See More Details: Efficient Image Super-Resolution by Experts Mining

Eduard Zamfir [1] **Zongwei Wu** [1 *] **Nancy Mehta** [1] **Yulun Zhang** [2 3 *] **Radu Timofte** [1]

Abstract

Reconstructing high-resolution (HR) images from low-resolution (LR) inputs poses a significant challenge in image super-resolution (SR). While recent approaches have demonstrated the efficacy of intricate operations customized for various objectives, the straightforward stacking of these disparate operations can result in a substantial computational burden, hampering their practical utility. In response, we introduce **S**eemo**R**e, an efficient SR model employing expert mining. Our approach strategically incorporates experts at different levels, adopting a collaborative methodology. At the macro scale, our experts address rank-wise and spatial-wise informative features, providing a holistic understanding. Subsequently, the model delves into the subtleties of rank choice by leveraging a mixture of low-rank experts. By tapping into experts specialized in distinct key factors crucial for accurate SR, our model excels in uncovering intricate intra-feature details. This collaborative approach is reminiscent of the concept of "see more", allowing our model to achieve an optimal performance with minimal computational costs in efficient settings. The source codes will be publicly made available at `https://github.com/eduardzamfir/seemoredetails`

1. Introduction

Single image super-resolution (SR) is a long-standing low-level vision endeavour that pursues the reconstruction of a high-resolution (HR) image from its degraded low-resolution (LR) counterpart. This challenging task has garnered considerable attention owing to the expeditious development of ultra-high definition devices and video streaming applications (Khani et al., 2021; Zhang et al., 2021a). Foreseeing the resource constraints, it is of substantial desire to design an efficient SR model for gauging the HR images to be perfectly visualized on these devices or platforms. Identifying the most plausible candidates for missing HR pixels poses a particular challenge for SR. In the absence of external priors, the primary approaches for SR involves exploring the intricate relationships among the neighboring pixels for reconstruction. Recent SR models exemplify this through methods such as **(a)** attention (Liang et al., 2021; Zhou et al., 2023; Chen et al., 2023), **(b)** feature mixing (Hou et al., 2022; Sun et al., 2023), and **(c)** global-local context modeling (Wang et al., 2023; Sun et al., 2022), yielding remarkable accuracy.

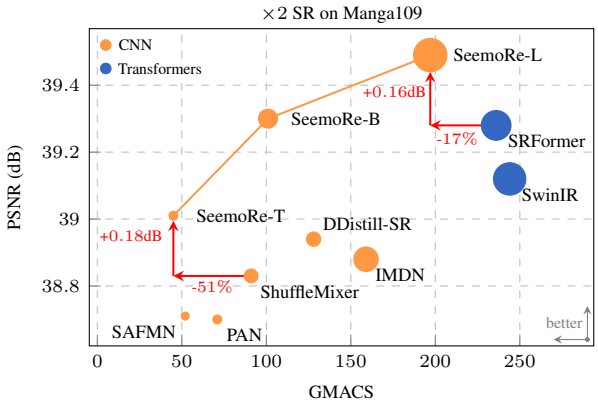

Figure 1. Model complexity trade-off. Visualization of PSNR, GMACS, and parameter counts on Manga109 dataset for ×2 task. Our proposed SeemoRe excels the state-of-the-art CNN-based and *lightweight* Transformer-based SR models. Marker size indicates parameter counts w.r.t SwinIR-Light (Liu et al., 2021).

Unlike other approaches in this work, we aim to avoid complex and disconnected blocks focusing on specific factors, opting instead for a unified learning module specialized for all aspects. However, an additional challenge arises due to the efficiency requirement, rendering implicit learning through a vast number of parameters unfeasible, especially in the context of devices with limited resources.

To achieve such an efficient unification, we introduce **S**eemo**R**e, which leverages the synergy of different experts to maximize intra-feature intertwining, collaboratively learn-

[1]Computer Vision Lab, CAIDAS & IFI, University of Würzburg, Germany [2]AI Institute, Shanghai Jiao Tong University, China [3]Computer Vision Lab, ETH Zurich, Switzerland. Correspondence to: Zongwei Wu <zongwei.wu@uni-wuerzburg.de>, Yulun Zhang <yulun100@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

ing a cohesive relation across LR pixels. Our motivation stems from the observation that image features often display diverse patterns and structures. Attempting to capture and model all these patterns with a single, monolithic model can be challenging. Collaborative experts, on the other hand, enable the network to specialize in different regions or aspects of the input space, enhancing its adaptability to various patterns and facilitating the modeling of LR-HR dependencies, akin to "See More".

Technically, our network is composed of stacked residual groups (RGs) for dynamically selecting the pivotal features via experts, focusing on two different aspects. At the macro level, each RG embodies two successive expert blocks: **(a)** *Rank modulating expert* (RME), expertized in dealing with the most informative features through low-rank modulation, and **(b)** *Spatial modulating expert* (SME), expertized in efficient spatial enhancement. At the micro level, we devise a Mixture of Low-Rank Expertise (MoRE) as the foundational component within RME to dynamically select the best and most suitable rank for different inputs and at different network depths while implicitly modeling the global contextual relationships. Furthermore, we design a Spatial Enhancement Expertise (SEE) as an efficient alternative to complex self-attention within SME for distinctly improving the spatial-wise local aggregation capabilities. Such a combination efficiently modulates the mutual dependencies within the feature attributes, enabling our model to extract high-level information, which is a key aspect of SR. By explicitly mining experts at different granularity for different expertise, our network navigates the intricacies between spatial and channel features, maximizing their synergistic contribution and thus accurately and efficiently reconstructing more details.

As shown in Figure 1, our network significantly outperforms the state-of-the-art (SOTA) efficient models such as DDistill-SR (Wang et al., 2022) or SAFMN (Sun et al., 2023) by a considerable margin, while utilizing only half or even less of the GMACS. Although our model is specifically designed for efficient SR, its scalability is evident as our larger model surpasses the SOTA lightweight transformer in performance while incurring lower computational costs. Overall, our key contributions are threefold:

- We propose SeemoRe which matches the versatility of Transformer-based methods and the efficiency of CNN-based methods.
- A Rank modulating expert (RME) is proposed to probe into the intricate inter-dependencies among the relevant feature projections in an efficient manner.
- A Spatial modulating expert (SME) is proposed to integrate the complementary features extracted by SME by encoding the local contextual information.

2. Related Works

CNN-based SR. In recent years, CNN-based techniques have outperformed traditional interpolation algorithms (Duchon, 1979) by learning a non-linear mapping between the input and target in an end-to-end training manner. The seminal SRCNN (Dong et al., 2014) introduced a three-layer convolutional approach for image super-resolution, later extended by works such as (Lim et al., 2017; Zhang et al., 2018b; Hui et al., 2019; Liang et al., 2021). VDSR (Kim et al., 2016) and EDSR (Lim et al., 2017) deepen networks using residual learning principles, with EDSR streamlining residual blocks for deeper training. Conversely, RCAN (Zhang et al., 2018a) introduces a novel residual-in-residual architecture for models exceeding 400 layers. While various spatial and channel attention mechanisms aim to enhance image reconstruction quality, CNN-based techniques still struggle to effectively utilize shared information across both dimensions. In this work, we aim to explore the interdependencies among the features in a computationally efficient way.

Transformer-based SR. Thanks to its remarkable performance in high-level tasks (Dosovitskiy et al., 2021), the Transformer architecture has found its way into low-level vision tasks, such as image SR. Contemporary Transformer-based approaches aim to alleviate the computational load by confining self-attention to local regions and incorporating a higher degree of locality bias into their network design. SwinIR (Liang et al., 2021) incorporates local window self-attention and a shift mechanism inspired by the Swin Transformer design(Liu et al., 2021). Meanwhile, others like ELAN (Zhang et al., 2022) or ESRT (Lu et al., 2022) reduce the feature dimensions by splitting or down-scaling to enhance the computational efficiency. Omni-SR (Wang et al., 2023) models pixel-interactions across different axes, creating universal correlations. SRFormer (Zhou et al., 2023) optimizes the computational efficiency by employing large window self-attention through the permutation of self-attention mechanisms. However, transformer-based methods typically demand significantly higher computational resources, even with smaller model capacities.

