Expanded SPAN for Efficient Super-Resolution

Qing Wang Yan Wang Hongyu An Yi Liu Liou Zhang Shijie Zhao†

ByteDance

{wangqing.keen,wangya.my,anhongyu,liuyi.chester,zhangliou,zhaoshijie.0526}@bytedance.com

Abstract

This work proposes ESPAN, an efficient super-resolution (SR) network that extracts robust representations with constrained parameters by incorporating innovations from three perspectives: self-distillation and progressive learning (SDPL), general re-parameterization (GRep), and frequency-aware loss. In detail, SDPL shares partial blocks between the student and teacher models and progressively removes the tail convolutions of the student model, which contributes to a stable training process and reasonable convergence. Regarding GRep, we provide a more general schema of re-parameterization with interpretable theoretical derivation to achieve more flexible expansion of re-parameterization complexity. The frequency-aware loss utilizes the discrete cosine transform and a high-pass filter, enforcing the model to focus more on important high-frequency areas. The experimental results demonstrate the effectiveness of the proposed strategies. Overall, ESPAN exhibits better generality and robustness than previous top-ranking solutions in the NTIRE ESR challenge (e.g., 0.33 dB higher than SPAN on Manga109) while maintaining inference and restoration performance.

1. Introduction

Single image super-resolution is known to be an ill-posed inverse problem, and previous works [8, 9, 24, 37, 38, 42] have come up with many methods which can obtain high quality HR results under different LR conditions. These models typically exhibit a high number of parameters and substantial FLOPs. Efficient super-resolution aims to achieve favorable subjective performance with constrained parameters and FLOPs. Previous studies [5, 14, 29, 36, 39, 44] primarily focused on architectural design and re-parameterization, achieving a balanced trade-off between model size and subjective performance. We posit that the key lies in guiding parameter-constrained models to prioritize critical regions, such as high-frequency textures, and learn more robust features, thereby avoiding gradual overfitting to the training set.

Hence, in the process of participating in the NTIRE 2025 Efficient Super-Resolution Challenge, we propose a model named ESPAN which, we think, can learn more robust features from the data with limited parameters. The contributions of the paper are mainly the following three points:

- We employ self-distillation and progressive learning to enable model learning more robust features in its backbone.
- We provide general re-parameterization from a more interpretable and robust perspective.
- A frequency aware loss is proposed to make the model pay more attention to important high frequency areas, which can achieve a higher PSNR.

As shown in Tab. 1, the PSNR results on public test sets demonstrate that our model exhibits better generalization than other top-ranking architectures such as SPAN (NTIRE 2024 ESR Top 1) [33]. For example, on the Manga109 [32] and Urban100 [20] datasets, our method achieves 0.33 dB and 0.18 dB improvements over SPAN [33], while maintaining comparable runtime and PSNR performance on the NTIRE 2025 ESR challenge test set (i.e. LD-test).

2. Related Work

2.1. Efficient Image Super Resolution

To achieve real-time super-resolution (SR) applications, extensive innovative works have improved the efficiency (runtime, FLOPs, parameters) of SR models. SRCNN [13] pioneered the design of a convolutional neural network (CNN) to address the SR task. DRCN [25] introduced intense recursive layers to increase the receptive field and eliminated the need for new parameters from extra convolutions. DRRN [35] adopted a recursive convolutional network with parameter sharing, and CARN [2] integrated a cascading mechanism with the residual network. Unfortunately, the aforementioned recurrent blocks are computationally intensive. Thus, IDN [22] and IMDN [23] employed information distillation blocks to simplify the network structure. Following the information distillation network, RFDN [30]

†Corresponding author.

Table 1. Quantitative comparative results of our method and other comparable models on the public test set for ×4 SR are presented. Using the official provided code, the running time is calculated by averaging the results of three runs on the LD-test dataset on an A100 GPU.

Methods	Latency	LD-test [1]	Set14 [45]	BSD100 [31]	Manga109 [32]	Urban100 [20]	Test2k [16]
SPAN (NTIRE 2024 ESR Top1) [36]	6.93ms	27.01	26.59	26.10	27.97	24.17	26.07
R2Net (NTIRE 2024 ESR Top2) [33]	8.53ms	27.00	26.66	26.15	28.16	24.33	26.12
ESPAN	7.05ms	27.00	**26.69**	**26.16**	**28.30**	**24.35**	**26.15**

proposed the feature distillation connection (FDC). The implementation of FDC was equivalent to the information multi-distillation block but was more lightweight and flexible, boasting better SR performance. The above studies artfully optimized the intricate inter-layer connections to achieve superior reconstruction results with limited computational resources. Recently, RLFN [26] put forward a simple model with a shallow-feature-oriented extractor and improved training strategies. SPAN [36] utilized a novel parameter-free attention mechanism to focus on high contribution information and suppress redundant information, thereby attaining a SOTA runtime while maintaining good image quality.

2.2. Re-parameterization

Re-parameterization has proven effective in high-level vision tasks [4, 10–12]. Arora *et al.* [4] found that the re-parameterization of the fully connected layer could accelerate the training process in deeper networks. ACNet [10] employed the asymmetric convolution as a type of structural re-parameterization. RepVGG [12] parameterized a normal 3×3 convolution into parallel branches with identity mapping, 1×1 operation, and 3×3 convolution, achieving competitive performance. Diverse Branch Block [11] explored the different scale representations of CNNs. However, the success of re-parameterization in high-level tasks cannot be directly transferred to the field of SR. Building upon such experimental results, ECBSR [48] designed an edge-aware filter block without introducing additional computation during the inference stage. Recently, most methods [26, 33, 36, 40, 44] have already adopted re-parameterization as a universal basic approach in the field of efficient SR. R2Net [33] further proposed a re-parameterization module to improve 1×1 convolutions by expanding the intermediate channel, which exploited the representation capability of complex structures.

