. H and W represent the size of the output feature map. K refers to the size of the convolution kernel. In SFFB, the computation is mainly for group convolution and ordinary convolution; therefore, we omit the SE Block to be concise. The ratio of the SFFB computation cost $Cost_{SFFB}$ to the 1×1 convolution computation cost $Cost_{conv1}$ is:

$$\frac{Cost_{SFFB}}{Cost_{conv1}} = \frac{\frac{C_{in} \times C_{out} \times H \times W \times 3 \times 3}{g}}{C_{in} \times C_{out} \times H \times W \times 1 \times 1} + \frac{C_{out} \times C_{out} \times H \times W \times k \times k}{C_{in} \times C_{out} \times H \times W \times 1 \times 1} \quad (9)$$

where g represents the number of groups and k denotes kernel size of last convolution. If we set $C_{in} = 128, g = 32$, $C_{out} = 32, k = 1$, the ratio will turn into:

$$\frac{Cost_{SFFB}}{Cost_{conv1}} = \frac{17}{32} \approx 0.53. \quad (10)$$

C. ENHANCED SHALLOW RESIDUAL BLOCK

To achieve a better super-resolution result, we further design an enhanced shallow residual block, as shown in Figure 4. ESRB consists of three parts: the feature refinement layer, SFFB and ESA. The feature refinement layer consists of several stacked convolutions to gradually refine the features. During the refining process, the corresponding similarity among multi-level features may decrease, which will affect the fusion effect of SFFB. Thus we integrate convolution with identity connection, which facilitates the maintenance of feature similarity on multi-level features in blocks and ensures that multi-level features' similarity is maintained throughout the network. Therefore, by introducing identity connections, ESRB and SFFB can be more compatible without additional parameters.

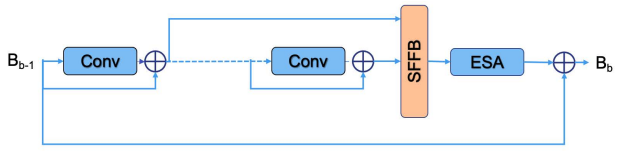


FIGURE 4. The structure of the enhanced shallow residual block.

As we all know, it is crucial for low-level computer vision tasks such as super-resolution to recover the textures. Although the SE block can model channel interdependence and adaptively weigh each channel's features, it cannot focus on the feature maps' spatial textures. Therefore, we introduce the enhanced spatial attention (ESA) [12] at the end of ESRB to enhance the detailed texture in the fusion block's output features. The structure of ESA is shown in Figure 5. ESA starts with 1×1 convolution for feature channel compression to reduce the feature map channel. The feature map then goes through a striding convolution (stride = 2) and a max-pooling layer successively to reduce the feature map's resolution. These three layers ensure that ESA is a lightweight module. After the feature map is filtered by several convolution layers, we use an up-sampling layer to increase the spatial resolution followed by a 1×1 convolution to increase the number of channels. Finally, the attention mask is generated in the sigmoid layer. The benefit from the stridden convolution and maxpooling layer is that ESA has a large receptive field to fully explore the spatial features.

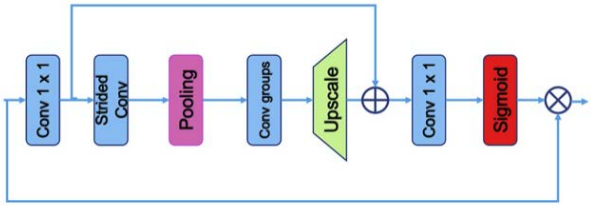


FIGURE 5. The structure of the enhanced spatial attention.