Efficiency in SR. In recent years, the pursuit of efficient SR techniques has gained significant momentum (Li et al., 2022; Ignatov et al., 2023; Li et al., 2023; Conde et al., 2023). Consequently, researchers have introduced streamlined neural architectures (Ignatov et al., 2021), network compression (Wang et al., 2022), reparameterization (Zhang et al., 2021b), and other training strategies to cater to the demand for efficiency. Initially, efficient SR methods utilized group convolutions and cascaded block designs to boost efficiency (Ahn et al., 2018; Hui et al., 2019). Subsequent advancements introduced convolution-based spatial or chan-

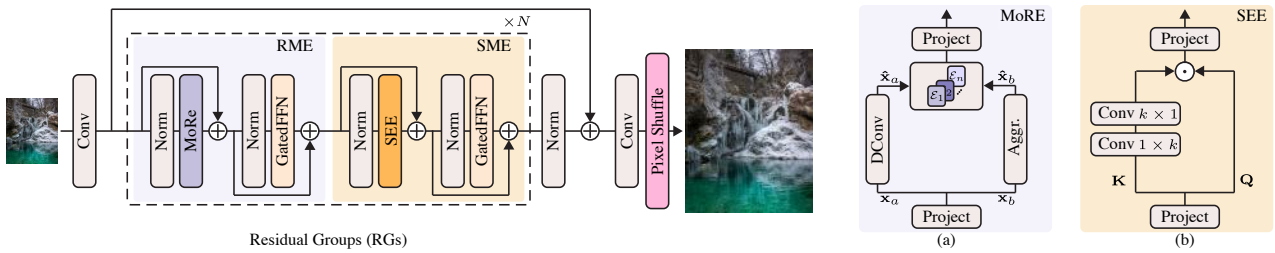

Figure 2. Architecture Overview. **S**eemo**R**e refines the feature representations via stacked Residual groups (RGs). Each RG consists of a Rank Modulating Exert (RME) and a Spatial Modulating Expert (SME). RME leverages the Mixture of Low Rank Expertise (MoRE) to refine the global texture, while SME employs spatial enhancement experts (SEE) to supplement RME with spatial cues.

nel enhancement modules (Liu et al., 2020b). More recently, ShuffleMixer (Sun et al., 2022) integrates large kernel convolutions and feature shuffling, improving both computational efficiency and high-resolution reconstruction. SAFMN (Sun et al., 2023) improves the efficiency by collecting non-local features using a shallow pyramid. Despite improvements in several efficiency aspects brought up by the aforementioned approaches, there is still scope for a better trade-off between model efficiency and the restoration performance.

Dynamic Networks. Dynamic networks have been extensively studied to optimize the balance between speed and performance across various tasks. Early research employed conditional computation to selectively activate network segments at different times (Bengio et al., 2013). More recently, Mixture-of-Experts (MoE) approaches with routing architecture (Shazeer et al., 2017; Riquelme et al., 2021; Puigcerver et al., 2024) have expanded model capacity without significantly increasing inference costs, primarily enhancing the feed-forward capacity of Transformers in Natural language processing (Shazeer et al., 2017) and high-level vision tasks (Riquelme et al., 2021; Puigcerver et al., 2024). A similar idea can be found in image restoration, where PathRestore (Yu et al., 2021) dynamically routes image patches to different network paths based on content and distortion, leveraging a difficulty-regulated reward function. In this work, our research explores the routing concept from an architecture design perspective for image super-resolution, aiming to discover the most efficient and appropriate expert to improve the feature modeling.

3. Methodology

In this section, we unveil the fundamental components of our proposed model tailored for efficient super-resolution. As demonstrated in Figure 2, our overall pipeline embodies a sequence of N residual groups (RGs) and an upsampler layer. The initial step involves applying a 3×3 convolution layer to generate the shallow features from the input low-resolution (LR) image. Subsequently, multiple stacked RGs are deployed to refine the deep features, easing the reconstruction of high-resolution (HR) images while maintaining efficiency. Each RG consists of a Rank modulating expert (RME) and a Spatial modulating expert (SME). Lastly, a global residual connection links the shallow features to the output of the deep features for capturing the high-frequency details and an up-sampler layer (3×3 and pixel-shuffle (Shi et al., 2016)) is deployed for faster reconstruction.

3.1. Rank Modulating Expert

Unlike large kernel convolution (Hou et al., 2022) or self-attention (Vaswani et al., 2017) that rely upon resource-intensive matrix operations for modelling the LR-HR dependencies, we opt for modulating the most relevant interactions in low-rank in our quest for efficiency. Our proposed Rank modulating expert (RME) (see Figure 2) explores a Transformer alike architecture using Mixture of Low-Rank Expertise (MoRE) for modelling the relevant global informative features efficiently and a GatedFFN (Chen et al., 2023) for refined contextual feature aggregation.

Mixture of Low-Rank Expertise. As illustrated in Figure 3, from a layer normalised input tensor $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, we use a 3×3 convolution for feature projection and then we split along the channel dimension to create two distinct views $\mathbf{x}_a$ and $\mathbf{x}_b \in \mathbb{R}^{H\times W\times C}$. To efficiently aggregate the pixel-wise cross-channel context, we leverage a recursive strided convolution t times followed by a refinement and upsampling step, resulting in the construction of the feature pyramid denoted as $\hat{\mathbf{x}}_b \in \mathbb{R}^{H\times W\times C}$. The process is formulated as follows:

$$|p|_{\downarrow h\times w} = \text{DConv}^{s}_{k\times k}(...(\text{DConv}^{s}_{k\times k}(\mathbf{x}_b)) \quad (1)$$

$$\hat{\mathbf{x}}_b = |\ \mathbf{W}_{C\to C}(\text{DConv}_{3\times 3}(|p|_{\downarrow h\times w}))\ |_{\uparrow H\times W}, \quad (2)$$

where $\text{DConv}_{k\times k}$ denotes a depth-wise convolution with kernel size k and stride s, $\mathbf{W}_{C\to C}$ denotes a linear layer, p represents the contextual feature pyramid. Simultaneously, a parallel depth-wise convolution extracts the local spatial context $\hat{\mathbf{x}}_a$ before feeding both the extracted feature maps into the mixture of low-rank expertise. This branched parallel design approach is chosen purposefully. In general, the

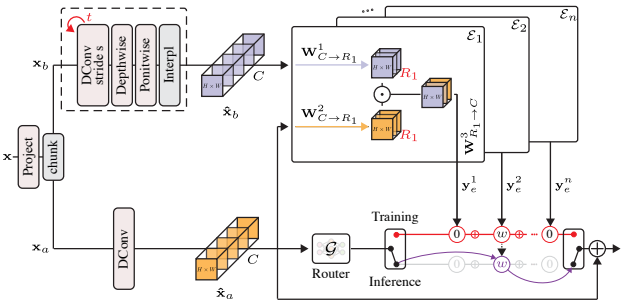

Figure 3. Illustration of the proposed Mixture of Low-Rank Expertise (MoRE) as a core block of the RME.

downsampling of the feature maps impacts the reconstruction performance of SR methods. Therefore, we maintain the same resolution for general feature extraction while incorporating an additional path to capture global contextual cues efficiently, thereby circumventing any information loss.

To further delve into the intricacies of the inter-dependencies among the extracted features for reducing complexity, we deploy low-rank decomposition for the inputs while modeling the global contextual relationships. As demonstrated in Figure 3, a single low-rank expert ($\mathcal{E}$), takes as input the spatial features, $\hat{\mathbf{x}}_a$, and encoded pixel-wise contextual cues, $\hat{\mathbf{x}}_b$ and is formulated as:

$$\mathcal{E}_i = \mathbf{W}^3_{R_i \to C}(\mathbf{W}^1_{C \to R_i}\hat{\mathbf{x}}_a \odot \mathbf{W}^2_{C \to R_i}\hat{\mathbf{x}}_b), \tag{3}$$

where the linear layers denoted as $\mathbf{W}_{C \to R_i}$, compress the encoded features along the channel dimension to their low-rank approximation R_i, where $i \in \{1, ..., n\}$. After adeptly modulating the spatial cues through element-wise multiplication with the contextual cues in low-dimensional space, another linear layer $\mathbf{W}^3_{R_i \to C}$ extends the features back to the original dimension C to extract the relevant channel-wise spatial content. Thereby, implicitly mixing the crucial spatial and channel dependencies in an efficient way.