2.3. Information Distillation

Knowledge distillation [18] is a model compression framework with teacher-student networks. The small student network is trained to predict the output of a deep teacher network. As for the task of SR, KDSR [15] calculated and propagated the intermediate features from the teacher network to the student SR network, which benefits the reconstruction performance of the efficient student SR model.

Table 2. Quantitative comparisons are made between different kernel sizes of the first convolution. The depth and channels are set to 5 and 32, respectively. The latency is calculated on the LD-valid dataset [1, 28] using an A100 GPU.

Channel & Depth & Kernel	Latency	#Params	#MACs
`c32d5k3`	6.12ms	166.9K	10.89G
`c32d5k9`	6.19ms	173.8K	11.34G

FAKD [17] leveraged the correlation within a feature map to supervise the training of lightweight student networks. Inspired by previous methods, PISR [27] designed an imitation loss for training the teacher network to enable the student network to learn the distilled knowledge. Recently, several studies [26, 30] have adopted information distillation as a paradigm to achieve super-resolution (SR) with low computational and memory overhead. Following that, DIPNet [44] incorporated the re-parameterizable topology into the feature extraction blocks. SRN [40] added a loss function between the outputs of the teacher and student networks to optimize the learning process through explicit knowledge transfer.

3. Methods

Based on SPAN [36], we propose ESPAN. Through an evaluation of the depth-channel combinations in SPAN, we determine that setting the number of channels to 32 results in higher efficiency, and a depth of 6 is selected. Additionally, a 9×9 convolution is used to replace the conventional 3×3 convolution at the network's input stage. As shown in Tab. 2, the 9×9 convolution (denoted as `c32d5k9`) has a comparable latency to the 3×3 convolution (denoted as `c32d5k3`). This phenomenon may be associated with the efficient scheduling of the GPU and the fact that cuDNN may select different yet more efficient algorithms for different convolution kernel sizes. Considering that a larger kernel size can expand the model's receptive field, we choose to set the kernel size of the first convolution to 9. We aim to enable the model with limited parameters to learn more robust features from three aspects: self distillation and progressive learning, more generalized re-parameterization, and a loss function that focuses on important frequencies.

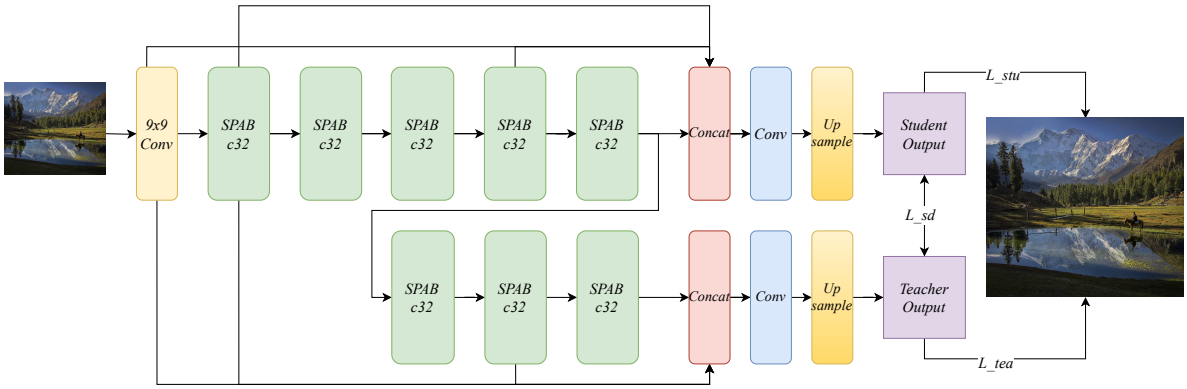

Figure 1. ESPAN with Self Distillation.

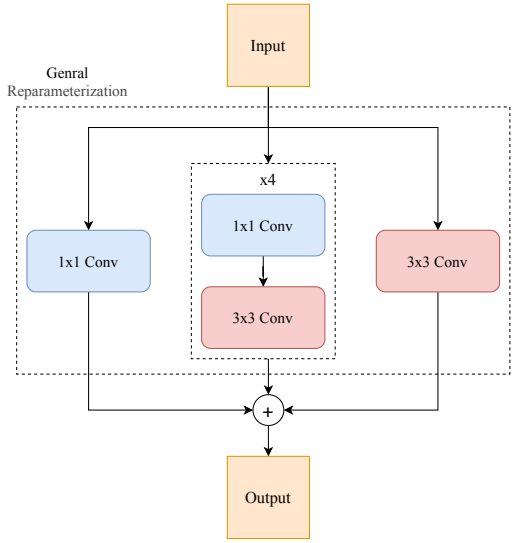

Figure 2. Illustration of the proposed general re-parameterization.

3.1. Self Distillation and Progressive Learning

Inspired by RIFE [21], self distillation is incorporated into our training pipeline. RIFE [21] demonstrates that certain intermediate information is more beneficial than extra privileged information in video frame interpolation (VFI) tasks. Similar investigations are conducted on super-resolution tasks. We believe that by sharing backbone components between the teacher and student models, self distillation can facilitate the learning of more robust features.

