OSFFNet: Omni-Stage Feature Fusion Network for Lightweight Image Super-Resolution

Yang Wang, Tao Zhang*

School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
yangwang@stu.jiangnan.edu.cn, taozhang@jiangnan.edu.cn

Abstract

Recently, several lightweight methods have been proposed to implement single-image super-resolution (SISR) on resource-constrained devices. However, these methods primarily focus on simplifying network structures without the full utilization of shallow features. The fact remains that shallow features encompass crucial details for the super-resolution task, including edges, textures, and colors. Therefore, developing a novel architecture that can effectively integrate features from different levels and capitalize on their mutual complementarity is necessary. We first analyze the relationship between multi-stage features and the restoration tasks in a classic lightweight SR method. Based on these observations, we propose an Omni-Stage Feature Fusion (OSFF) architecture, which incorporates Original Image Stacked Initialisation, Shallow Feature Global Connection, and Multi-Receptive Field Dynamic Fusion. An Attention-Enhanced Feature Distillation module is also designed to enhance the model performance. Finally, leveraging these contributions, we construct an Omni-Stage Feature Fusion Network (OSFFNet). Through extensive experiments on various benchmark datasets, the proposed model outperforms state-of-the-art methods. Notably, it achieves a 0.26dB PSNR improvement over the second-best method for ×2 SR on the Urban100 dataset.

Introduction

As a typical branch of low-level vision methods, single-image super-resolution (SISR) aims to reconstruct a high-resolution (HR) image from a degraded low-resolution (LR) image. Compared to traditional methods, deep learning (DL) methods have achieved outstanding performance and realistic visual effects thanks to their learnable feature representations (Wang, Chen, and Hoi 2020). As a result, some studies have aimed to improve image restoration quality by increasing convolutional layers or adopting complex network topologies. These characteristics often restrict practical applications as resource-constrained mobile devices struggle with intensive SR methods. Therefore, the design of lightweight and efficient SR models has become essential and also a challenging problem.

Researchers have proposed several lightweight and efficient SR methods to address this challenge. The main idea of these methods is to design efficient network structures, which involve strategies such as parameter sharing, feature distillation, attention mechanisms, and convolution optimization. These SR methods have successfully balanced reconstruction performance and computational cost. However, these networks often rely on sequentially simplifying network structures to improve computational efficiency, which leads to the underutilization of low-level features during the forward propagation process. Figure 1 illustrates how feature maps in IMDN (Hui et al. 2019) contain detailed information, with low-level features emphasizing edges and textures near the network input, while deep-level features focus on extracting local features with fewer details near the output. Neglecting low-level features may lead to the loss of crucial details, affecting SR model performance. Therefore, a new network architecture needs to be designed to effectively integrate feature maps from different stages and leverage their complementarity. Such an architecture will be able to maintain computational efficiency and enhance SR performance. To achieve this, we propose a novel Omni-Stage Feature Fusion (OSFF) architecture to efficiently fuse multi-level features. OSFF includes Original Image Stacked Initialization (OISI), Shallow Feature Global Connection (SFGC), and Multi-Receptive Field Dynamic Fusion (MFDF). OISI duplicates and stacks the original image multiple times along the channel dimension to enrich details in shallow features. SFGC combines shallow features with

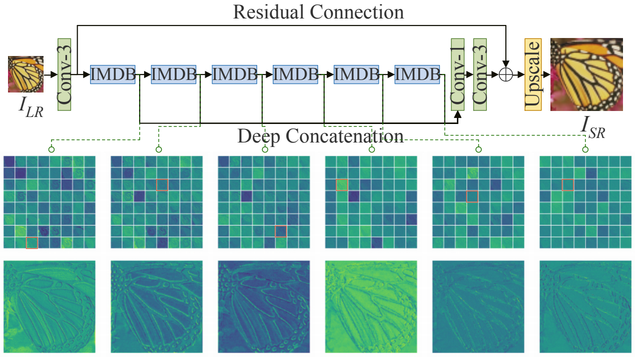


Figure 1: Visualization of feature maps at different stages in IMDN.