However, manually determining the optimal low-rank (R) may not fully leverage all the inherent information for modulation, leading to underutilized model capacity. Thus, we employ a dynamic approach using a mixture of different low rank experts, with a routing network ($\mathcal{G}$) that systematically explores the search space to identify the ideal low-rank expert based on the input and network depth. Following (Shazeer et al., 2017), the final output $\mathbf{y}$ of the mixture of low-rank experts is as follows:

$$\mathbf{y} = \sum_{i}^{n} \mathcal{G}(\hat{\mathbf{x}}_a)\mathcal{E}_i(\hat{\mathbf{x}}_a, \hat{\mathbf{x}}_b) + \hat{\mathbf{x}}_a, \tag{4}$$

where $\mathcal{G}(\cdot)$ and $\mathcal{E}_i(\cdot)$ denote the learned routing function and the output of the i-th expert, respectively. The sparsity inherent in the router function $\mathcal{G}(\cdot)$ optimizes computation

Algorithm 1 Mixture of Low-Rank Experise

Input: Input feature $\hat{\mathbf{x}}_a$, semantic cues $\hat{\mathbf{x}}_b$
Parameters: n Experts $\mathcal{E}$, Router $\mathcal{G}$, Low-Rank dimensions $R_i = 2^{i+1}$ with $i \in \{1, ..., n\}$, top-1 expert $k = 1$
Compute router outputs: $\mathbf{g} = \mathcal{G}(\hat{\mathbf{x}}_a)$
Normalize weights: $\mathbf{w} = \text{Softmax}(\mathbf{g})$
Select top-1 expert: $w_{\text{top-1}} = \text{topk}(\mathbf{w}, k = 1)$
Set all other weights to zero: $\mathbf{w}_i = 0$ for $i \neq$ top-1
if training **then**
  for each $e \in \mathcal{E}$ **do**
    $\mathbf{y}^i_e = \mathbf{W}^3_{R_i \to C}(\mathbf{W}^1_{C \to R_i}\hat{\mathbf{x}}_a \odot \mathbf{W}^2_{C \to R_i}\hat{\mathbf{x}}_b)$
  end for
  Compute final output: $\mathbf{y} = \sum_{i=1}^{n} w_i \cdot \mathbf{y}^i_e$
else
  Compute final output: $\mathbf{y} = w_{\text{top-1}} \cdot \mathbf{y}^{\text{top-1}}_e$
end if
Output: Final output $\mathbf{y}$

by assigning greater weights to the top-k low-rank experts. While at training time, our method learns from different experts, during inference, only the selected top-k expert is utilized for computation, further enhancing the efficiency. More specifically, the inference complexity is not proportional to the number of experts.

Adhering to the MoE concept with $k > 1$, our routing function for optimal low-rank representation extends sparse routing principles (Shazeer et al., 2017) by selecting only the top-1 expert. As our work is pioneering in this domain, we emphasize a more interpretable top-1 design, as shown in Figure 5b, which allows us to streamline the model architecture and computational process, creating an efficient yet powerful image super-resolution model. Technically, both training and inference leverage dynamic expert selection based on input and model depth; however, only the top-1 expert per layer is utilized, with contributions from other experts weighted at zero. During inference, inactive experts are disregarded to efficiently exploit contextual information using the optimal input-dependent expert chosen by the router. This ensures consistency between training and inference, as only one expert per layer remains active, thereby mitigating potential discrepancies. In Table 9 found in the supplementary, we show that augmenting the number of top-k experts can slightly improve the performance, at the cost of increased computational complexity. We hope that our network can serve as a fresh baseline for future development.

Additionally, the design choices contributing towards the selection of the number of low-rank experts ($\mathcal{E}_i$) and the rank dimension (R) for memory-efficient reconstruction is illustrated in Table 5a of the ablation study. We also provide the pseudocode for the proposed MoRE block in Algorithm 1.

In addition to the primary analyses presented in the main text, the supplementary material offer further insights and experiments that substantiate the design decisions of our proposed MoRE module. For detailed information, refer to Tables 11 and 14.

3.2. Spatial Modulating Expert

We observe that the rank modulating expert is more dedicated towards investigating the global channel-wise contextual information, and its effectiveness would be complemented by the spatial-wise local information. Inspired by the previous work in classification (Yang et al., 2022; Hou et al., 2022), we design a spatial modulating expert (SME) (see Figure 2) comprising of a spatial enhancement expertise (SEE) block that efficiently captures the spatial-wise coupling followed by a GatedFFN (Chen et al., 2023) for feature refinement.

Spatial Enhancement Expertise. While the vanilla self-attention (SA) mechanism (Vaswani et al., 2017) creates connections among all the input pixels, effectively capturing the relevant context, its quadratic computational complexity with image size poses limitations, particularly in high-resolution scenarios like image SR. Thus, our spatial enhancement expertise simplifies the computation of the similarity matrix $\mathbf{A}$ between keys $\mathbf{K}$ and queries $\mathbf{Q}$ by utilizing a striped depth-wise convolution with a large kernel, sequentially convolving the feature maps with $\mathbf{k_1} \in \mathbb{R}^{[1,k]}$ followed by $\mathbf{k_2} \in \mathbb{R}^{[k,1]}$. Specifically, we compute the locally enhanced spatial-wise features as follows:

$$\mathbf{x}_{out} = \text{DConv}^{s}_{k\times k}(\mathbf{W}^{4}_{C\to C}\mathbf{x}_{in}) \odot \mathbf{W}^{5}_{C\to C}\mathbf{x}_{in}, \quad (5)$$

where $\odot$ is the Hadamard product, $\mathbf{W}^{4}_{C\to C}$ and $\mathbf{W}^{5}_{C\to C}$ are linear (project) layers, $\text{DConv}^{s}_{k\times k}$ denotes the striped depth-wise convolution, and $\mathbf{x}_{in}$ is the layer normalised output of the RME. The use of a large-kernel convolution facilitates a localized correlation among the pixels within the $k \times k$ window, emulating the window-based SA layers frequently employed in image restoration (Liu et al., 2021; Zamir et al., 2022; Chen et al., 2023), all the while preserving the efficiency benefits associated with convolutional layers as demonstrated in Table 4a.

4. Experiments

Datasets and Evaluation. Following the SR literature (Liang et al., 2021; Chen et al., 2023), we utilize DIV2K (Agustsson & Timofte, 2017) and Flickr2K (Lim et al., 2017) datasets for training. We produce LR images using bicubic downscaling of HR images. When testing our method, we assess its performance on canonical benchmark datasets for SR - Set5 (Bevilacqua et al., 2012), Set14 (Zeyde et al., 2010), BSD100 (Martin et al., 2001), Urban100 (Huang et al., 2015) and Manga109 (Matsui et al., 2017). We calculate PSNR and SSIM results on the Y-channel from the YCbCr color space.

Implementation Details. We augment our training data with randomly extracted 64×64-sized crops, with random rotation, horizontal and vertical flipping. Similar to (Sun et al., 2022; 2023), we minimize the L1-Norm between SR output and HR ground truth in the pixel and frequency domain using Adam (Kingma & Ba, 2017) optimizer for 500K iterations with a batch size of 32 and initial learning rate of 1×10^{-3} halving it at following milestones: $[250K, 400K, 450K, 475K]$. All experiments are conducted with the PyTorch framework on NVIDIA RTX 4090 GPUs. We design our smallest model (SeemoRe-T) with 6 RGs. The feature dimension and channel expansion factor in GatedFFN are set to 36 and 2, respectively. For all MoRE sub-modules, we select an exponential growth of the channel dimensionality and choose in total of 3 experts. The kernel size in SEE is set to 11×11. More details can be found in the supplemental, *c.f.* Table 6.

4.1. Comparison to State-of-the-Art Methods

We present quantitative results for $\times 2$, $\times 3$, and $\times 4$ image SR, comparing against current efficient state-of-the-art models in Table 1, including CARN-M (Ahn et al., 2018), IMDN (Hui et al., 2019), PAN (Zhao et al., 2020), DRSAN (Park et al., 2021), DDistill-SR (Wang et al., 2022), ShuffleMixer (Sun et al., 2022), and SAFMN (Sun et al., 2023). Additionally, we evaluate against lightweight variants of popular Transformer-based SR models such as SwinIR (Liu et al., 2021), ELAN (Zhang et al., 2022), and SRFormer (Zhou et al., 2023) in Table 2. Our proposed SeemoRe-T stands out as the most efficient method, consistently surpassing all other methods across all benchmarks and scale factors. For instance as clear from Table 1, on the Urban100 and Manga109 benchmarks ($\times 2$), SeemoRe-T outperforms SAFMN (Sun et al., 2023) by 0.41dB and 0.30dB, respectively. Furthermore, with 47% fewer parameters and 65% fewer GMACS than DDistill-SR (Wang et al., 2022), SeemoRe-T achieves on average 0.12 dB higher PSNR results across all benchmarks ($\times 4$). Scaling our method up to a comparable size with lightweight Transformers yields comparable or superior results. As demonstrated in Table 2, our SeemoRe-L outperforms SwinIR-Light and SRFormer-Light on Manga109 ($\times 4$) by 0.57dB and 0.31dB, while requiring fewer GMACS.

Visual Results. We show visual comparisons ($\times 4$) in Figure 4. In some challenging scenarios, the previous methods may suffer blurring artifacts, distortions, or inaccurate texture restoration. Contrary to others, our SeemoRe alleviates the blurring artifacts better and maintains structural fidelity.

Table 1. Comparison to efficient SR models. PSNR (dB ↑) and SSIM (↑) metrics are reported on the Y-channel. Best and second best performances are highlighted. GMACS (G) are computed by upscaling to a 1280 × 720 HR image. SeemoRe-T achieves state-of-the-art performance across all benchmarks with the lowest parameter count and computational demand. '-' represents unreported results.