As shown in Fig. 1, the student model is based on SPAN with 32 channels and a depth of 5. The teacher model shares the identical backbone as the student model but includes 3 additional SPAB blocks appended to the student's backbone. A self distillation loss analogous to RIFE's formulation is adopted to co-train the teacher and student networks. Since the student model contains limited number of parameters, we aim to enable the student model to focus more on the areas that are not recovered as well as those by the teacher model through the use of a pixel-wise loss mask $\mathcal{M}$. Specifically, the pixel-wise loss mask $\mathcal{M}$ is defined as Eq. (1) to identify regions where the teacher's output outperforms the student's output,

$$\mathcal{M} = \mathbb{I}\left(|I_{stu} - I_{GT}|, |I_{tea} - I_{GT}|, 0.01\right) \tag{1}$$

$$\mathbb{I}\left(a, b, c\right) = \begin{cases} 1 & a - b > c \\ 0 & else \end{cases} \tag{2}$$

where I_{stu} and I_{tea} are outputs of student model and teacher model respectively, I_{GT} is ground-truth and $|\cdot|$ calculates absolute value. Then self distillation loss L_{sd} is calculated according to Eq. (3),

$$L_{sd} = ||(I_{tea} - I_{stu}) * \mathcal{M}||_2 \tag{3}$$

where $||\cdot||_2$ calculates mean square errors.

And the total loss L_{all} in the self-distillation stage is defined as Eq. (4),

$$\begin{aligned} L_{all} &= L_{stu} + L_{tea} + 0.001 * L_{sd} \\ L_{stu} &= ||I_{stu} - I_{GT}||_1 \\ L_{tea} &= ||I_{tea} - I_{GT}||_1 \end{aligned} \tag{4}$$

where L_{stu} is the L1 Loss between the output of the student model and the ground-truth, and L_{tea} is the L1 Loss between the output of the teacher model and the ground-truth.

After the self distillation phase, the student loss L_{stu} and the self distillation loss L_{sd} components are removed, and the entire teacher model is finetuned using the teacher loss L_{tea}. As shown in Fig. 3, leveraging the pretrained robust teacher, progressive learning is employed. The extra convolution layers or SPAB blocks are gradually removed from the teacher's backbone. We also attempted to load the self

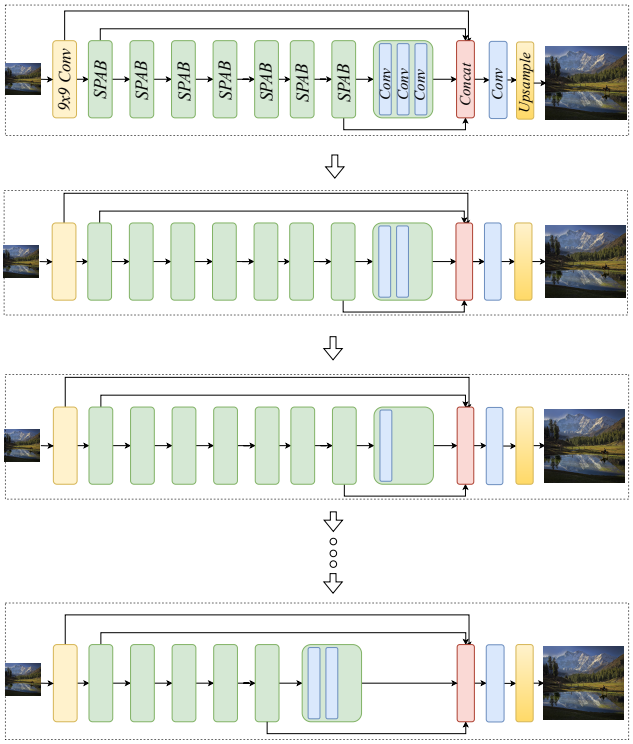

Figure 3. ESPAN with Progressive Learning.

distilled model and then increase the model depth one by one. However, we find that the performance is not as good as training a deeper model at the beginning and then reducing the model depth one by one. More details can be found in Sec. 4.3.1 and Sec. 4.3.2.

3.2. General Re-parameterization

To endow more representative capability to a single convolution, we employ reparamterization strategy [11, 12] that trains a complex module and infer it as convolution layer via equivalent transformation. Existing researches [14, 39, 41] have exploited varied re-parameterizable structures in SR tasks and achieved remarkable success. However, they are artificially designed with complicated topology, *e.g.*, residual in residual [46] block and mobilenet [19] block, making them hard to expand. Hence, we propose a general reparameterization (GRrep) block that simply parallelizes vanilla convolution, point-wise convolution, and multiple sequential convolutions as shown in Fig. 2. Given the input $\mathbf{x}$, 3×3 convolution's weights and bias $\{\mathbf{w}_1, b_1\}$, 1×1 convolution $\{w_2, b_2\}$, and several sequential convolutions (3×3-1×1 and 1×1-3×3) $\{\mathbf{w}_k w_k / w_k \mathbf{w}_k, b_k\}$, the output $\mathbf{y}$ of GRep $\{\mathbf{w}, b\}$ can be calculated by:

$$\begin{aligned}\mathbf{y} &= \overbrace{\mathbf{w}_1\mathbf{x}+b_1}^{3\times3\text{ Conv}} + \overbrace{w_2\mathbf{x}+b_2}^{1\times1\text{ Conv}} + \overbrace{\sum_{k\in\mathcal{K}}\mathbf{w}_k w_k\mathbf{x}+b_k}^{\text{SeqConv}} \\ &= \underbrace{\left[\mathbf{w}_1 + w_2 + \sum_{k\in\mathcal{K}}\mathbf{w}_k w_k\right]}_{\text{Rep weight}}\mathbf{x} + \underbrace{\left[b_1 + b_2 + \sum_{k\in\mathcal{K}} b_k\right]}_{\text{Rep bias}}.\end{aligned} \tag{5}$$