[*] Corresponding author

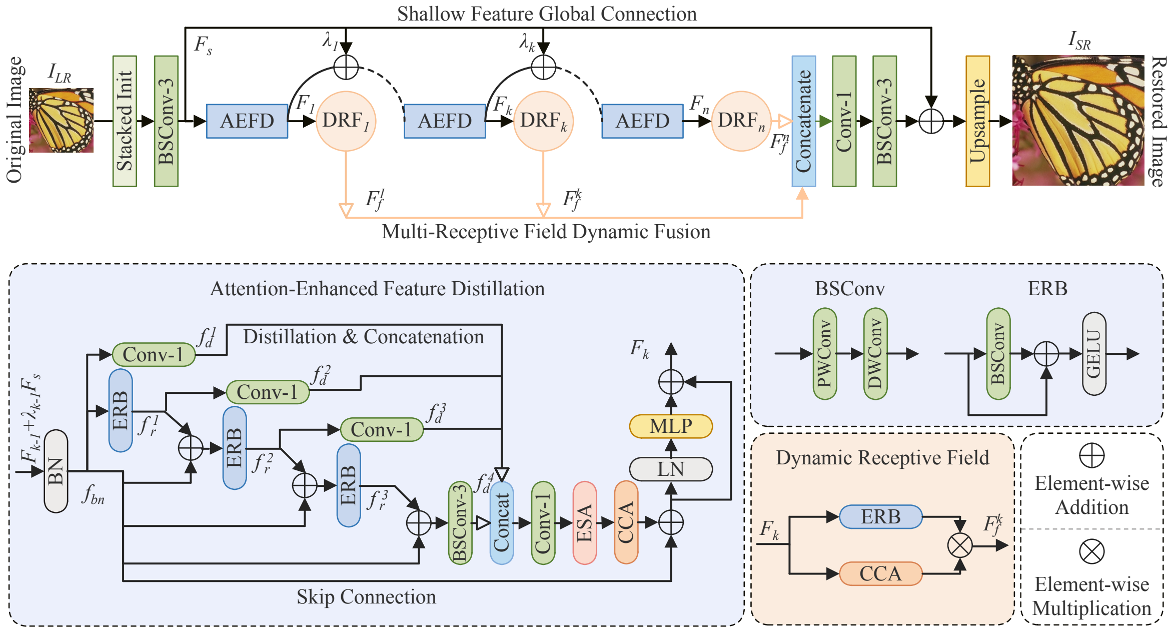


Figure 2: The overall architecture of Omni-Stage Feature Fusion Network (OSFFNet).

multi-stage features through element-wise addition to stabilize the training process and utilize texture details. MFDF integrates different-stage features with dynamic receptive-field modules to enhance and merge features from different stages. Furthermore, we improve model performance by designing an Attention-Enhanced Feature Distillation (AEFD) module based on IMDB (Hui et al. 2019), which has shown effectiveness in previous super-resolution tasks. We also integrate Blueprint Separation Convolutions (BSConv) (Haase and Amthor 2020) to reduce redundancy. Building on these achievements, we develop an Omni-Stage Feature Fusion Network (OSFFNet) for lightweight image super-resolution, which outperforms current state-of-the-art methods based on experimental results on benchmark datasets, especially with significant improvements on the Urban100 dataset. The main contributions of this paper are as follows:

- We propose an Omni-Stage Feature Fusion architecture that enhances the contribution of shallow features by employing Original Image Stacked Initialization, Shallow Feature Global Concatenation, and Multi-Stage Feature Dynamic Fusion. This enables effective complementarity among multiple feature levels.
- We design an Attention-Enhanced Feature Distillation module and integrate Blueprint Separation Convolution as the core building block. This combination enhances network performance while maintaining lightweight computational overhead.
- Our extensive experiments on benchmark datasets confirm the effectiveness of our methods. The results show that OSFFNet reaches a new state-of-the-art performance in lightweight SISR, with significant improvements observed on the Urban100 dataset in particular.

Related Work

SISR for Reconstruction Quality

Dong et al. (Dong et al. 2014) initially introduced SRCNN, a three-layer CNN for LR to HR image mapping. SRCNN demonstrated significant improvements over traditional methods in SR reconstruction, inspiring the development of subsequent CNN-based methods (Lepcha et al. 2023) for further enhancement.

VDSR (Kim, Lee, and Lee 2016a) utilized residual learning with stacked convolutional layers to improve SR accuracy. EDSR (Lim et al. 2017) introduced a deeper and wider residual network to enhance feature representation capability. However, residual learning-based methods have limitations and inefficiencies. To address these, MSRN (Li et al. 2018) introduced a multi-scale residual network to adaptively detect image features at different scales, improving feature representation. SRDenseNet (Tong et al. 2017) introduced dense connections to address gradient vanishing and feature information loss, enabling effective feature propagation and reuse between layers.

As the networks become larger and deeper, the introduction of various attention mechanisms has become another trend in deep super-resolution research. For instance, RCAN (Zhang et al. 2018) used channel attention, PAN (Zhao et al. 2020) employed pixel attention, SAN (Dai et al. 2019) utilized second-order attention, and ENLCN (Xia et al. 2022) utilized efficient non-local sparse attention. Additionally, self-attention mechanisms have shown remarkable performance in image reconstruction. SwinIR (Liang et al. 2021) leveraged Swin Transformer (Liu et al. 2021) architecture, multi-scale feature representation, hybrid attention mechanisms, and local-global feature interactions to achieve ex-

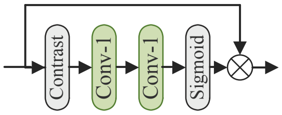


Figure 3: The architecture of CCA block.

cellent SISR. HAT (Chen et al. 2023) combined channel attention and self-attention mechanisms to fully exploit the global information induction ability of channel attention and the powerful representation capability of self-attention, ultimately aiming to activate more pixel information.

SISR for Computational Efficiency

To meet the demands of edge devices, it is imperative to develop lightweight and efficient SR models (Gendy, He, and Sabor 2023). Many researchers have designed various lightweight SR algorithms to reduce parameter count and computational complexity. Initially, researchers identified the computational redundancy in the pre-upsampling stage of SRCNN and began designing lightweight models using a post-upsampling infrastructure. For example, FSRCNN (Dong, Loy, and Tang 2016) broke the computational bottleneck by restoring feature maps to the desired high-resolution size using deconvolution layers at the end of the network. ESPCN (Shi et al. 2016) achieved efficient image super-resolution reconstruction in the pixel reconstruction stage using sub-pixel convolution without the need for additional interpolations, resulting in fast and high-quality reconstructions. Therefore, sub-pixel convolution has been widely adopted in lightweight super-resolution.