	Method	Params	GMACS	Set5		Set14		BSD100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2	Bicubic	-	-	33.66	.9299	30.24	.8688	29.56	.8431	26.88	.8403	30.80	.9339
	CARN-M (Ahn et al., 2018)	412K	91	37.53	.9583	33.26	.9141	31.92	.8960	31.23	.9193	-	-
	IMDN (Hui et al., 2019)	694K	159	38.00	.9605	33.63	.9177	32.19	.8996	32.17	.9283	38.88	.9774
	PAN (Zhao et al., 2020)	261K	71	38.00	.9605	33.59	.9181	32.18	.8997	32.01	.9273	38.70	.9773
	DRSAN (Park et al., 2021)	370K	86	37.99	.9606	33.57	.9177	32.16	.8999	32.10	.9279	-	-
	DDistill-SR (Wang et al., 2022)	414K	128	38.03	.9606	33.61	.9182	32.19	.9000	32.18	.9286	38.94	.9777
	ShuffleMixer (Sun et al., 2022)	394K	91	38.01	.9606	33.63	.9180	32.17	.8995	31.89	.9257	38.83	.9774
	SAFMN (Sun et al., 2023)	228K	52	38.00	.9605	33.54	.9177	32.16	.8995	31.84	.9256	38.71	.9771
	SeemoRe-T (*ours*)	220K	45	38.06	.9608	33.65	.9186	32.23	.9004	32.22	.9286	39.01	.9777
×3	Bicubic	-	-	30.39	.8682	27.55	.7742	27.21	.7385	24.46	.7349	26.95	.8556
	CARN-M (Ahn et al., 2018)	415K	46	33.99	.9236	30.08	.8367	28.91	.8000	27.55	.8385	-	-
	IMDN (Hui et al., 2019)	703K	72	34.36	.9270	30.32	.8417	29.09	.8046	28.17	.8519	33.61	.9445
	PAN (Zhao et al., 2020)	261K	39	34.40	.9271	30.36	.8423	29.11	.8050	28.11	.8511	33.61	.9448
	DRSAN (Park et al., 2021)	410K	43	34.41	.9272	30.27	.8413	29.08	.8056	28.19	.8529	-	-
	DDistill-SR (Wang et al., 2022)	414K	57	34.37	.9275	30.34	.8420	29.11	.8053	28.19	.8528	33.69	.9451
	ShuffleMixer (Sun et al., 2022)	415K	42	34.40	.9272	30.37	.8423	29.12	.8051	28.08	.8498	33.69	.9448
	SAFMN (Sun et al., 2023)	233K	23	34.34	.9267	30.33	.8418	29.08	.8048	27.95	.8474	33.52	.9437
	SeemoRe-T (*ours*)	225K	20	34.46	.9276	30.44	.8445	29.15	.8063	28.27	.8538	33.92	.9460
×4	Bicubic	-	-	28.42	.8104	26.00	.7027	25.96	.6675	23.14	.6577	24.89	.7866
	CARN-M (Ahn et al., 2018)	415K	33	31.92	.8903	28.42	.7762	27.44	.7304	25.62	.7694	-	-
	IMDN (Hui et al., 2019)	715K	41	32.21	.8948	28.58	.7811	27.56	.7353	26.04	.7838	30.46	.9075
	PAN (Zhao et al., 2020)	272K	28	32.13	.8948	28.61	.7822	27.59	.7363	26.11	.7854	30.51	.9095
	DRSAN (Park et al., 2021)	410K	31	32.15	.8935	28.54	.7813	27.54	.7364	26.06	.7858	-	-
	DDistill-SR (Wang et al., 2022)	434K	33	32.23	.8960	28.62	.7823	27.58	.7365	26.20	.7891	30.48	.9090
	ShuffleMixer (Sun et al., 2022)	411K	28	32.21	.8953	28.66	.7827	27.61	.7366	26.08	.7835	30.65	.9093
	SAFMN (Sun et al., 2023)	240K	14	32.18	.8948	28.60	.7813	27.58	.7359	25.97	.7809	30.43	.9063
	SeemoRe-T (*ours*)	232K	12	32.31	.8965	28.72	.7840	27.65	.7384	26.23	.7883	30.82	.9107

For instance, in image *img60* and *img73* from Urban100, certain methods like DDistill-SR, SwinIR-Light and DAT-Light fail to accurately reconstruct shadow patterns or window struts, whereas our method exhibits strong recovery of fine details. These visual comparisons highlight SeemoRe's ability to reconstruct high-quality images by effectively leveraging local and contextual information. Coupled with quantitative comparisons, these findings underscore the effectiveness of our method. More visual results can be found in the Supplementary material.

4.2. Model Complexity Trade-Off

In the vision domain, scalability becomes more paramount. We strive to expand the limits of our SeemoRe framework, optimizing for both reconstruction fidelity and efficiency. The framework provides three complexity scales—tiny (T), base (B), and large (L)—with progressively improved reconstruction performance, *c.f* Figure 1. In Table 3, we present comparisons of memory usage and running time, demonstrating that our SeemoRe-T outperforms representative state-of-the-art methods. By using the low-rank feature modulation and simultaneous aggregation of the channel-spatial dependencies, the GPU consumption of our SeemoRe-T is 3% less than DDistill-SR, while being 2 times faster. Additionally, Table 3 highlights the significant efficiency advantage of SeemoRe over lightweight Transformers. Further results are provided in the Supplemental. To further underscore our method's capability, we align SwinIR-Light and SRFormer-Light with a size and computational demand similar to ours, followed by retraining these downsized networks using our schedule. The results presented in Table 3 highlight that SeemoRe-T significantly outperforms both Transformer-based models by a considerable margin.

4.3. Ablation Study

We conduct detailed studies on the components within our approach. All experiments are conducted on the ×2 setting.

Macro Architecture. As reported in Table 4a, we evaluate the effectiveness of our proposed key architectural components by comparing them with a baseline model consisting solely of depthwise and pointwise convolutions, more details in Supplemental. After adding the proposed modules into the baseline model, their is a notable and persistent

Table 2. Comparison to lightweight SR Transformers. PSNR (dB ↑) and SSIM (↑) metrics are reported on the Y-channel. Best and second best performances are highlighted. GMACS (G) are computed by upscaling to a 1280 × 720 HR image. SeemoRe-L outperforms or achieves comparable performance to compared Transformers while being more efficient. ×3 results are in the Supplemental.

	Method	Params	GMACS	Set5		Set14		BSD100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2	Bicubic	-	-	33.66	.9299	30.24	.8688	29.56	.8431	26.88	.8403	30.80	.9339
	SwinIR-Light (Liang et al., 2021)	910K	244	38.14	.9611	33.86	.9206	32.31	.9012	32.76	.9340	39.12	.9783
	ELAN-Light (Zhang et al., 2022)	621K	201	38.17	.9611	33.94	.9207	32.30	.9012	32.76	.9340	39.11	.9782
	SRFormer-Light (Zhou et al., 2023)	853K	236	38.23	.9613	33.94	.9209	32.36	.9019	32.91	.9353	39.28	.9785
	ESRT (Lu et al., 2022)	751K	191	38.03	.9600	33.75	.9184	32.25	.9001	32.58	.9318	39.12	.9774
	SwinIR-NG (Choi et al., 2023)	1181K	274	38.17	.9612	33.94	.9205	32.31	.9013	32.78	.9340	39.20	.9781
	DAT-Light (Chen et al., 2023)	553K	194	38.24	.9614	34.01	.9214	32.34	.9019	32.89	.9346	39.49	.9788
	SeemoRe-L (*ours*)	931K	197	38.27	.9616	34.01	.9210	32.35	.9018	32.87	.9344	39.49	.9790
×4	Bicubic	-	-	28.42	.8104	26.00	.7027	25.96	.6675	23.14	.6577	24.89	.7866
	SwinIR-Light (Liang et al., 2021)	897K	64	32.44	.8976	28.77	.7858	27.69	.7406	26.47	.7980	30.91	.9151
	ELAN-Light (Zhang et al., 2022)	601K	54	32.43	.8975	28.78	.7858	27.69	.7406	26.54	.7982	30.92	.9150
	SRFormer-Light (Zhou et al., 2023)	873K	63	32.51	.8988	28.82	.7872	27.73	.7422	26.67	.8032	31.17	.9165
	ESRT (Lu et al., 2022)	751K	68	32.19	.8947	28.69	.7833	27.69	.7379	26.39	.7962	30.75	.9100
	SwinIR-NG (Choi et al., 2023)	1201K	63	32.44	.8980	28.83	.7870	27.73	.7418	26.61	.8010	31.09	.9161
	DAT-Light (Chen et al., 2023)	573K	50	32.57	.8991	28.87	.7879	27.74	.7428	26.64	.8033	31.37	.9178
	SeemoRe-L (*ours*)	969K	50	32.51	.8990	28.92	.7888	27.78	.7428	26.79	.8046	31.48	.9181

Table 3. Complexity Analysis. Runtime (ms, ↓) and memory consumption (M, ↓) averaged across 200 samples using a NVIDIA RTX 4090 device.