Inspired by [7], we further exhibit the weight updating process to show how the proposed GRep influencing training. Specifically, by chain rule, $\nabla_{\mathbf{w}^{(t)}} = \nabla_{\mathbf{w}_1^{(t)}} = \nabla_{w_2^{(t)}} = (w_k^{(t)})^{-1}\nabla_{\mathbf{w}_k^{(t)}}$, we can abstract GRep training procedure as using time varying momentum $\gamma^{(t)}$ for its complex sequential convolution and adaptive learning rate $\rho^{(t)}$:

$$\begin{aligned}\mathbf{w}^{(t+1)} &= \mathbf{w}_1^{(t+1)} + w_2^{(t+1)} + \sum_{k\in\mathcal{K}}\mathbf{w}_k^{(t+1)} w_k^{(t+1)} \\ &\leftarrow \mathbf{w}_1^{(t)} - \eta\nabla_{\mathbf{w}_1^{(t)}} + w_2^{(t)} - \eta\nabla_{w_2^{(t)}} \\ &\quad + \sum\left(\mathbf{w}_k^{(t)} - \eta\nabla_{\mathbf{w}_k^{(t)}}\right)\left(w_k^{(t)} - \eta\nabla_{w_k^{(t)}}\right) \\ &= \mathbf{w}^{(t)} - \eta\left(\nabla_{\mathbf{w}_1^{(t)}} + \nabla_{w_2^{(t)}} + \sum\mathbf{w}_k^{(t)}\nabla_{w_k^{(t)}}\right. \\ &\quad \left. + \sum w_k^{(t)}\nabla_{\mathbf{w}_k^{(t)}}\right) + \mathcal{O}(\eta^2) \\ &= \mathbf{w}^{(t)} - \eta\left(2 + \sum_{k\in\mathcal{K}}(w_k^{(t)})^2\right)\nabla_{\mathbf{w}^{(t)}} - \eta\sum_{k\in\mathcal{K}}\nabla_{w_k^{(t)}}\mathbf{w}_k^{(t)} \\ &= \mathbf{w}^{(t)} - \rho^{(t)}\nabla_{w_k^{(t)}} - \gamma_{\mathcal{K}}^{(t)}\mathbf{w}_{\mathcal{K}}^{(t)}.\end{aligned} \tag{6}$$

Since $(w_k^{(t)})^2 \geqslant 0$, we can observe the learning rate $\gamma^{(t)} = \eta(2 + \sum_{k\in\mathcal{K}}(w_k^{(t)})^2)$ is correspondingly increasing when using more sequential convolution in Eq. (6), enabling a faster convergence. Moreover, the optimization of each additional $\mathbf{w}_k$ will be calibrated with a time-related momentum $\gamma_k^{(t)} = \eta\nabla_{w_k^{(t)}}$ to guide better updating orientation and suppress oscillations. Compared to existing reparameterization methods, the proposed GRep is more interpretable and ensures the flexibility and robustness to expand or squeeze with varying complexities.

3.3. Frequency-Aware Loss Function

Frequency-domain analysis has already been widely considered in the field of super-resolution (SR) [3, 9]. To enhance the ESPAN network's capability to reconstruct high-frequency (HF) texture details, we incorporate a frequency-aware loss function into ESPAN. Specifically, we extract high-frequency information through a Gaussian Blurring operation. The residuals before and after filtering are used to simulate details that are difficult for ESPAN to recover.

Moreover, we constrain the aforementioned residuals during training. The HF loss is formulated as follows:

$$L_{HF} = \left\| (I_{HR} - \mathrm{B}(I_{HR})) - (I_{SR} - \mathrm{B}(I_{SR})) \right\|_1, \quad (7)$$

where I_{HR} and I_{SR} indicate the ground-truth high-resolution (HR) image and reconstructed SR image, $\mathrm{B}(\cdot)$ denotes a 5×5 kernel Gaussian Blur process. Furthermore, we compute the loss between HR and SR images in the Discrete Cosine Transform (DCT) frequency domain to ensure that different frequency components are treated equally, thereby effectively preserving high-frequency (HF) visual details. Compared with other frequency representation methods, e.g., Fast Fourier Transform (FFT), (Discrete Wavelet Transform) DWT, the energy of DCT is more concentrated and can better preserve the local characteristics of images. The results of subsequent ablation studies also demonstrate this point. Specifically, the DCT loss is formulated as follows:

$$L_{DCT} = \left\| \mathrm{DCT}(I_{HR}) - \mathrm{DCT}(I_{SR}) \right\|_1. \quad (8)$$

The overall frequency-aware loss function of the proposed ESPAN network is defined as:

$$L = L_2 + \alpha L_{HF} + \beta L_{DCT}, \quad (9)$$

where the weight parameters α and β are empirically preset to 1 and 0.001. L_{DCT} switches the loss calculation to the frequency domain, encouraging the model to better preserve different frequency components, while L_{HF} encourages the model to preserve high-frequency parts. Through the aforementioned loss function, ESPAN becomes more adept at capturing structural information for reconstructing finer details.