IV. EXPERIMENTS

A. DATASETS AND METRICS

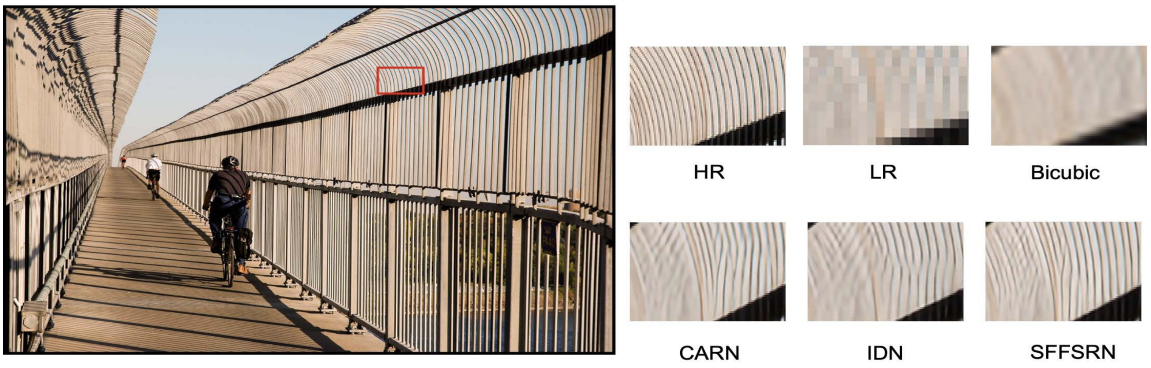
We use DIV2K [13] as the training set, which contains 800 high-resolution (HR) images. Low-resolution (LR) images are obtained from the HR images by bicubic down-sampling. As for testing, we use four standard and widely used benchmark datasets: Set5 [14], Set14 [15], B100 [16] and Urban100 [17]. We use PSNR and SSIM [19] to measure the results on the Y-channel of the transformed YCbCr color space for a fair comparison.

B. IMPLEMENTATION DETAILS

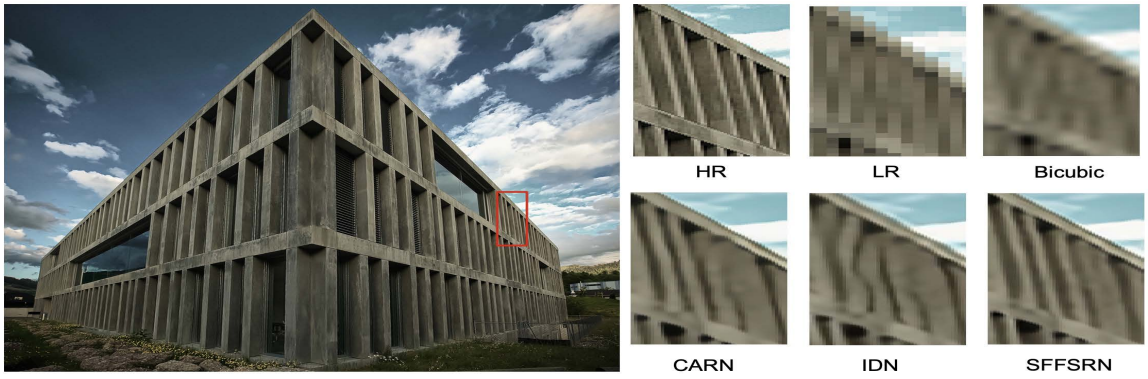
Our proposed method SFFN includes a total of 4 feature fusion groups, each of which contains 4 ESRBs. The number of feature map channels is 32. All convolutions in the network are initialized using the method of He *et al.* [19]. To achieve better performance, we chose the L_1 loss as our loss function. Our model is trained with the ADAM optimizer by setting $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate is set to 1×10^{-3}, which decreases by half every 200 epochs. The batch-size and patch-size of our input images are 16 and 48×48, respectively. We performed standard data augmentation on the training data set, including horizontal flip, vertical flip, and random rotation of 90°, 180°, and 270°.