Method	Input	Scale	Time	GPU Memory
DAT-Light			210.12	8715.1
SwinIR-Light			131.25	6175.3
SRFormer-Light			103.95	7270.1
SeemoRe-L (*ours*)	[320, 180]	×4	**17.99**	9531.6
ShuffleMixer			7.40	**1380.7**
DDistill-SR			11.20	2822.1
SeemoRe-T (*ours*)			**5.66**	2744.9

(a) Runtime and memory consumption.

Scale	Method	Params.	GMACS	Urban100	Manga109
×2	SwinIR*	191K	48	31.56	38.07
	SRFormer*	188K	49	31.60	38.59
	DAT*	115K	44	31.91	38.80
	SeemoRe-T	**220K**	**45**	**32.22**	**39.01**

(b) PSNR (dB ↑) on the Y-Channel. ∗ denotes retrained models.

Table 4. Ablation on Blocks. GMACS (↓) are computed by upscaling to a 1280 × 720 HR image. We show results for ×2 upscaling.

Method	RME	SME	Params.	GMACS	Urban100	Manga109
Baseline	-	-	157K	35	31.61	38.55
	✓	-	199K	40	31.96	38.75
SeemoRe-T	-	✓	178K	40	31.97	38.87
	✓	✓	**220K**	**45**	**32.22**	**39.01**

(a) Contribution of components.

Method	Macro Expert Order	Urban100	Manga109
SeemoRe-T	Spatial - Rank	32.17	38.93
	Rank - Spatial	**32.22**	**39.01**

(b) Block order.

Method	Kernel k	Params.	GMACS	Urban100	Manga109
SeemoRe-T	3	217K	44	32.04	38.84
	7	219K	45	32.10	38.96
	11	**220K**	**45**	**32.22**	**39.01**

(c) Kernel size (k) variation.

improvement in the results. The incorporation of RME or SME results in improvements of 0.26 dB or 0.32 dB on Urban100 over the baseline, respectively. Although both modules individually outperform the baseline with only a marginal increase in parameters, alternating the insertion of both the modules within each RG fully unleashes the model's capabilities while enhancing the overall efficiency. Overall, our SeemoRe-T obtains a compelling gain of 0.49 dB and 0.38 dB on Urban100 and Manga109, respectively. Moreover, Table 4b empirically justifies the chosen block ordering, showcasing the Rank-Spatial macro order design's superiority over permuted Spatial-Rank macro order. This empirical evidence supplements the qualitative justifications in Section 5 regarding the individual importance of MoRE and SEE blocks.

Design choices of SME. The main component of SME module, SEE deploys striped convolutions with large-kernel sizes to effectively module the spatial cues. Table 4 demonstrates that deploying large kernel sizes improves the overall performance of the model. In particular, the PSNR shows a notable gain of 0.18 dB on Urban100 dataset when increasing the kernel size from 3×3 to 11×11 (keeping other settings intact), with only 3K increase in parameters. It

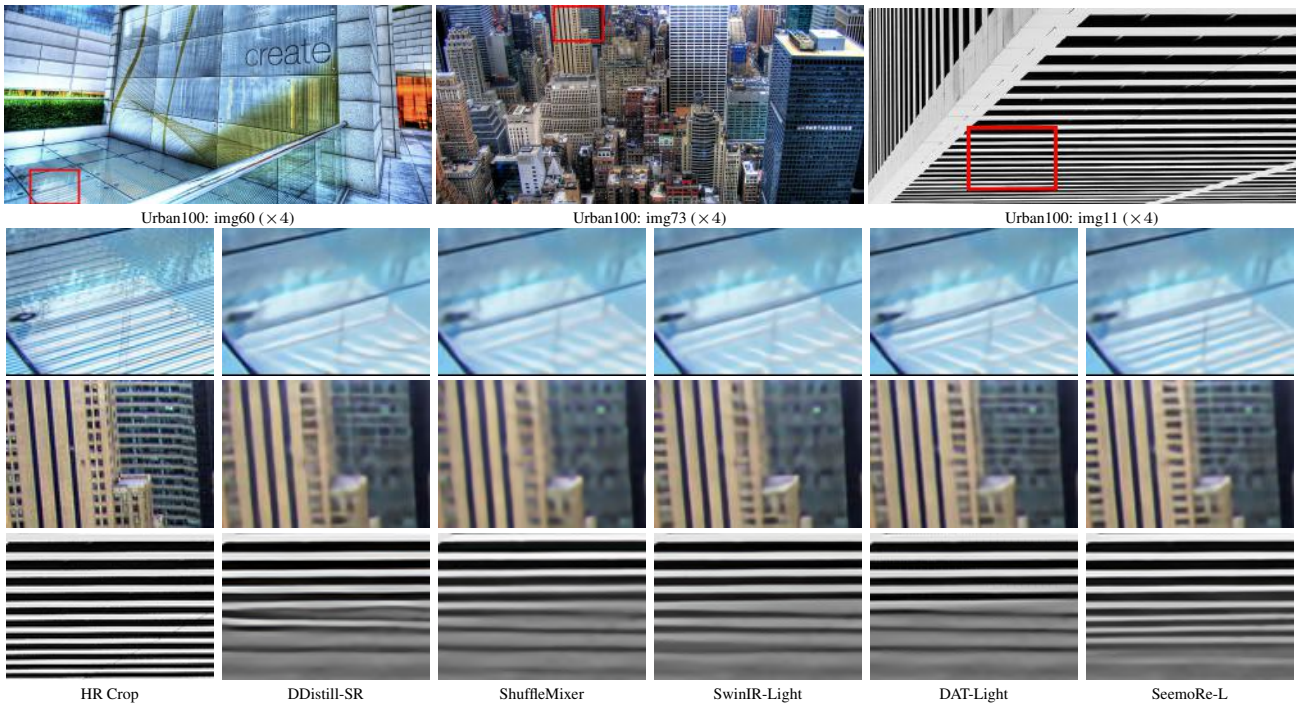

Figure 4. Visual comparison of SeemoRe with state-of-the-art methods on challenging cases for ×4 SR from the Urban100 benchmark.

Table 5. Ablation on MoRE. Exponential growth yields best performance in terms of parameter counts and PSNR. $\#\mathcal{E}$ denotes the number of experts and *Dim.* the rank dimensionality. We show results for ×2 upscaling.

	Method	$\#\mathcal{E}$	Dim.	Params.	Urban100	Manga109
	SeemoRe-T					
Growth	$i+2$	6	2,3,4,5,6,7	229K	32.12	39.01
	$2*i+2$	4	2,4,6,8	224K	32.11	38.98
	2^i	4	2,4,8,16	231K	32.21	**39.02**
$\#\mathcal{E}$	2^i	3	$2, 4, 8$	**220K**	**32.22**	39.01
		2	$2, 4$	214K	32.19	39.00
		1	2	211K	32.16	38.92

(a) Low-rank expert design.

Method	Recursive Steps t	Urban100	Manga109
SeemoRe-T	1	32.10	38.99
	2	**32.22**	**39.01**
	3	31.04	38.13

(b) Recursive step (t) variation.

clearly proves that such a design benefits in the efficient use of the relevant information to augment the restoration of sharp regions spatially.

Design choices of RME. We motivate our design choices for the MoRE module in RME by varying the growth function and the number of experts as depicted in Table 5a. When pursuing a dynamic solution for determining the optimal low-rank dimensionality, it becomes necessary to design the corresponding search space. First, we present results for ×2 upscaling on Urban100 and Manga109 using different growth functions. Based on the observed outcomes, it is evident that an exponentially increasing low-rank dimensionality yields the best performance with marginal increase in the parameters. Hence, we opt to retain this search space design in all further experimentation. Next, we analyze the reconstruction quality based on the number of experts in each MoRE module, while exponentially increasing the low-rank dimensionality. Based on these experiments, we assert that the efficient results are obtained when we have three, as our total number of experts. Further, we ablate the choice of recursive steps for SeemoRe-T in Table 5b, where our plain version takes ($t = 2$). It can be seen that lower ($t = 1$) and higher ($t = 3$) values either fail to capture sufficient contextual information or overly compromise spatial image features.

5. Discussion on Experts

Our model integrates experts at varying levels, each specializing in crucial factors for SR. In this section, we aim to elucidate their expertise.