4. Experiments

4.1. Settings

4.1.1 Implementation Details

To achieve a better trade-off between running time and restoration quality, we determine the optimal number of blocks and the number of channels through grid searching. As shown in Tab. 4, a configuration with 32 channels is more cost-effective than those with 28 and 30 channels. This is because NVIDIA GPU data is generally aligned in multiples of 32. When the channel is set to 32, it can be better stored and read in an aligned manner. The GPU can retrieve data from the memory more efficiently, reducing the number of memory accesses and latency. Regarding model depth, slightly increasing the model depth increases the latency but effectively improves the PSNR. To this end, we develop an ESPAN with 32 channels and a depth of 5 or 6 for efficiency.

4.1.2 Datasets and Metrics

Following [29, 33, 41], we leverage a mixed dataset of DIV2K [1] and LSDIR [28] to train ESPAN, comprising a total of 85791 high-quality images. Low-resolution (LR) images are synthesized via bicubic interpolation. During the test phase, PSNR and SSIM [43] are employed to assess super-resolution performance on several widely used benchmark datasets, including LD-valid [33], Set5 [6], Set14 [45], B100 [31], Urban100 [20], Manga109 [32], Test2K [16], and Test4K [16].

4.1.3 Training Details for NTIRE 2025 ESR

I. At the first stage, we employ self distillation to train the teacher model.

- **Step 1**. We first train a 2× super-resolution model. HR patches of size 256×256 are randomly cropped from HR images, and the mini-batch size is set to 64. L1 loss and self distillation loss with AdamW optimizer are used and the initial learning rate is set to 0.0001 and halved at every 100k iterations. The total iterations is 500k. This step is repeated twice. And then we follow the same training setting and use 2× super-resolution model as pretrained model to train a 4× super-resolution model. This step is repeated twice.
- **Step 2**. HR patches of size 512×512 are randomly cropped from HR images, and the mini-batch size is set to 16. MSE loss, frequency-aware loss and self distillation loss with AdamW optimizer are used and the initial learning rate is set to 0.0001 and halved at every 100k iterations. The total iterations is 500k. This step is also repeated twice.
- **Step 3**. We only train the teacher model. HR patches of size 512×512 are randomly cropped from HR images, and the mini-batch size is set to 16. MSE loss and frequency-aware loss with AdamW optimizer are used and the initial learning rate is set to 0.00005 and halved at every 100k iterations. The total iterations is 500k. This step is also repeated twice.

II. At the second stage, progressive learning is applied to derive the final student model.

- **Step 4**. We drop the additional convolution layer or SPAB block gradually. In the procedure, HR patches with size 512×512 are randomly cropped from HR images, and the mini-batch size is set to 16. L1 loss with AdamW optimizer are used and the initial learning rate is initialized as 0.0001 and halved at every 100k iterations. The total iterations is 500k.
- **Step 5**. We repeat the following training process many times until convergence. HR patches of size 512×512 are randomly cropped from HR images, and the mini-batch size is set to 16. MSE loss and frequency-aware loss with

Table 3. Quantitative comparison (average PSNR/SSIM on RGB, Parameters, MACs, Latency, Memory Footprint, and Activations) with state-of-the-art approaches for efficient image SR (×4). Overall best results are in **bold**. MACs are measured under the setting of the input image to 256×256. Latency and Memory footprint are reported on the DIV2K-valid [1] dataset with Tesla V100.

Methodss	Latency (ms)	Para (K)	MACs (G)	Mem (M)	Acts (M)	DIV2KV [1] PSNR/SSIM	Test2K [16] PSNR/SSIM	Test4K [16] PSNR/SSIM
PCEVA [49]	20.9	402	24.73	372.5	72.09	28.68/0.8104	26.01/0.7511	27.41/0.7981
EFDN [39]	20.4	276	16.73	710.0	111.12	29.00/0.8187	26.17/0.7580	27.63/0.8049
ECBSR [47]	19.7	622	40.66	231.6	77.59	28.86/0.8190	26.13/0.7564	27.57/0.8033
FMEN [14]	18.2	341	22.28	**205.9**	72.09	29.00/0.8190	26.17/0.7586	27.62/0.8051
QuickSRNet [5]	15.3	436	28.51	296.5	56.62	28.78/0.8140	26.08/0.7552	27.49/0.8019
PFDN [29]	16.3	272	16.76	344.3	65.10	28.95/0.8176	26.17/0.7578	27.63/0.8047
DIPNet [44]	16.0	243	14.90	550.4	72.97	29.04/0.8184	26.11/0.7550	27.52/0.8018
R2Net [33]	14.3	215	13.04	306.7	52.52	28.91/0.8162	26.12/0.7561	27.57/0.8029
SPAN [36]	14.2	151	**9.83**	705.2	**41.68**	28.87/0.8143	26.07/0.7539	27.48/0.8005
ESPAN	**14.0**	192	12.56	1795	48.08	28.89/0.8159	26.15/0.7569	27.58/0.8033

Table 4. Quantitative comparisons are made between different block configurations and depth numbers on the LD-valid dataset [1, 28]. For fairness, all models are retrained under the same settings. The latency is calculated using an A6000 GPU.

Channel & Depth	Latency	#Params	#MACs	LD-valid
`c28d6` [36]	15.25ms	150.7K	9.84G	26.72
`c24d6`	14.52ms	112.4K	7.33G	26.66
`c24d7`	15.41ms	128.0K	8.00G	26.69
`c22d7`	14.95ms	108.4K	6.78G	26.65
`c22d8`	15.57ms	121.6K	7.64G	26.68
`c28d5`	15.21ms	129.4K	8.45G	26.68
`c30d5`	15.26ms	147.6K	9.63G	26.71
`c32d5`	14.56ms	166.9K	10.89G	26.73

AdamW optimizer are used and the initial learning rate is set to 0.00005 and halved at every 100k iterations. The total iterations is 500k.