C. COMPARISONS WITH STATE-OF-THE-ART METHODS

We compared SFFN with five state-of-the-art SISR methods, including DRRN, MemNet, CARN, IDN, and MADNet. The results are shown in Table 2. The parameters of the models and multi-adds are also provided for the comparison of memory usage and computational expenditures. Multi-adds are estimated on 720p HR images. The results show that our proposed method achieves the best or second-best results on all data compared to other SR methods, with the second-lowest multi-adds. Our method even outperforms MemNet and DRRN, which are ten times more computationally expensive than ours. The results show that our approach is simple yet effective.

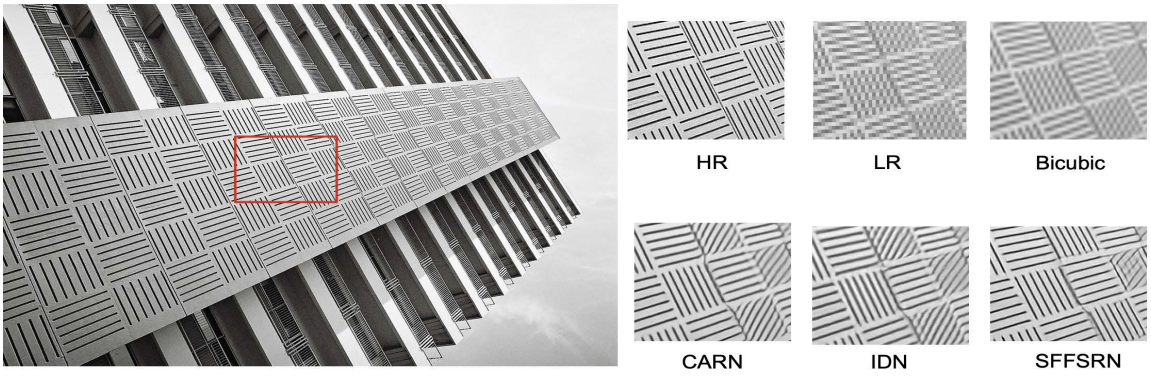


(a) img024



(b) img027

(c) img062



(d) img092

FIGURE 6. Visual comparison for ×4 SR on "img024", "img027", "img062", "img092" from the Urban100 Dataset.

TABLE 2. Quantitative comparisons of existing methods on four datasets. Red/blue text: best/second-best.

Network	Scale	Params(K)	Multi-Adds(G)	Set5	Set14	B100	Urban100
				PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
DRRN [8]	2	297	6796.9	37.74 / 0.9591	33.23 / 0.9136	32.05 / 0.8973	31.23 / 0.9188
MemNet [20]	2	677	2,662.4	37.78 / 0.9597	33.28 / 0.9142	32.08 / 0.8978	31.31 / 0.9195
IDN [21]	2	579	124.6	37.85 / 0.9598	33.58 / 0.9178	32.11 / 0.8989	31.95 / 0.9266
CARN [9]	2	1,592	222.8	37.76 / 0.9590	33.52 / 0.9166	32.09 / 0.8978	31.92 / 0.9256
CARN-M [9]	2	412	91.2	37.53 / 0.9583	33.26 / 0.9141	31.92 / 0.8960	31.23 / 0.9193
MADNet [10]	2	878	187.1	37.94 / 0.9604	33.46 / 0.9167	32.10 / 0.8988	31.74 / 0.9246
SFFN(ours)	2	912	138.7	38.02 / 0.9606	33.59 / 0.9177	32.20 / 0.9000	32.34 / 0.9298
DRRN	3	297	6796.9	34.03 / 0.9244	29.96 / 0.8349	28.95 / 0.8004	27.53 / 0.8378
MemNet	3	677	2,662.4	34.09 / 0.9248	30.00 / 0.8350	28.96 / 0.8001	27.56 / 0.8376
IDN	3	588	56.3	34.24 / 0.9260	30.27 / 0.8408	29.03 / 0.8038	27.99 / 0.8489
CARN	3	1,592	118.8	34.29 / 0.9255	30.29 / 0.8407	29.06 / 0.8034	28.06 / 0.8493
CARN-M	3	412	46.1	33.99 / 0.9236	30.08 / 0.8367	28.91 / 0.8000	27.55 / 0.8385
MADNet	3	930	88.4	34.26 / 0.9262	30.29 / 0.8410	29.04 / 0.8033	27.91 / 0.8464
SFFN(ours)	3	916	69.35	34.42 / 0.9274	30.34 / 0.8419	29.11 / 0.8055	28.26 / 0.8543
DRRN	4	297	6796.9	31.68 / 0.8888	28.21 / 0.7720	27.38 / 0.7284	25.44 / 0.7638
MemNet	4	677	2,662.4	31.74 / 0.8893	28.26 / 0.7723	27.40 / 0.7281	25.50 / 0.7630
IDN	4	600	32.3	31.99 / 0.8928	28.52 / 0.7794	27.52 / 0.7339	25.92 / 0.7801
CARN	4	1,592	90.9	32.13 / 0.8937	28.60 / 0.7806	27.58 / 0.7349	26.07 / 0.7837
CARN-M	4	412	32.5	31.92 / 0.8903	28.42 / 0.7762	27.44 / 0.7304	25.62 / 0.7694
MADNet	4	1002	54.1	32.11 / 0.8939	28.52 / 0.7799	27.52 / 0.7340	25.89 / 0.7782
SFFN(ours)	4	923	34.6	32.23 / 0.8950	28.58 / 0.7813	27.56 / 0.7361	26.15 / 0.7877