Mixture of Low-Rank Experts. The decision-making process of the router at different network depths is illustrated in Figure 5a. Notably, earlier blocks showcase a diverse

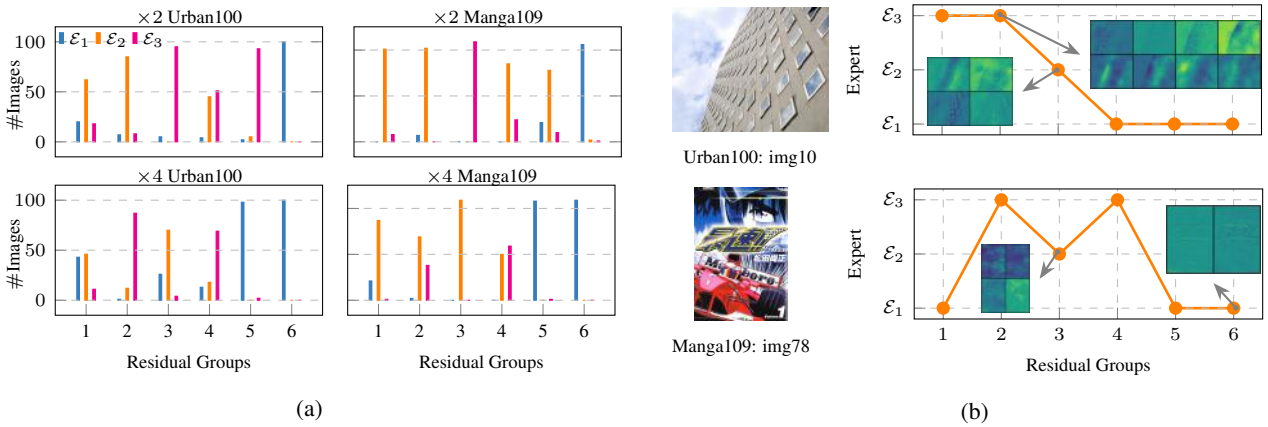

Figure 5. Low-Rank Analysis. (a) We plot the decisions made by the routing function for SeemoRe-T over the depth of the network.(b) We visualize the low-rank features of SeemoRe-T for ×4 SR given example images from Urban100 and Manga109.

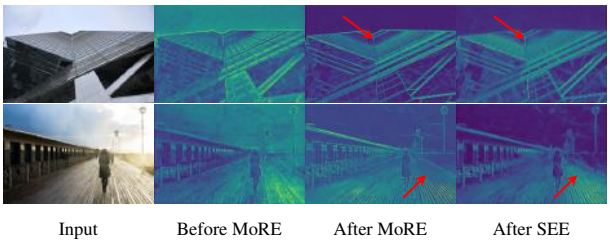

Figure 6. Feature Visualization. We present visualizations of feature maps before and after our proposed modules. Clearly, our MoRE block notably enhances activation sharpness via contextual feature modulation. Moreover, our SEE module improves learned representations by integrating spatial cues effectively.

range of rank choices ($\mathcal{E}_1$,$\mathcal{E}_2$,$\mathcal{E}_3$), while deeper layers tend to favor lower ranks ($\mathcal{E}_1$) (Please note that for every $\mathcal{E}_i$, the corresponding rank dimension is 2^i). This phenomenon can be attributed to the hierarchical feature learning nature of deep neural networks, aligning with our expectations. In fact, earlier layers typically capture low-level details and, at times, unwanted noise while reconstructing details in the input LR image, thus resulting in wide variations in the rank choices. In contrast, deeper layers focus on the main structures and key features required for SR. Hence, higher ranks at deeper layers are less favored, as they may introduce redundancy or noise that does not significantly contribute to the overall quality of the reconstructed image. This design aspect provides our method with the flexibility to adapt to the complexity of the task, a capability that, to the best of our knowledge, has not yet been explored in the image reconstruction community. In Figure 5b, we further visualize the routing decisions and the corresponding low-rank feature maps for two exemplary input images. It is noteworthy that each individual rank carries distinct information while being mutually complementary. As the model depth increases, the network becomes proficient in restructuring these representations.

How important are MoRE and SEE? To substantiate the significance of the proposed MoRE and SEE modules, we analyze the feature maps before and after integrating both blocks into the RME module as depicted in Figure 6. This analysis vividly showcases the advantages of leveraging MoRE for contextual information mining within RME. Notably, the activations exhibit reduced noise and enhanced sharpness. Additionally, we observe a synergistic interaction between MoRE and SEE at marked locations (indicated by red arrows): MoRE effectively refines global textures by filtering out noise, while SEE supplements over-filtered regions with critical local details.

6. Conclusion

We propose a novel ConvNet, named **S**eemo**R**e, for efficient and accurate image super-resolution. Our **S**eemo**R**e excels in modeling local and contextual information, surpassing both previous CNN-based and lightweight Transformer approaches in terms of efficiency and reconstruction fidelity. Unlike other approaches, we empirically demonstrate both the scalability of efficiency and reconstruction performance. In our approach, we intricately design the rank modulation expert to discern the most pivotal features, enhancing this compressed representation with valuable contextual cues. Our spatial enhancement expert efficiently integrates local spatial-wise information, unlocking the full potential of our architecture. This novel approach optimally exploits the low information regime in the input image, enhancing detail reconstruction while improving efficiency. Extensive experiments on image super-resolution demonstrate that our proposed **S**eemo**R**e achieves consistent superior performance over recent state-of-the-art efficient methods on all considered SR benchmarks, while even being on par with the lightweight Transformers in terms of reconstruction fidelity.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, specifically efficient image super-resolution. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here. However, applying super-resolution methods in AI-assisted software raises ethical concerns about privacy invasion and increased surveillance capabilities. Adherence to transparency, accountability, and privacy rights is crucial to mitigate potential harm and ensure responsible deployment in alignment with societal values.

Acknowledgments

This work was supported by The Alexander von Humboldt Foundation.

References

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

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

Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

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

Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., and Yu, F. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Choi, H., Lee, J., and Yang, J. N-gram in swin transformers for efficient lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2071–2081, 2023.

Conde, M. V., Zamfir, E., and Timofte, R. Efficient deep models for real-time 4k image super-resolution. ntire 2023 benchmark and report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1495–1521, June 2023.

Dong, C., Loy, C. C., He, K., and Tang, X. Learning a deep convolutional network for image super-resolution. In *Proceeding of the European Conference on Computer Vision*, pp. 184–199. Springer, 2014.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Duchon, C. E. Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology and Climatology*, 1979.

Hou, Q., Lu, C.-Z., Cheng, M.-M., and Feng, J. Conv2former: A simple transformer-style convnet for visual recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

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

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

Ignatov, A., Timofte, R., Denna, M., and Younes, A. Real-time quantized image super-resolution on mobile npus, mobile ai 2021 challenge: Report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2525–2534, 2021.

Ignatov, A., Timofte, R., Denna, M., Younes, A., Gankhuyag, G., Huh, J., Kim, M. K., Yoon, K., Moon, H.-C., Lee, S., et al. Efficient and accurate quantized image super-resolution on mobile npus, mobile ai & aim 2022 challenge: report. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pp. 92–129. Springer, 2023.

Khani, M., Sivaraman, V., and Alizadeh, M. Efficient video compression via content-adaptive super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4521–4530, 2021.

Kim, J., Lee, J. K., and Lee, K. M. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1637–1645, 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Kong, F., Li, M., Liu, S., Liu, D., He, J., Bai, Y., Chen, F., and Fu, L. Residual local feature network for efficient super-resolution. In *Proceedings of the European Conference on Computer Vision Workshops*, 2022.

Li, Y., Zhang, K., Timofte, R., Van Gool, L., Kong, F., Li, M., Liu, S., Du, Z., Liu, D., Zhou, C., et al. Ntire 2022 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1062–1102, 2022.

Li, Y., Zhang, Y., Van Gool, L., Timofte, R., et al. NTIRE 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Liang, J., Zeng, H., and Zhang, L. Efficient and degradation-adaptive network for real-world image super-resolution. In *European Conference on Computer Vision*, pp. 574–591. Springer, 2022.

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

Liu, J., Tang, J., and Wu, G. Residual feature distillation network for lightweight image super-resolution. In *Proceedings of the European Conference on Computer Vision Workshops*, 2020a.

Liu, J., Zhang, W., Tang, Y., Tang, J., and Wu, G. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2359–2368, 2020b.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., and Zeng, T. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pp. 416–423. IEEE, 2001.

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

Park, K., Soh, J. W., and Cho, N. I. Dynamic residual self-attention network for lightweight single image super-resolution. *IEEE Transactions on Multimedia*, 2021.

Puigcerver, J., Ruiz, C. R., Mustafa, B., and Houlsby, N. From sparse to soft mixtures of experts. In *The Twelfth International Conference on Learning Representations*, 2024.

Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., and Houlsby, N. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34: 8583–8595, 2021.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

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

Sun, L., Pan, J., and Tang, J. Shufflemixer: An efficient convnet for image super-resolution. *Advances in Neural Information Processing Systems*, 35:17314–17326, 2022.

Sun, L., Dong, J., Tang, J., and Pan, J. Spatially-adaptive feature modulation for efficient image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, H., Chen, X., Ni, B., Liu, Y., and jinfan, L. Omni aggregation networks for lightweight image super-resolution. In *Conference on Computer Vision and Pattern Recognition*, 2023.