Quantitative comparison of the ESPAN and the baseline of NTIRE 2025 ESR Challenge is shown as Tab. 5.

4.2. Comparative Results

4.2.1 Quantitative Comparison

Following [41], we compare the ESPAN with more efficient SR models in Tab. 3, including PCEVA [49], EFDN [39], ECBSR [48], FMEN [14], QuickSRNet [5], PFDN [29], DIPNet [44], R2Net [33], and SPAN [36]. In particular, we utilize DIV2K [1], Test2K [16], and Test4K [16] for evaluation to avoid the over-fitting on the dataset of the challenge. Generally, our ESPAN performs a better robustness than recent NTIRE solutions. Specifically, compared to EFDN [39], the SPAN drops 0.1dB on Test2K while the proposed ESPAN slightly degrades with 0.02dB.

4.2.2 Qualitative Comparison

We present a visual comparison of ESPAN with SPAN [33] and R2Net [33] in Fig. 4. By learning more robust features and focusing on critical high-frequency regions, our method produces fewer artifacts and reconstructs periodic textures closer to the ground truth (first and second rows of Fig. 4). Additionally, outlines in text and facial regions (last three rows of Fig. 4) are sharper and more defined.

4.3. Ablation Studies

We conduct extensive ablation studies to analyze the contributions of individual components in proposed methods.

4.3.1 Effect of Self Distillation

To demonstrate the effect of self distillation, we compare the training settings with and without self distillation. As shown in Fig. 5, the training process with self distillation converges faster than that without self distillation and achieves a higher PSNR on the validation set. The teacher model with more parameters can learn better super-resolution results from the ground truth. However, simply using the teacher model's output to guide the student model's output is difficult. Sharing some parts of the backbone between the teacher model and the student model enables the student model to learn knowledge more easily since it has a similar feature representation to the teacher model. And this is what we consider to be robust features for both the student model and the teacher model.

4.3.2 Effect of Progressive Learning

We leverage checkpoints during training to demonstrate the effectiveness of the proposed progressive learning. As shown in Fig. 1, after self-distillation training, a teacher model and a student model sharing the first 5 SPAB blocks

Table 5. Comparisons between ESPAN and the baseline of the NTIRE 2025 ESR on the LD-valid/test [1, 28]. The data is sourced from official report [34].

Model	Runtime[ms]	#Params[M]	FLOPs[G]	LD-valid PSNR	LD-test PSNR
Baseline	22.18	0.27	16.70	26.93	27.01
ESPAN	9.51	0.19	12.56	26.90	27.00

Table 6. Comparision of different training settings for progressive learning.

Training Setting	Validation PSNR
`d5d6d6`	26.847
`d5d7d6`	26.850

LR SPAN [33] R2Net [33] **ESPAN** (Ours) HR

Figure 4. Visual comparison of ESPAN with other top ranking methods of the NTIRE 2024 ESR Challenge for the $\times 4$ super-resolution (SR) task on the public datasets.

are obtained. Two training settings are considered. One is loading the shared 5 blocks, increasing the depth to 6, and training for additional iterations (denoted as `d5d6d6`). The other is loading the shared 5 blocks, increasing the depth to 7, and then decreasing the depth to 6 (denoted as `d5d7d6`). As indicated in Tab. 6, `d5d7d6` slightly outperforms `d5d6d6`, suggesting that initializing with a deeper model and gradually condensing it into a smaller architecture may yield better results by retaining more informative features.

4.3.3 Effect of General Re-parameterization

To evaluate the effectiveness of the proposed general reparameterization (GRep), we compare it with existing reparameterization approaches (RepVGG [12], RRRB [14], and RepMBConv [41]) by applying them to the proposed ESPAN under the same training settings. As shown in Tab. 7, even the simplest GRep [1,0] surpasses other reparameterizable blocks. Moreover, the proposed GRep achieves faster convergence and higher PSNR with an increased number of SeqConvs. Specifically, using 4 SeqConvs, the PSNR of the model reaches 26.67 dB within 300k iterations, out-

Table 7. Ablation study of proposed general re-paramterization.

Methods	3×3	1×1	Identity	SeqConv	300k/500k
Vanilla Conv	✓				26.355/26.484
RepVGG [12]	✓	✓	✓		26.563/26.618
EDBB [39]	✓	✓	✓	*5 branches*	26.658/26.689
RRRB [14]	*Residual in Residual Block*				26.620/26.677
RepMBC [41]	*MobileNetv3 Block*				26.605/26.663
GRep	✓	✓		[1,0]	26.633/26.690
	✓	✓		[2,0]	26.672/26.707
	✓	✓		[4,0]	**26.674**/26.705
	✓	✓		[2,2]	26.669/**26.709**
	✓	✓		[0,4]	26.672/26.700

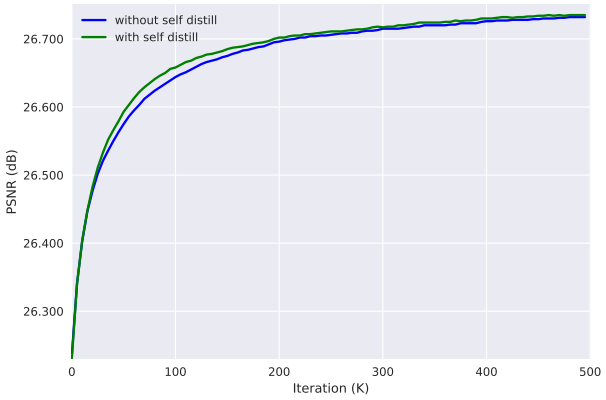

Figure 5. Validation PSNR during the training process.