TABLE 3. The ablation study of SFFB and ESRB for the scale factor ×4 on the Urban100 dataset.

	SFFN-NF	SFFN-NB	SFFN
SFFB	✗	✓	✓
ESRB	✓	✗	✓
Multi-adds	34.6G	33.9G	34.6G
PSNR/SSIM	26.09/0.7854	26.05/0.7841	26.15/0.7877

We also provide a visual comparison, shown in Figure 6. It can be observed that our method generates the most visually pleasing images, which contain more plausible textures. Compared to other methods, our super-resolution results are the clearest with the least errors.

D. MODEL ANALYSIS

Table 3 shows the results of the ablation experiments of SFFN. In SFFN-NF, we replace SFFB by 1×1 convolution. SFFN-NB replaces ESRB by 4 cascaded 3×3 convolutions and an SFFB. We calculate the number of operations (multi-adds) as the computational complexity. Multi-adds are estimated on 720p HR images. Compared with SFFN-NF, SFFN improves the super-resolution recovery accuracy without increasing the computational operations, which indicates that SFFB can indeed retain more useful features than 1×1 convolution does during feature compression. Meanwhile, the performance of SFFN-NB is not as good as that of SFFN, which indicates that ESRB can indeed cooperate better with SFFB and extract more effective feature maps. Furthermore, we test the results of removing the identity connection from the network ESR's refining process. On the Urban100 dataset with the scale factor of 4, the corresponding PSNR is 26.09, which is 0.06 lower than the original model, while the SSIM is 0.7854, which is 0.0023 lower than the original model. These results demonstrate the importance of including the identity connection.

V. CONCLUSION

This paper proposes a novel lightweight, similar feature fusion network for single image super-resolution. Our approach focuses on using similarity in feature maps to reduce the information loss and computational effort during feature fusion. Specifically, the proposed similar feature fusion block incorporates more contextual and global information without increasing the computational effort by utilizing the similarity between features to group and fuse multi-level features and channel attention mechanism. We also propose an enhanced shallow residual block as the base module of the whole network, which contains the convolution with the identity connection to reinforce multi-level feature similarity, the similar feature fusion block to fusion features, and the spatial attention module to enhance detailed spatial features. Extensive experiments on benchmark datasets illustrate the effectiveness of our SFFN in image super-resolution. Dynamic clustering can be tried in subsequent work to group cross-layer features in order to improve feature fusion and enhance super-resolution results.