Wang, X., Xie, L., Dong, C., and Shan, Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1905–1914, 2021.

Wang, Y., Su, T., Li, Y., Cao, J., Wang, G., and Liu, X. Ddistill-sr: Reparameterized dynamic distillation network for lightweight image super-resolution. *IEEE Transactions on Multimedia*, 2022.

Yang, J., Li, C., and Gao, J. Focal modulation networks. arXiv, 2022.

Yu, K., Wang, X., Dong, C., Tang, X., and Loy, C. C. Path-restore: Learning network path selection for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7078–7092, 2021.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M.-H. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022.

Zeyde, R., Elad, M., and Protter, M. On single image scale-up using sparse-representations. In *Proceedings of International Conference on Curves and Surfaces*, pp. 711–730. Springer, 2010.

Zhang, K., Li, D., Luo, W., Ren, W., Stenger, B., Liu, W., Li, H., and Yang, M.-H. Benchmarking ultra-high-definition image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14769–14778, 2021a.

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

Zhang, X., Zeng, H., Guo, S., and Zhang, L. Efficient long-range attention network for image super-resolution. In *Proceedings of the European Conference on Computer Vision*, 2022.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision*, 2018a.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., and Fu, Y. Residual dense network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018b.

Zhao, H., Kong, X., He, J., Qiao, Y., and Dong, C. Efficient image super-resolution using pixel attention. In *Proceedings of the European Conference on Computer Vision Workshops*, pp. 56–72, 2020.

Zhou, Y., Li, Z., Guo, C.-L., Bai, S., Cheng, M.-M., and Hou, Q. Srformer: Permuted self-attention for single image super-resolution. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

See More Details: Efficient Image Super-Resolution by Experts Mining – Appendix –

Table 6. Implementation Details.

Parameter	SeemoRe-T	SeemoRe-B	SeemoRe-L
Num. RGs	6	8	16
Channel dimension	36	48	48
MLP-Ratio		2	
LR dimensionality growth		exponential	
Num. Experts $\mathcal{E}$		3	
Top-k experts		1	
SFM kernel size		11	
Recursion steps	2	2	1
Training Dataset		DIV2K + Flickr2K	
Optimizer		Adam	
Batch size		32	
Total Num. Iterations		$500k$	
FFT Loss weight		0.1	
LR-Rate		1e.3	
LR-Decay Rate		0.5	
LR-Decay Milestones		$[250K,400K,450K,475K]$	

A. Further Implementation Details

Table 6 outlines the architectural configurations and training settings employed to achieve the reported results in this study. Throughout all our experiments, we maintained a fixed random seed for reproducibility purposes. We based our implementation on the public PyTorch-based *BasicSR*[1] framework for architecture development and training. We use *fvcore*[2] Python package for computing GMACS and parameter counts.

Baseline for Architecture Contribution. Here we provide more details about the baseline method for the ablation in Tables 4a and 7. In the main text, we evaluate SeemoRe-T by sequentially removing the proposed RME and SME blocks, resulting in a plain baseline model with fewer parameters and GMACs. To ensure a fair comparison, we adjust the baseline configuration to match our plain SeemoRe-T model. To ensure roughly equivalent parameter counts and computational complexity, we adopt 5 RGs with a channel dimensionality of 48. Within each RG, we integrate simple convolutional operators from our RME submodule without the MoRE module, while the SME module is simplified to a pointwise convolution.

[1] https://github.com/XPixelGroup/BasicSR
[2] https://github.com/facebookresearch/fvcore

Table 7. Ablation on contribution of components. GMACS (↓) are computed by upscaling to a 1280 × 720 HR image. (*) denotes modified configuration from proposed SeemoRe-T model.

Method	RME	SME	Params.	GMACS	Urban100	Manga109
Baseline*	-	-	232K	52	31.73	38.63
	✓	-	249K	48	31.99	38.75
SeemoRe-T	-	✓	238K	53	32.05	38.96
	✓	✓	**220K**	**45**	**32.22**	**39.01**

Comparison to lightweight SR Transformers ($\times 3$). In Table 8, we present the performance of our SeemoRe-L model for $\times 3$ upscaling, extending the results from Table 2 in the main text. Our SeemoRe-L consistently outperforms other lightweight Transformers and demonstrates only a slightly lower performance compared to DAT-Light (Chen et al., 2023).

B. More Ablations

B.1. Architecture Design

SEE block compared to prior designs. The results in Table 10 prove that our SEE block design outperforms the FusedMB-Conv block proposed by ShuffleMixer (Sun et al., 2022) in terms of reconstruction abilities while maintaining higher efficiency. Moreover, substituting the large-kernel convolution in the Conv2Former (Hou et al., 2022) block with our striped large-kernel variant not only enhances efficiency but also improves the reconstruction capabilities of high-frequency information, as evident from Urban100 results.

MoRE block design. Our rationale behind the MoRE design involves the aggregation of valuable contextual information. Similar to prior works (Liu et al., 2020b), we assign more learning parameters to enhance the high-frequency features while keeping the simple DConv-branch as residual to facilitate the optimization. We further support this rationale with empirical evidence provided in Table 11. The results show that adding the extended feature to the output of DConv performs better than with and without the aggregation output.

Optimization function. In Table 12, we explore the impact of using the L1-Norm in FFT space to compare the

Table 8. Comparison to lightweight SR Transformers. Extension of Table 2. PSNR (dB ↑) and SSIM (↑) metrics are reported on the Y-channel. GMACS ↓) are computed by upscaling to a 1280 × 720 HR image.

	Method	Params	GMACS	SET5		SET14		BSD100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×3	Bicubic	-	-	30.39	.8682	27.55	.7742	27.21	.7385	24.46	.7349	26.95	.8556
	SwinIR-Light (Liang et al., 2021)	918K	111	34.62	.9289	30.54	.8463	29.20	.8082	28.66	.8624	33.98	.9478
	ELAN-Light (Zhang et al., 2022)	629K	90	34.61	.9288	30.55	.8463	29.21	.8081	28.69	.8624	34.00	.9478
	SRFormer-Light (Zhou et al., 2023)	861K	105	34.67	.9296	30.57	.8469	29.26	.8099	28.81	.8655	34.19	.9489
	ESRT (Lu et al., 2022)	770K	96	34.42	.9268	30.43	.8433	29.15	.8063	28.46	.8574	33.95	.9455
	SwinIR-NG (Choi et al., 2023)	1190K	114	34.64	.9293	30.58	.8471	29.24	.8090	28.75	.8639	34.22	.9488
	DAT-Light (Chen et al., 2023)	629K	89	34.76	.9299	30.63	.8474	29.29	.8103	28.89	.8666	34.55	.9501
	SeemoRe-L (*ours*)	959K	87	34.72	.9297	30.60	.8469	29.29	.8101	28.86	.8653	34.53	.9496

Table 9. Ablation on the top-k experts. PSNR (dB ↑) and SSIM (↑) metrics are reported on the Y-channel for ×2 upscaling. GMACS (↓) and memory consumption (M, ↓) are computed by upscaling to a 1280 × 720 HR image using a NVIDIA RTX 4090 device.

Method	Params	GMACS	GPU Memory	SET5		SET14		BSD100		Urban100		Manga109	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SeemoRe-T													
$k = 1$	**220K**	**44.83**	**10972**	38.06	.9608	33.65	**.9186**	**32.23**	**.9004**	32.22	.9286	39.01	.9777
$k = 2$	220K	45.22	11233	38.09	.9608	33.61	.9184	32.22	.9003	32.23	.9286	39.00	.9777
$k = 3$	220K	46.12	11494	**38.10**	**.9609**	**33.66**	.9185	32.23	.9004	**32.24**	**.9289**	**39.08**	**.9779**

Table 10. Analysis of proposed SEE block. We have conducted the following experiment by replacing our proposed SEE with the spatial enhancement module, Fused-MBConv in Shufflemixer (Sun et al., 2022), and Conv Block in Conv2Former (Hou et al., 2022) on ×2 scale.

Method	Params.	GMACS	Urban100	Manga109
SeemoRe-T				
FusedMB (Sun et al., 2022)	304K	64	32.18	38.99
C2F (Hou et al., 2022)	**220K**	46	32.19	**39.04**
SEE (*ours*)	**220K**	**45**	**32.22**	39.01

Table 11. Analysis of MoRE design. We provide further insights in the design decisions of our SeemoRe framework for ×2 upscaling.

Method	Residual t	Urban100	Manga109
SeemoRe-T	No aggregation	32.11	38.96
	Aggregation output	32.15	38.99
	DConv output	**32.22**	**39.01**

model output with high-quality GT images. Compared to utilizing only the traditional L1 loss in RGB space, we observe an average performance improvement of 0.09 dB on Urban100 and Manga109 datasets while using the combined losses. We acknowledge that only a few previous methods incorporate the same FFT loss (Sun et al., 2022; 2023); however, other efficient image super-resolution methods either employ a more intricate training schedule with multiple stages (Liu et al., 2020a; Kong et al., 2022) or utilize large-scale models for knowledge distillation (Wang et al., 2022).