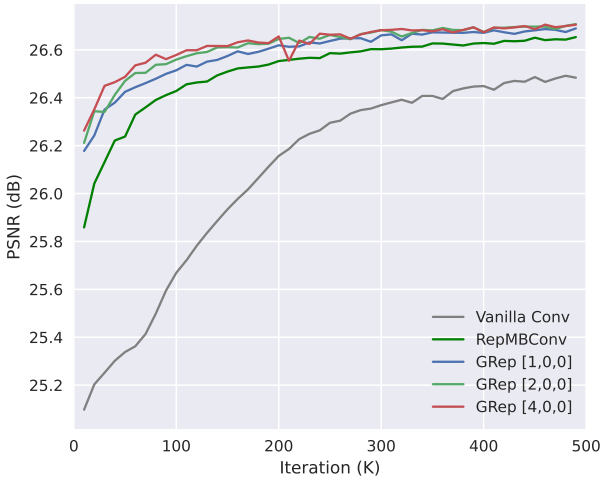

Figure 6. Validation PSNR with varied repconv during training.

performing vanilla convolution by 0.31 dB or saving 40% training time compared to RRRB. We also compared varied SeqConv ([a,b] represents a 1×1-3×3 and b 3×3-1×1). The GRep with [4,0] achieves the highest PSNR in 300k training iterations, while GRep with [2,2] negligibly outperforms other permutations. In Fig. 6, we visualize the PSNR curve of training process for more comprehensive comparison, where the proposed GRep surpasses RepMBConv and improves with the expanded number of sequential convolution. Overall, GRep provides approximately 0.32 dB improvement over vanilla convolution.

Table 8. Ablation study of the frequency-aware loss.

Methods	PSNR	Methods	PSNR
L1 Loss	26.609	HF Loss	16.740
L2 Loss	26.592	DCT Loss	26.570
L1 + HF Loss	26.622	L1 + FFT Loss	26.598
L1 + DCT Loss	26.636	L1 + DWT Loss	26.619
L1 + HF + DCT Loss	26.641	L2 + HF + DCT Loss	**26.658**

4.3.4 Effect of Frequency-Aware Loss Function

To evaluate the effectiveness of different combinations of the proposed high frequency (HF) loss and discrete cosine transform (DCT) loss fairly, a series of ablation studies are conducted. Specifically, we conduct × 4 super-resolution (SR) experiments on the validation dataset of the NTIRE 2025 ESR challenge. Each ablation branch has the same network architecture and training strategy and only the loss function is changed. As can be seen from Table 8, directly replacing the L1 or L2 loss with a frequency-domain loss will result in a degradation of performance. Especially, the HF loss leads to the failure of training and we depict intermediate results. Introducing the HF loss and DCT loss improves the baseline by 0.013 dB and 0.027 dB, respectively. For other frequency-domain transforms, the DCT loss performs better than the FFT and DWT loss as it processes in a block-wise manner that better preserves local correlation. Notably, when directly replacing the L1 loss with the L2 loss, the reconstruction quality decreases. Overall, the combination of the L2 loss and the frequency-aware loss achieves the best restoration quality. These results indicate that the proposed frequency-aware loss can effectively enhance the ESR reconstruction performance by preserving more structural high-frequency details.

5. Conclusion

In this paper, we propose ESPAN, which learns more robust features from general reparameterization, frequency-aware loss, self-distillation, and progressive learning. Quantitative results show a better convergence curve during training, competitive PSNR on the test set of the challenge, and better generalization capabilities on other public test sets. We believe ESPAN demonstrates that studying loss functions, reparameterization, and knowledge distillation can also be very helpful in exploring the potential of efficient super-resolution compared to studying model structure. We hope the proposed ESPAN can enlighten more research in efficient SR in the future.

References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, pages 1122–1131, 2017. 2, 5, 6, 7

[2] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, pages 256–272, 2018. 1

[3] Hongyu An, Xinfeng Zhang, Shijie Zhao, and Li Zhang. Fato: Frequency attention transformer for omnidirectional image super-resolution. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–7, 2024. 4

[4] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International conference on machine learning*, pages 244–253. PMLR, 2018. 2

[5] Guillaume Berger, Manik Dhingra, Antoine Mercier, Yashesh Savani, Sunny Panchal, and Fatih Porikli. Quicksrnet: Plain single-image super-resolution architecture for faster inference on mobile platforms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2187–2196, 2023. 1, 6

[6] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, pages 1–10, 2012. 5

[7] Kartikeya Bhardwaj, Milos Milosavljevic, Liam O'Neil, Dibakar Gope, Ramon Matas Navarro, Alex Chalfin, Naveen Suda, Lingchuan Meng, and Danny Loh. Collapsible linear blocks for super-efficient super resolution. In *Proceedings of Machine Learning and Systems*, Santa Clara, USA, 2022. 4

[8] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, pages 3086–3095, 2019. 1

[9] Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In *European Conference on Computer Vision*, pages 669–687. Springer, 2022. 1, 4