LapSRN (Lai et al. 2017) leveraged a pyramid structure and skip connections to utilize multi-scale image information, and gradually improved resolution through hierarchical reconstruction. DRCN (Kim, Lee, and Lee 2016b) introduced recursive learning into super-resolution and reduced parameter count through weight-sharing strategies. CARN (Ahn, Kang, and Sohn 2018) adopted a cascading mechanism based on residual learning and used grouped convolution instead of standard convolution to reduce the number of model parameters. IDN (Hui, Wang, and Gao 2018) employed an information distillation strategy to better utilize hierarchical features by independently processing the current feature map. IMDN (Hui et al. 2019) improved model efficiency through more effective feature distillation mechanisms and efficient adaptive cropping strategies.

Recently, researchers have been optimizing the convolutional approach to develop lighter and more efficient SR models. For instance, ECBSR (Zhang, Zeng, and Zhang 2021) applied the structure reparameterization technique from RepVGG (Ding et al. 2021) to effectively extract edge and texture information. FMEN (Du et al. 2022) prioritized memory efficiency and utilized structure reparameterization to further accelerate network inference. BSRN (Li et al. 2022) introduced blueprint separation convolution, a variant of depthwise separable convolution, combined with residual learning to achieve efficient super-resolution.

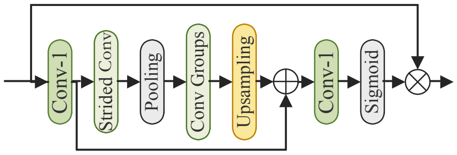


Figure 4: The architecture of ESA block.

Proposed Method

Overall Network Architecture

The overall network architecture of the Omni-Stage Feature Fusion Network (OSFFNet) is shown in Figure 2. OSFFNet consists of three main stages: an initial feature extraction, a multi-stage feature extraction, and a high-resolution reconstruction. Here, I_{LR} represents the original image input to OSFFNet, and I_{SR} represents the restored image. Prior to the high-resolution reconstruction stage, the feature maps maintain the same resolution as I_{LR}. We perform stack initialization on the original image in the initial feature extraction stage. Specifically, we replicate and stack the input original image multiple times along the channel dimension. Then, we apply a 3×3 Blueprint Separation Convolution (BSConv) to extract shallow features, allowing for a richer representation of texture details in the shallow feature maps. This process can be expressed as:

$$F_s = H_{SF}(Concat_{i=1}^{o} I_{LR}), \tag{1}$$

where H_{SF} represents shallow feature extraction using a 3×3 BSConv (Haase and Amthor 2020). $Concat$ denotes channel-wise concatenation, and o indicates the number of copies of the original image. Next, F_s extracts multi-stage features using AEFDs and Dynamic Receptive Field (DRF) modules. This process can be expressed as:

$$F_k = \begin{cases} H_{AEFD}(F_s) & k = 1 \\ H_{AEFD}(F_{k-1} + \lambda_{k-1} F_s) & 1 < k \leq n \end{cases}, \tag{2}$$

$$F_f^k = H_{DRF}^k(F_k), \quad 1 \leq k \leq n\,, \tag{3}$$

where H_{AEFD} denotes an AEFD module, H_{DRF}^k is the k-th DRF module. λ denotes a learnable weight parameter, and n is the number of AEFDs. Then, the F_f^ks are concatenated and fused. This process can be expressed as:

$$F_{fused} = H_{Fusion}(Concat_{k=1}^{n} F_f^k), \tag{4}$$

where H_{fusion} denotes the fusion operation which consists of a 1×1 convolution and a 3×3 BSConv. Lastly, the reconstruction module is applied to generate the final high-resolution image I_{SR} as:

$$I_{SR} = H_{rec}(F_{fused} + F_s), \tag{5}$$

where H_{rec} denotes the reconstruction which consists of a 3×3 convolution and a sub-pixel operation (Shi et al. 2016).