Table 12. Optimization function. SeemoRe-T was trained on DIV2K and Flickr2K. We report PSNR (dB ↑) on the Y-Channel for ×2 upscaling.

Method	L1		FFT		BSD100	Urban100	Manga109
SeemoRe-T	✓	1.0	-	0.0	32.21	32.14	38.90
	✓	1.0	✓	0.1	32.23	**32.22**	39.01
	✓	1.0	✓	0.2	32.22	32.16	**39.02**

Table 13. Model size. PSNR (dB ↑) is reported on the Y-channel. GMACS are computed by upscaling to a 1280 × 720 HR image. N and C denote number of RGs and channel features, respectively.

Method	Config	Params.	GMACS	Urban100	Manga109
SeemoRe-T	N:6 C:36	220K	45	32.22	39.01
SeemoRe-B	N:8 C:48	490K	101	32.52	39.30
SeemoRe-L	N:16 C:48	931K	197	32.87	39.49

Scaling the model size. In Table 13, we detail the architecture, efficiency, and PSNR results across different model sizes on Urban100 and Manga109 datasets. Starting with SeemoRe-T, which has 220K parameters and 45 GMACS, each subsequent complexity stage doubles these figures. Notably, all model stages achieve state-of-the-art performance within their weight classes, with SeemoRe-L matching or even surpassing recent lightweight Transformer-based SR models.

Futhermore, we investigate increasing the number of ex-

Table 14. Scaling up the numbers of experts. We analyze the impact of the number of experts on SeemoRe-T's performance.

Scale	Method	$\#\mathcal{E}$	Growth	Params.	Urban100	Manga109
$\times 2$	SeemoRe-T	8	$2*i+2$	261K	32.18	**39.02**
		4	2^i	231K	32.21	**39.02**
		3	2^i	**220K**	**32.22**	39.01

Table 15. Real SR performance. NIQE and BRISQUE are reported on the real image collection provided by SwinIR (Liang et al., 2021). DIV2K-I and DIV2K-II performance reported as PSNR.

Method	NIQE (↓)	BRISQUE (↓)	DIV2K-I	DIV2K-II
Bicubic	7.65	58.29	26.30	25.71
SAFMN	7.19	51.39	26.80	26.77
SeemoRe-T	**6.53**	**45.53**	**27.07**	**27.01**

perts to 8 as shown in Table 14 the impact on the overall model performance. The results indicate that increasing the number of experts adds complexity, however it doesn't consistently improve the reconstruction fidelity. Balancing the low-rank space and the expert count offers to fine-tune the performance trade-off. Though, our emphasis here is on efficiency, we aim to explore more complex designs in future research.

B.2. Evaluation on Real SR

We conduct experiments for Real SR ($\times 4$) using the Real-ESRGAN (Wang et al., 2021) degradation model on SeemoRe-T and the current efficient SOTA SR model SAFMN (Sun et al., 2023), see Table 15. Both SAMFN and SeemoRe-T are initialized from the $\times 4$ bicubic checkpoints, we reduce the number of iterations on the DF2K_OST dataset by half (250k) and train only using the L1 loss. We report the popular NR-IQA metrics (NIQE and BRISQUE) on the commonly used real-world image collection given in SwinIR (Liang et al., 2021). Additionally, we conduct a cross-dataset evaluation using testsets with more realistic degradation of different severity levels (Type I and Type II), as provided by (Liang et al., 2022).

C. Future work and limitations

The proposed approach, employing a mixture of experts for feature modulation, is versatile for tasks with limited input information, such as low-light enhancement and denoising. Additionally, SeemoRe's efficient design makes it a valuable solution for dynamic and resource-intensive environments. Expanding the number of experts in our network's low-rank aspect poses challenges due to rapid feature dimensionality growth. Thus, our approach is currently limited to a small number of experts, contrasting with other fields leveraging larger expert ensembles. Despite the improving trade-off between efficiency and reconstruction fidelity, as depicted in Figures 7 and 8, our SeemoRe model still contends with blur artifacts. However, similar artifacts can also be observed in Transformer-based super-resolution alternatives, albeit at a higher computational cost (in terms of inference time). While our model represents a pioneering effort in utilizing a mixture of low-rank experts for super-resolution, significant opportunities for further research exist. For instance, exploring explicit constraints on the features learned by different experts presents intriguing research directions with potential applications across a spectrum of restoration problems. We wish our network serve as a straightforward yet effective baseline, stimulating continued exploration in the field.

D. Visual Results.

We provide additional visual comparisons ($\times 4$) in Figure 7 for the Manga109 benchmark and in Figure 8 for the Urban100 benchmark. Our SeemoRe framework consistently produces visually pleasing results, even on artistic images. In contrast to previous methods which exhibit flawed texture and character reconstruction, our proposed approach effectively reconstructs missing details, as illustrated in Figure 7, across all exemplary images considered. More concretely, when examining the example image *img25* our SeemoRe network proficiently reconstructs the capital letter "I" within the text prompt "COMIC," whereas SwinIR-Light and DAT-L encounter difficulty in producing any readable output. Additionally, in example image *img04* our model significantly outperforms others in reconstructing the pattern with higher fidelity. Moreover, our model's reconstruction of *img92* in Figure 8 demonstrates reduced blurring and more distinct edges, enhancing overall visibility.

Figure 7. Visual comparison of SeemoRe with state-of-the-art methods on challenging cases for ×4 SR from the Manga109 benchmark.

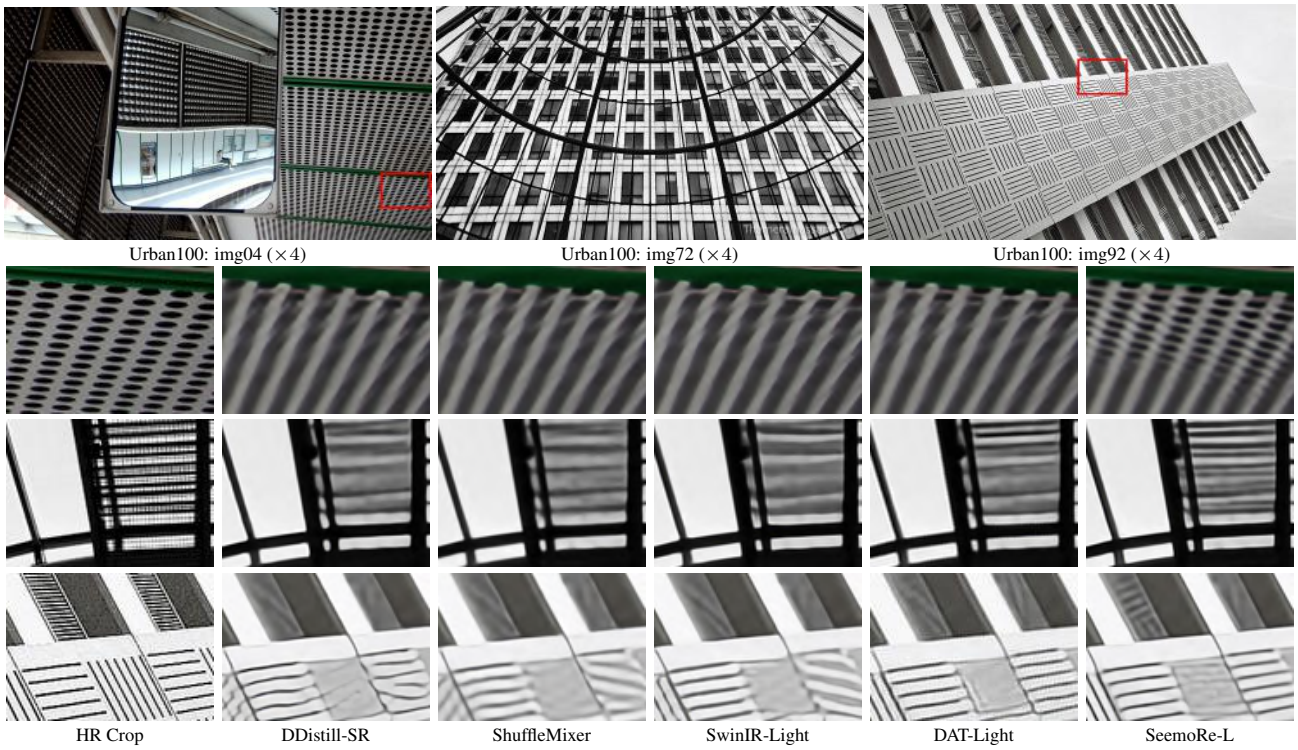

Figure 8. Visual comparison of SeemoRe with state-of-the-art methods on challenging cases for ×4 SR from the Urban100 benchmark.