[10] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks. In *ICCV*, pages 1911–1920, Seoul, Korea (South), 2019. IEEE. 2

[11] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Diverse branch block: Building a convolution as an inception-like unit. In *CVPR*, pages 10886–10895, virtual, 2021. Computer Vision Foundation / IEEE. 2, 4

[12] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *CVPR*, pages 13733–13742, virtual, 2021. Computer Vision Foundation / IEEE. 2, 4, 7, 8

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 38(2):295–307, 2016. 1

[14] Zongcai Du, Ding Liu, Jie Liu, Jie Tang, Gangshan Wu, and Lean Fu. Fast and memory-efficient network towards efficient image super-resolution. In *CVPRW*, pages 853–862, 2022. 1, 4, 6, 7, 8

[15] Qinquan Gao, Yan Zhao, Gen Li, and Tong Tong. Image super-resolution using knowledge distillation. In *Asian Conference on Computer Vision*, pages 527–541. Springer, 2018. 2

[16] Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. In *ICCVW*, pages 3512–3516, 2019. 2, 5, 6

[17] Zibin He, Tao Dai, Jian Lu, Yong Jiang, and Shu-Tao Xia. Fakd: Feature-affinity based knowledge distillation for efficient image super-resolution. In *2020 IEEE international conference on image processing (ICIP)*, pages 518–522. IEEE, 2020. 2

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[19] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 4

[20] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015. 1, 2, 5

[21] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[22] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *CVPR*, pages 723–731, Salt Lake City, USA, 2018. Computer Vision Foundation / IEEE Computer Society. 1

[23] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACM MM*, pages 2024–2032, 2019. 1

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1

[25] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, pages 1637–1645, 2016. 1

[26] Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *CVPRW*, pages 766–776, 2022. 2

[27] Wonkyung Lee, Junghyup Lee, Dohyung Kim, and Bumsub Ham. Learning with privileged information for efficient image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 465–482. Springer, 2020. 2

[28] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Deman-

dolx, et al. Lsdir: A large scale dataset for image restoration. In *CVPRW*, pages 1775–1787, 2023. 2, 5, 6, 7

[29] Yawei Li, Yulun Zhang, Radu Timofte, Luc Van Gool, Lei Yu, Youwei Li, Xinpeng Li, Ting Jiang, Qi Wu, Mingyan Han, et al. Ntire 2023 challenge on efficient super-resolution: Methods and results. In *CVPRW*, pages 1922–1960, 2023. 1, 5, 6

[30] Jie Liu, Jie Tang, and Gangshan Wu. Residual feature distillation network for lightweight image super-resolution. In *ECCVW*, pages 41–55, 2020. 1, 2

[31] David R. Martin, Charless C. Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, pages 416–425, 2001. 2, 5

[32] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multim. Tools Appl.*, 76(20):21811–21838, 2017. 1, 2, 5

[33] Bin Ren and et al. The ninth ntire 2024 efficient super-resolution challenge report, 2024. 1, 2, 5, 6, 7

[34] Bin Ren and et al. The tenth ntire 2025 efficient super-resolution challenge report, 2025. 7

[35] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, pages 2790–2798, 2017. 1

[36] Cheng Wan, Hongyuan Yu, Zhiqi Li, Yihang Chen, Yajun Zou, Yuqing Liu, Xuanwu Yin, and Kunlong Zuo. Swift parameter-free attention network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6246–6256, 2024. 1, 2, 6

[37] Boyang Wang, Fengyu Yang, Xihang Yu, Chao Zhang, and Hanbin Zhao. Apisr: anime production inspired real-world anime super-resolution. In *CVPR*, pages 25574–25584, 2024. 1

[38] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*, pages 0–0, 2018. 1

[39] Yan Wang. Edge-enhanced feature distillation network for efficient super-resolution. In *CVPRW*, pages 777–785, 2022. 1, 4, 6, 8

[40] Yucong Wang and Minjie Cai. A single residual network with esa modules and distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1971–1981, 2023. 2

[41] Yan Wang, Yusen Li, Gang Wang, and Xiaoguang Liu. Plainusr: Chasing faster convnet for efficient super-resolution. In *Proceedings of the Asian Conference on Computer Vision*, pages 4262–4279, 2024. 4, 5, 6, 7, 8

[42] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *CVPR*, pages 25796–25805, 2024. 1

[43] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 5

[44] Lei Yu, Xinpeng Li, Youwei Li, Ting Jiang, Qi Wu, Haoqiang Fan, and Shuaicheng Liu. Dipnet: Efficiency distillation and iterative pruning for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1701, 2023. 1, 2, 6

[45] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces - 7th International Conference*, pages 711–730, 2010. 2, 5

[46] Ke Zhang, Miao Sun, Tony X Han, Xingfang Yuan, Liru Guo, and Tao Liu. Residual networks of residual networks: Multilevel residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(6):1303–1314, 2017. 4

[47] Xindong Zhang, Hui Zeng, and Lei Zhang. Edge-oriented convolution block for real-time super resolution on mobile devices. In *ACM MM*, pages 4034–4043, Virtual Event, China, 2021. ACM. 6

[48] Xindong Zhang, Hui Zeng, and Lei Zhang. Edge-oriented convolution block for real-time super resolution on mobile devices. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4034–4043, 2021. 2, 6

[49] Zhou Zhou, Jiahao Chao, Jiali Gong, Hongfan Gao, Zhenbing Zeng, and Zhengfeng Yang. Enhancing real-time super resolution with partial convolution and efficient variance attention. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5348–5357, 2023. 6