REFERENCES

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2015.

[2] 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.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

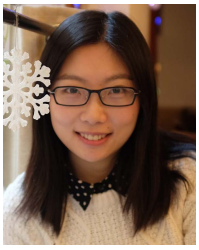
[4] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.

[5] 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.

[6] 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.
[7] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.
[8] 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.
[9] 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.
[10] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "MADNet: A fast and lightweight network for single-image super resolution," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1443–1453, Mar. 2020.
[11] 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.
[12] 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. 2359–2368.
[13] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 114–125.
[14] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L.-A. Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. Procedings Brit. Mach. Vis. Conf.*, 2012, pp. 1–10.
[15] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surf.* Cham, Switzerland: Springer, 2010, pp. 711–730.
[16] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Jul. 2001, pp. 416–423.
[17] 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.
[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[20] 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.
[21] Z. Hui, X. Wang, and X. Gao, "Fast and accurate single image super-resolution via information distillation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 723–731.
[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
[23] 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.
[24] J. Yu, Y. Fan, J. Yang, N. Xu, Z. Wang, X. Wang, and T. Huang, "Wide activation for efficient and accurate image super-resolution," 2018, *arXiv:1808.08718*.
[25] C. Wang, Z. Li, and J. Shi, "Lightweight image super-resolution with adaptive weighted learning network," 2019, *arXiv:1904.02358*.
[26] W. Yang, W. Wang, X. Zhang, S. Sun, and Q. Liao, "Lightweight feature fusion network for single image super-resolution," *IEEE Signal Process. Lett.*, vol. 26, no. 4, pp. 538–542, Apr. 2019.
[27] P. Shamsolmoali, M. Zareapoor, E. Granger, H. Zhou, R. Wang, M. E. Celebi, and J. Yang, "Image synthesis with adversarial networks: A comprehensive survey and case studies," *Inf. Fusion*, vol. 72, pp. 126–146, Aug. 2021.
[28] D. Song, C. Xu, X. Jia, Y. Chen, C. Xu, and Y. Wang, "Efficient residual dense block search for image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 12007–12014.
[29] Y. Wang, L. Wang, H. Wang, and P. Li, "Resolution-aware network for image super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1259–1269, May 2019.

ZIRUI WANG was born in Heilongjiang, China, in 1995. He received the B.S. degree in communication engineering from Sun Yat-sen University, Guangzhou, China, in 2017. He is currently pursuing the master's degree with the Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. His current research interests include single image super-resolution and generative adversarial networks.

YUNMENG LIU was born in Jiangsu, China, in 1978. He received the Graduate degree from the School of Mechatronics Engineering, Harbin Institute of Technology, in 2005. Since 2005, he has been working with the Shanghai Institute of Technical Physics, CAS. He was promoted to a Professor, in 2015. He is mainly engaged in the research of photoelectric remote sensing and application technology, target detection, and identification technology.



RUI ZHU received the Ph.D. degree in statistics from University College London, in 2017. She is currently a Senior Lecturer with the Faculty of Actuarial Science and Insurance, City, University of London. Her research interests include spectral data analysis, hyperspectral image analysis, subspace-based classification methods, and image quality assessment.

WENMING YANG (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Zhejiang University, in 2006. He is currently an Associate Professor with the Shenzhen International Graduate School/Department of Electronic Engineering, Tsinghua University. His research interests include image processing, pattern recognition, computer vision, and AI in medicine.

QINGMIN LIAO (Senior Member, IEEE) received the Ph.D. degree in signal processing and telecommunications from the University of Rennes 1, France, in 1994. Since 2002, he has been a Professor with the Department of Electronic Engineering, Tsinghua University. Since 2010, he has been the Director of the Division of Information Science and Technology, Graduate School at Shenzhen, Tsinghua University. His research interests include image/video processing, transmission, analysis, biometrics, and their applications.