Method	Scale	#Params	#Multi-Adds	Set5		Set14		BSD100		Urban100		Manga109	
		[K]	[G]	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SRCNN	×2	8	52.7	36.66	0.9542	32.45	0.9067	31.36	0.8879	29.50	0.8946	35.60	0.9663
VDSR		666	612.6	37.53	0.9587	33.03	0.9124	31.90	0.8960	30.76	0.9140	37.22	0.9750
EDSR		1335	307.9	37.99	0.9604	33.57	0.9175	32.16	0.8994	31.98	0.9272	35.85	0.9436
CARN		1592	222.8	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
IMDN		694	158.8	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283	38.88	0.9774
HNCT		357	82.4	38.08	0.9608	33.65	0.9182	32.22	0.9001	32.22	0.9294	38.87	0.9774
FMEN		748	172.0	38.10	0.9609	**33.75**	0.9192	32.26	0.9007	32.41	0.9311	38.95	0.9778
BSRN		332	73.0	38.10	0.9610	33.74	**0.9193**	32.24	0.9006	32.34	0.9303	**39.14**	**0.9782**
RepRFN		386	85.12	38.07	**0.9612**	33.63	0.9184	32.22	0.9009	32.10	0.9274	39.00	0.9779
OSFFNet		516	83.2	**38.11**	0.9610	33.72	0.9190	**32.29**	**0.9012**	**32.67**	**0.9331**	39.09	0.9780
SRCNN	×3	8	52.7	32.75	0.9090	29.30	0.8215	28.41	0.7863	26.24	0.7989	30.48	0.9117
VDSR		666	612.6	33.66	0.9213	29.77	0.8314	28.82	0.7976	27.14	0.8279	32.01	0.9340
CARN		1592	118.9	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
IMDN		703	71.5	34.36	0.9270	30.32	0.8417	29.09	0.8046	28.17	0.8519	33.61	0.9445
HNCT		363	37.8	34.47	0.9275	30.44	0.8439	29.15	0.8067	28.28	0.8557	33.81	0.9459
FMEN		757	77.2	34.45	0.9275	30.40	0.8435	29.17	0.8063	28.33	0.8562	33.86	0.9462
FDIWN		645	51.5	34.52	0.9281	30.42	0.8438	29.14	0.8065	28.36	0.8567	-	-
BSRN		340	33.3	34.46	0.9277	30.47	0.8449	29.18	0.8068	28.39	0.8567	**34.05**	0.9471
RepRFN		392	38.4	34.45	0.9280	30.39	0.8430	29.13	0.8068	28.06	0.8494	33.76	0.9451
OSFFNet		524	37.8	**34.58**	**0.9287**	**30.48**	**0.8450**	**29.21**	**0.8080**	**28.49**	**0.8595**	34.00	**0.9472**
SRCNN	×4	8	52.7	30.48	0.8626	27.50	0.7513	26.90	0.7101	24.52	0.7221	27.58	0.8555
VDSR		666	612.6	31.35	0.8838	28.01	0.7674	27.29	0.7251	25.18	0.7524	28.83	0.8870
EDSR		1778	102.9	32.09	0.8938	28.58	0.7813	27.57	0.7357	26.04	0.7849	30.21	0.8336
CARN		1592	90.9	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
IMDN		715	40.9	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838	30.45	0.9075
HNCT		373	22.0	32.31	0.8957	28.71	0.7834	27.63	0.7381	26.20	0.7896	30.70	0.9112
FMEN		769	44.2	32.24	0.8955	28.70	0.7839	27.63	0.7379	26.28	0.7908	30.70	0.9107
FDIWN		664	28.4	32.23	0.8955	28.66	0.7829	27.62	0.7380	26.28	0.7919	-	-
BSRN		352	19.4	32.35	0.8966	28.73	0.7847	27.65	0.7387	26.27	0.7908	**30.84**	0.9123
RepRFN		402	22.1	32.28	0.8969	28.68	0.7836	27.65	0.7389	26.18	0.7858	30.79	0.9102
OSFFNet		537	22.0	**32.39**	**0.8976**	**28.75**	**0.7852**	**27.66**	**0.7393**	**26.36**	**0.7950**	**30.84**	**0.9125**

Table 1: Quantitative results of the state-of-the-art models on four benchmark datasets. The best result is marked with bold, and the second-best result is underlined.

Code	OISI	SFGC	MFDF	#Params	#Multi-Adds	Set5		Set14		BSD100		Urban100	
				[K]	[G]	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
000	✘	✘	✘	486	19.4	31.92	0.8908	28.45	0.7786	27.48	0.7336	25.76	0.7775
100	✔	✘	✘	488	19.5	32.11	0.8936	28.58	0.7818	27.57	0.7365	25.98	0.7846
010	✘	✔	✘	486	19.4	31.97	0.8914	28.47	0.7794	27.50	0.7342	25.80	0.7789
001	✘	✘	✔	535	21.9	31.95	0.8921	28.50	0.7806	27.52	0.7352	25.86	0.7818
110	✔	✔	✘	488	19.5	32.11	0.8937	28.58	0.7818	27.56	0.7365	25.98	0.7847
101	✔	✘	✔	536	22.0	**32.25**	**0.8958**	**28.65**	**0.7833**	**27.61**	**0.7379**	**26.12**	0.7890
011	✘	✔	✔	535	21.9	31.89	0.8912	28.46	0.7795	27.49	0.7343	25.80	0.7795
111	✔	✔	✔	537	22.0	**32.25**	**0.8958**	**28.65**	**0.7833**	**27.61**	**0.7379**	**26.12**	**0.7891**

Table 2: Different Configurations of OSFF.

Omni-Stage Feature Fusion architecture

Original Image Stacked Initialization The image stacking technique (Lee and Tai 2016) is commonly used in photography post-process for denoising and enhancement. It utilizes the randomness of noise by capturing a series of images without camera movement, resulting in slightly different noise patterns. Combining these images intelligently can remove noise without sacrificing details. In CNN-based super-resolution models, we can simulate this noise randomness by duplicating the original images and initializing convolutional kernels randomly.

We have designed an Original Image Stacked Initialization (OISI) method based on this phenomenon. As expressed in Equation 1, we stack multiple original images as input to the network and then use a 3×3 BSConv to extract shallow features. This initialization method of network input allows the shallow features to contain richer texture information, thus improving the model's ability to reconstruct images. BSRN's research (Li et al. 2022) has already validated the effectiveness of this approach.

Shallow Feature Global Connection Inspired by residual learning (He et al. 2016), we introduce the Shallow Feature Global Connection (SFGC) strategy, depicted in Figure 2. This strategy involves additive blending of shallow features with input features at the beginning of each feature extraction block. Before each addition operation, the shallow features are multiplied by a learnable weight parameter. This shared residual connection enables direct element-wise addition of shallow features with features at various levels, facilitating the transmission of low-frequency features and non-local information. Furthermore, these residual connections effectively address gradient vanishing and exploding issues, leading to faster convergence and more accurate model training. Although the SFGC may lead to an increase in computational and memory overhead for lightweight SISR networks, these additional costs are acceptable and do not significantly impact performance.

Multi-Receptive Field Dynamic Fusion To integrate features from various stages and leverage the complementary benefits of shallow and deep features, we introduce a Multi-Receptive Field Dynamic Fusion (MFDF) strategy. In this approach, as depicted in Figure 2, MFDF comprises multiple DRFs (matched with AEFDs). Lower-level features are processed using DRFs with smaller convolution kernels to better restore image texture and structures. Conversely, deeper-level features are processed using DRFs with larger convolution kernels to capture a broader range of contextual information and enhance overall structural recovery in the image. This process can be expressed as:

$$\begin{aligned} F_f^k &= H_{DRF}^k(F_k) \\ &= H_{ERB}^d(F_k) \times H_{CCA}(F_k), \end{aligned} \tag{6}$$

where H_{ERB}^d denotes an Efficient Residual Block (ERB) based on a $d \times d$ BSConv, with $d = \lfloor \frac{k}{2} \rfloor \times 2 + 1$. H_{CCA} denotes the Contrast-aware Channel Attention (CCA) block, which adaptively adjusts the importance of different channels in the feature map. Figure 3 illustrates how CCA enhances the expressive power of multi-level features by prioritizing crucial features in the deeper dimension.

Attention-Enhanced Feature Distillation

To further enhance the OSFFNet's feature extraction capability, we design an Attention-Enhanced Feature Distillation (AEFD) module based on the information distillation mechanisms (Hui et al. 2019). Firstly, we combine the refinement blocks based on ERB and the distillation blocks based on 1×1 convolution to perform feature distillation. The distilled features have half the number of channels compared to the input features. This process can be expressed as:

$$\begin{aligned} f_{bn} &= H_{BN}(f_{input}), \\ f_d^1, f_r^1 &= H_D^1(f_{bn}), H_R^1(f_{bn}), \\ f_d^2, f_r^2 &= H_D^2(f_r^1), H_R^2(f_r^1 + f_{bn}), \\ f_d^3, f_r^3 &= H_D^3(f_r^2), H_R^3(f_r^2 + f_{bn}), \\ f_d^4 &= H_D^4(F_R^3), \end{aligned} \tag{7}$$

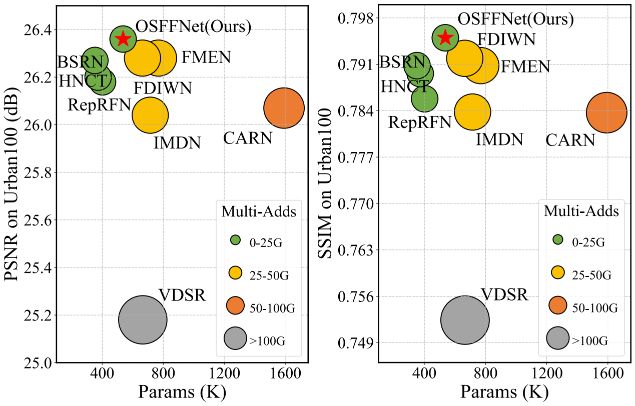


Figure 5: The performance and complexity comparisons on Urban100 for ×4 SR.

where H_{BN} represents denotes the batch normalization, H_D denotes an ERB based on a 3×3 BSConv, and H_R denotes a 1×1 standard convolution. Then, these distilled features are fused by:

$$f_{fused} = H_{fusion}(Concat_{i=1}^4 f_d^i), \tag{8}$$

where H_{fusion} denotes a 1×1 convolution layer used to adjust the number of channels. Next, we linearly combine the Enhanced Spatial Attention (ESA) (Liu et al. 2020) and CCA (Hui et al. 2019) blocks to achieve attention enhancement. The architecture of ESA is shown in Figure 4. In the following experiments, applying visual attention modules to super-resolution tasks has been proven to be highly effective. This process can be expressed as:

$$f_{att} = H_{CCA}(H_{ESA}(f_{fused})) + f_{bn}, \tag{9}$$

where H_{ESA} denotes the ESA module. During the final phase of AEFD, a feedforward network (FFN) (Liu et al. 2021) consisting of layer normalization (LN) and multi-layer perceptron (MLP) is used to enhance the features further. Thus the whole process of AEFD can be expressed as:

$$\begin{aligned} f_{output} &= H_{AEFD}(f_{input}) \\ &= H_{MLP}(H_{LN}(f_{att})) + f_{att}, \end{aligned} \tag{10}$$

where H_{MLP} denotes the MLP module, and H_{LN} denotes the operation of LN layer.

Experiments

Benchmarks

Datasets Drawing from previous works (Du et al. 2022; Li et al. 2022; Deng et al. 2023), we utilize two widely used training datasets DIV2K (Agustsson and Timofte 2017) and Flickr2K (Agustsson and Timofte 2017). Bicubic downsampling is used to generate corresponding LRs during training, and 64×64 patches are obtained through random cropping. Additionally, we introduce random flips and rotations (90°, 180°, 270°) to the images to enhance data diversity. For evaluation, we employ five common benchmark datasets: Set5 (Bevilacqua et al. 2012), Set14 (Zeyde, Elad, and Protter 2012), BSD100 (Martin et al. 2001), Urban100 (Huang, Singh, and Ahuja 2015), and Manga109 (Matsui et al. 2017).

Metrics We evaluate all the SR results by PSNR↑ and SSIM↑ metrics on the Y channel of the YCbCr color space. Furthermore, to evaluate the complexity of the model, we also measure the number of parameters (#Params) and the number of multiply-add operations (#Multi-Adds) performed by the model on RGB HRs of size 1280×720.

Implementation Details

Hyperparameters To determine a complete OSFFNet, we need to configure several hyperparameters, including the number of stacked original images (O), the number of channels in the hidden layers (C), the number of AEFD modules (N), and the kernel size in each DRF (D). After balancing network complexity and reconstruction performance, we propose an OSFFNet with $O = 8, C = 64, N = 8, Ds = (1, 1, 3, 3, 5, 5, 7, 7)$, respectively.

Optimizer & Scheduler For the optimization of our proposed OSFFNet, we utilize the Adam optimizer (Kingma and Ba 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1 \times 10^{-8}$ to minimize the mean absolute error (MAE) loss. As for the scheduler, the initial learning rate is set to 1×10^{-4} and decayed to the final value of 1×10^{-7} over 5×10^5 cosine annealing iterations. Lastly, the training is performed on an RTX3090 GPU with a mini-batch size of 64. Specifically, the models for ×3 and ×4 are the results of fine-tuning based on the entirety of training on ×2.

Comparisons with State-of-the-art Methods

Quantitative Comparison To demonstrate the effectiveness of the proposed OSFFNet, we conduct a quantitative objective comparison with some state-of-the-art (SOTA) lightweight SR methods, which include SRCNN (Dong et al. 2014), VDSR (Kim, Lee, and Lee 2016a), EDSR (Lim et al. 2017), CARN (Ahn, Kang, and Sohn 2018), IMDN (Hui et al. 2019), HNCT (Fang et al. 2022), FMEN (Du et al. 2022), FDIWN (Gao et al. 2022), BSRN (Li et al. 2022) and RepRFN (Deng et al. 2023). The quantitative comparison results are shown in Table 1, where our model achieves the best metrics on most datasets. The performance and complexity comparisons on Urban100 for ×4 SR are also shown in Figure 5. Based on these comparisons, we can see that OSFFNet achieves the best results across all datasets at the ×4 scale, highlighting its superior adaptability for high-scale ratio SR tasks. Furthermore, our model demonstrates notable improvements on the Urban100 dataset across all three scales, potentially due to OSFF's integration of multiple strategies to enhance shallow feature contribution, particularly suited for repetitive structures in urban scenes. However, OSFFNet falls short in achieving the best performance on the ×2 downscaled Set14 dataset, indicating a need for further improvement in local feature extraction.

Qualitative Comparison The qualitative comparisons of OSFFNet with other SOTA methods for ×4 SR are shown in Figure 6, featuring images from Set14, BSD100, Urban100, and Manga109. Notably, OSFFNet excels in accurately restoring book structure information in *barbara*, and produces clearer road textures in *78004* compared to other methods. It also demonstrates superior feature representation in *img011* and *Samayoeru*, resulting in finer HR images. For instance, in *img011*, only OSFFNet accurately rebuilds the texture of the small area with satisfactory visualization, consistent with quantitative results. Additionally, in *Samayoeru*, the female character depicted in the OSFFNet reconstruction results shows more pronounced eyebrows.

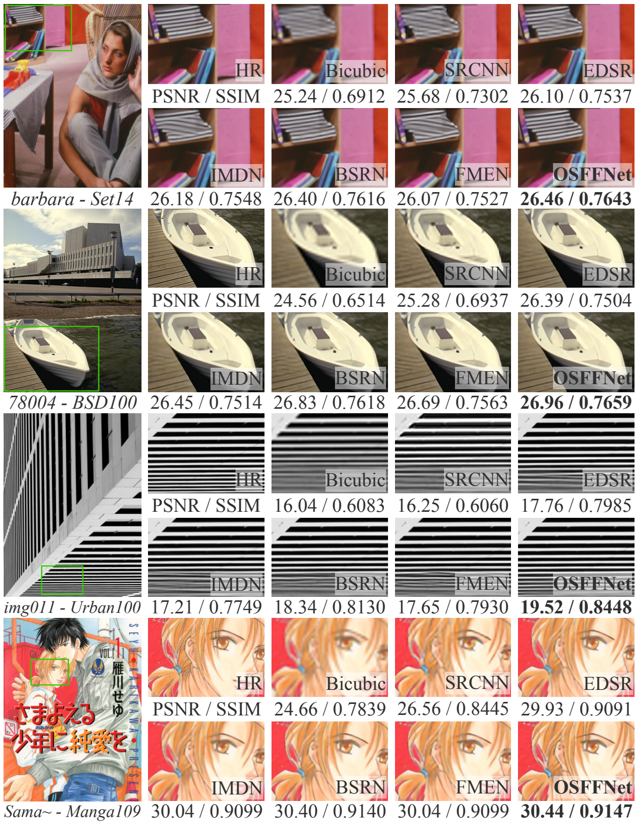


Figure 6: Qualitative comparison of OSFFNet with the state-of-the-art methods for ×4 SR.

Ablation Studies

Study on Configurations of OSFF To investigate the impact of different configurations of individual modules in OSFF on network performance, we have implemented several variants of OSFFNet, each named based on its code designation. Then, the experiments are performed at the ×4 scale, and all models are trained for only 1×10^5 iterations with the same training configuration. The training process is illustrated in Figure 7, with PSNR calculated on the Set5. The final results are reported in Table 2. Notably, significant improvements are achieved with OISI when using a single module. SFGC demonstrates good performance gains at minimal cost, making it the most cost-effective option. Our network achieves the best performance when combining MFDF with the other two modules.

Effectiveness of OSFF Architecture After configuring the OSFF architecture, we apply it to IMDN to evaluate its

Model	Scale	#Params [K]	#Multi-Adds [G]	Set5 PSNR	Set5 SSIM	Set14 PSNR	Set14 SSIM	BSD100 PSNR	BSD100 SSIM	Urban100 PSNR	Urban100 SSIM
w/o	×2	694	158.8	38.00	0.9605	**33.63**	0.9177	32.19	0.8996	32.17	0.9283
w/		739	168.8	**38.05**	**0.9606**	33.59	**0.9178**	**32.22**	**0.9000**	**32.20**	**0.9287**
w/o	×3	703	71.5	34.36	0.9270	**30.32**	**0.8417**	29.09	0.8046	28.17	0.8519
w/		748	75.8	**34.45**	**0.9277**	**30.32**	0.8415	**29.13**	**0.8056**	**28.19**	**0.8528**
w/o	×4	715	40.9	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838
w/o (channel-extended)		805	46.0	31.97	0.8917	28.46	0.7780	27.47	0.7325	25.71	0.7740
w/o (depth-extended)		823	47.2	**32.27**	0.8956	28.63	0.7821	**27.59**	0.7361	**26.11**	0.7856
w/		760	43.4	32.26	**0.8957**	**28.64**	**0.7822**	**27.59**	**0.7363**	**26.11**	**0.7862**

Table 3: IMDN without *vs.* with OSFF.

Model	ESA	CCA	#Params [K]	#Multi-Adds [G]	Set5 PSNR	Set5 SSIM	Set14 PSNR	Set14 SSIM	BSD100 PSNR	BSD100 SSIM	Urban100 PSNR	Urban100 SSIM
Net-*w/oAtt*	✘	✘	499	20.7	32.11	0.8941	28.57	0.7816	27.56	0.7357	25.96	0.7825
Net-*ESA*	✔	✘	532	22.0	32.30	0.8968	28.71	0.7845	27.65	0.7386	26.28	0.7925
Net-*CCA*	✘	✔	504	20.7	32.30	0.8963	28.66	0.7834	27.62	0.7377	26.17	0.7887
OSFFNet	✔	✔	537	22.0	**32.39**	**0.8976**	**28.75**	**0.7852**	**27.66**	**0.7393**	**26.36**	**0.7950**

Table 4: Evaluation results of Attention Mechanism.

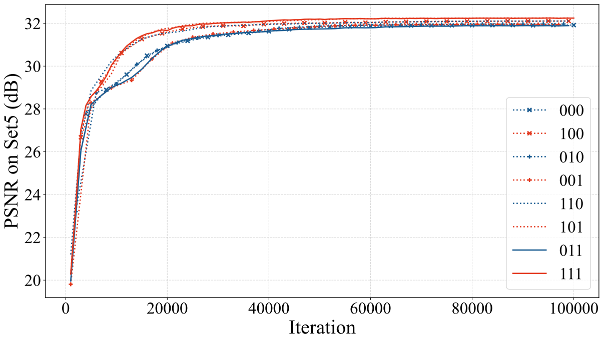
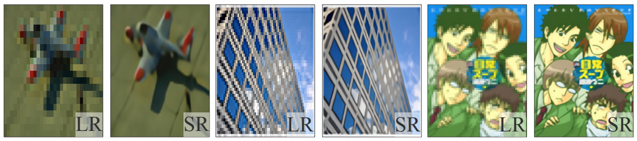


Figure 7: The training process of OSFFs.

effectiveness. The results of the ablation study are given in Table 3, demonstrating that integrating the OSFF architecture into IMDN significantly improves metrics, with only a slight increase in #Params and #Multi-Adds. To ensure fairness in parameter count and computational complexity, we extend the original IMDN in both channel and depth aspects and conduct experiments at the ×4 scale. As shown in Table 3, the depth-extended IMDN outperforms the OSFF-based IMDN, partially due to its higher #Params.

Evaluation of Attention Mechanism We conduct an ablation study on the attention mechanism for the OSFFNet to analyse the performance before and after applying the attention mechanism. By varying the presence of ESA and CCA, we obtain three variants: Net-*w/oAtt*, Net-*ESA*, and Net-*CCA*. These variants are evaluated on the benchmark dataset at the ×4 scale, and the results are summarized in Table 4. Notably, the model without the attention mechanism (#Params) exhibits a noticeable decline in performance. The inclusion of either ESA or CCA individually shows some performance improvement. Additionally, the best performance is achieved when ESA and CCA are combined as an enhanced attention mechanism, resulting in a 0.40dB PSNR increase on Urban100 compared to the Net-*w/oAtt* variant.



37073 in BSD100 *img035* in Urban100 *Nichijou~ in Manga109*

Figure 8: Visual results of OSFFNet for ×16 SR.

Performance at the high scale Figure 8 shows the remarkable super-resolution visual effect at ×6 scale. Despite the extreme degradation, with the LR image reduced to a very small resolution, our OSFFNet demonstrates a notably impressive super-resolution reconstruction ability. This achievement is attributed to the meticulously designed OSFF architecture, which safeguards crucial details in its underlying functionality, enabling OSFFNet to achieve outstanding performance even under severe degradation.

Conclusion

This paper focuses on the under-utilization of low-level features, a practical scenario rarely explored in previous lightweight SR studies. Consequently, we propose an Omni-Stage Feature Fusion Network (OSFFNet) to leverage the advantages of multi-level features, particularly shallow features. The OSFF architecture, featuring OISI, SFGC, and MFDF, seamlessly integrates features from different levels and incorporates an attention-enhanced feature distillation module to enhance model performance. The experimental results undoubtedly demonstrate the superiority of our model in lightweight SR reconstruction. Ablation experiments further confirm the effectiveness of the modules, offering valuable guidance for future research.

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, 126–135.

Ahn, N.; Kang, B.; and Sohn, K.-A. 2018. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 252–268.

Bevilacqua, M.; Roumy, A.; Guillemot, C.; and Alberi-Morel, M. L. 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, 135.1–135.10. BMVA Press.

Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; and Dong, C. 2023. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22367–22377.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11065–11074.

Deng, W.; Yuan, H.; Deng, L.; and Lu, Z. 2023. Reparameterized Residual Feature Network for Lightweight Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1712–1721.

Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; and Sun, J. 2021. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13733–13742.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014*, 184–199. Springer.

Dong, C.; Loy, C. C.; and Tang, X. 2016. Accelerating the super-resolution convolutional neural network. In *Computer Vision–ECCV 2016*, 391–407. Springer.

Du, Z.; Liu, D.; Liu, J.; Tang, J.; Wu, G.; and Fu, L. 2022. Fast and memory-efficient network towards efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 853–862.

Fang, J.; Lin, H.; Chen, X.; and Zeng, K. 2022. A hybrid network of cnn and transformer for lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1103–1112.

Gao, G.; Li, W.; Li, J.; Wu, F.; Lu, H.; and Yu, Y. 2022. Feature distillation interaction weighting network for lightweight image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 661–669.

Gendy, G.; He, G.; and Sabor, N. 2023. Lightweight image super-resolution based on deep learning: State-of-the-art and future directions. *Information Fusion*, 94: 284–310.

Haase, D.; and Amthor, M. 2020. Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14600–14609.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Huang, J.-B.; Singh, A.; and Ahuja, N. 2015. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5197–5206.

Hui, Z.; Gao, X.; Yang, Y.; and Wang, X. 2019. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2024–2032.

Hui, Z.; Wang, X.; and Gao, X. 2018. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 723–731.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016a. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1646–1654.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016b. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1637–1645.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 624–632.

Lee, M.; and Tai, Y.-W. 2016. Robust all-in-focus super-resolution for focal stack photography. *IEEE Transactions on Image Processing*, 25(4): 1887–1897.

Lepcha, D. C.; Goyal, B.; Dogra, A.; and Goyal, V. 2023. Image super-resolution: A comprehensive review, recent trends, challenges and applications. *Information Fusion*, 91: 230–260.

Li, J.; Fang, F.; Mei, K.; and Zhang, G. 2018. Multi-scale residual network for image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 517–532.

Li, Z.; Liu, Y.; Chen, X.; Cai, H.; Gu, J.; Qiao, Y.; and Dong, C. 2022. Blueprint separable residual network for efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 833–843.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, 136–144.

Liu, J.; Zhang, W.; Tang, Y.; Tang, J.; and Wu, G. 2020. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2359–2368.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. 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*, volume 2, 416–423. IEEE.

Matsui, Y.; Ito, K.; Aramaki, Y.; Fujimoto, A.; Ogawa, T.; Yamasaki, T.; and Aizawa, K. 2017. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76: 21811–21838.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. 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*, 1874–1883.

Tong, T.; Li, G.; Liu, X.; and Gao, Q. 2017. Image super-resolution using dense skip connections. In *Proceedings of the IEEE international conference on computer vision*, 4799–4807.

Wang, Z.; Chen, J.; and Hoi, S. C. 2020. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3365–3387.

Xia, B.; Hang, Y.; Tian, Y.; Yang, W.; Liao, Q.; and Zhou, J. 2022. Efficient non-local contrastive attention for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2759–2767.

Zeyde, R.; Elad, M.; and Protter, M. 2012. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, 711–730. Springer.

Zhang, X.; Zeng, H.; and Zhang, L. 2021. Edge-oriented convolution block for real-time super resolution on mobile devices. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4034–4043.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 286–301.

Zhao, H.; Kong, X.; He, J.; Qiao, Y.; and Dong, C. 2020. Efficient image super-resolution using pixel attention. In *Computer Vision–ECCV 2020 Workshops*, 56–72. Springer.