Progressive Focused Transformer for Single Image Super-Resolution

Supplementary Material

In this supplementary material, we provide additional details on model training, inference time efficiency comparisons, and more comprehensive visual results. Specifically, in Section A, we present the training details for the PFT and PFT-light models. Subsequently, in Section B, we compare the inference time efficiency of different models. Finally, in Section C, we provide more detailed visualizations of the model's results.

A. Training Details

For training the PFT model, we use the DF2K dataset, which combines DIV2K [32] and Flickr2K [23], as our training set. To ensure fair comparisons, we adopt the same training configurations as those employed in recent super-resolution (SR) studies [6, 31, 41]. Our model is optimized using the AdamW optimizer with parameters set to $(\beta_1 = 0.9, \beta_2 = 0.99)$, a weight decay coefficient $\lambda = 0.0001$, and an initial learning rate of 2×10^{-4}. The $\times 2$ model is trained for 500K iterations. During training, the input patch size is fixed at 64×64, and a MultistepLR scheduler is applied to halve the learning rate at predefined iterations $[250000, 400000, 450000, 475000]$. The batch size is set to 32 for all training processes. To enhance robustness, the training data is augmented with random horizontal and vertical flips as well as random rotations of $90°$. For the $\times 3$ and $\times 4$ models, we apply fine-tuning based on the pre-trained $\times 2$ model to save time, training these models for only 250K iterations. The initial learning rate is set to 2×10^{-4}, and a MultistepLR scheduler is used to halve the learning rate at predefined iterations $[100000, 150000, 200000, 225000, 240000]$. We evaluate our method on five standard benchmark datasets: Set5 [2], Set14 [38], BSD100 [26], Urban100 [16], and Manga109 [27]. Additionally, the computational cost of all models presented in this paper is measured at an output resolution of 1280×640. For training the PFT-light model, only the DIV2K dataset is used, excluding Flickr2K. The initial learning rate for training $\times 2$ SR is set to 5×10^{-4}. All other training strategies remain consistent with those used for the PFT model.

B. Comparison of inference time

We compare the inference time of our PFT model with several state-of-the-art SR methods, including HAT [6], IPG [31], and ATD [41]. In this experiment, the inference time for all models is measured on a single NVIDIA GeForce RTX 4090 GPU at an output resolution of 512×512. As shown in Tab. 6, the inference time of our PFT model is comparable to existing methods. At the $\times 2$ and $\times 3$ scales, our model takes more time than HAT and ATD but is significantly faster than IPG. At the $\times 4$ scale, PFT outperforms both ATD and IPG in terms of inference speed. This improvement can be attributed to the efficient SMM CUDA kernels we developed to accelerate sparse matrix multiplication. Notably, despite the minor differences in inference time, our PFT model achieves lower computational complexity and delivers the best reconstruction performance.

Scale	Method	Params	FLOPs	Inference time
× 2	HAT [6]	20.6M	5.81T	1078ms
	ATD [41]	20.1M	6.07T	1394ms
	IPG [31]	18.1M	5.35T	2320ms
	PFT (Ours)	19.6M	5.03T	1594ms
× 3	HAT [6]	20.8M	2.58T	799ms
	ATD [41]	20.3M	2.69T	1038ms
	IPG [31]	18.3M	2.39T	1651ms
	PFT (Ours)	19.8M	2.23T	1158ms
× 4	HAT [6]	20.8M	1.45T	725ms
	ATD [41]	20.3M	1.52T	867ms
	IPG [31]	17.0M	1.30T	1060ms
	PFT (Ours)	19.8M	1.26T	852ms

Table 6. Inference efficiency comparison of models.

C. More Visual Examples.

C.1. Visual of attention distributions.

The visualization of attention distributions across different layers of the PFT-light model is shown in Fig. 5. As the network deepens, the PFA module progressively filters out tokens irrelevant to the current query and concentrates attention on the most critical regions. This mechanism not only reduces the influence of irrelevant features on reconstruction performance but also lowers computational costs, enabling the model to perform feature interactions over a larger spatial scope.

C.2. Visual Comparisons of PFT-light.

To qualitatively evaluate the reconstruction performance of our PFT and PFT-light models in comparison with other methods, we provide visual examples in Fig. 6, Fig. 7, and Fig. 8. These comparisons emphasize the strengths of our approach in restoring sharp edges and fine textures from severely degraded low-resolution inputs. The PFT-light model, in particular, excels at capturing edge details. Its selective focus on critical regions allows it to produce cleaner edges and achieve more accurate and visually reasonable reconstructions.

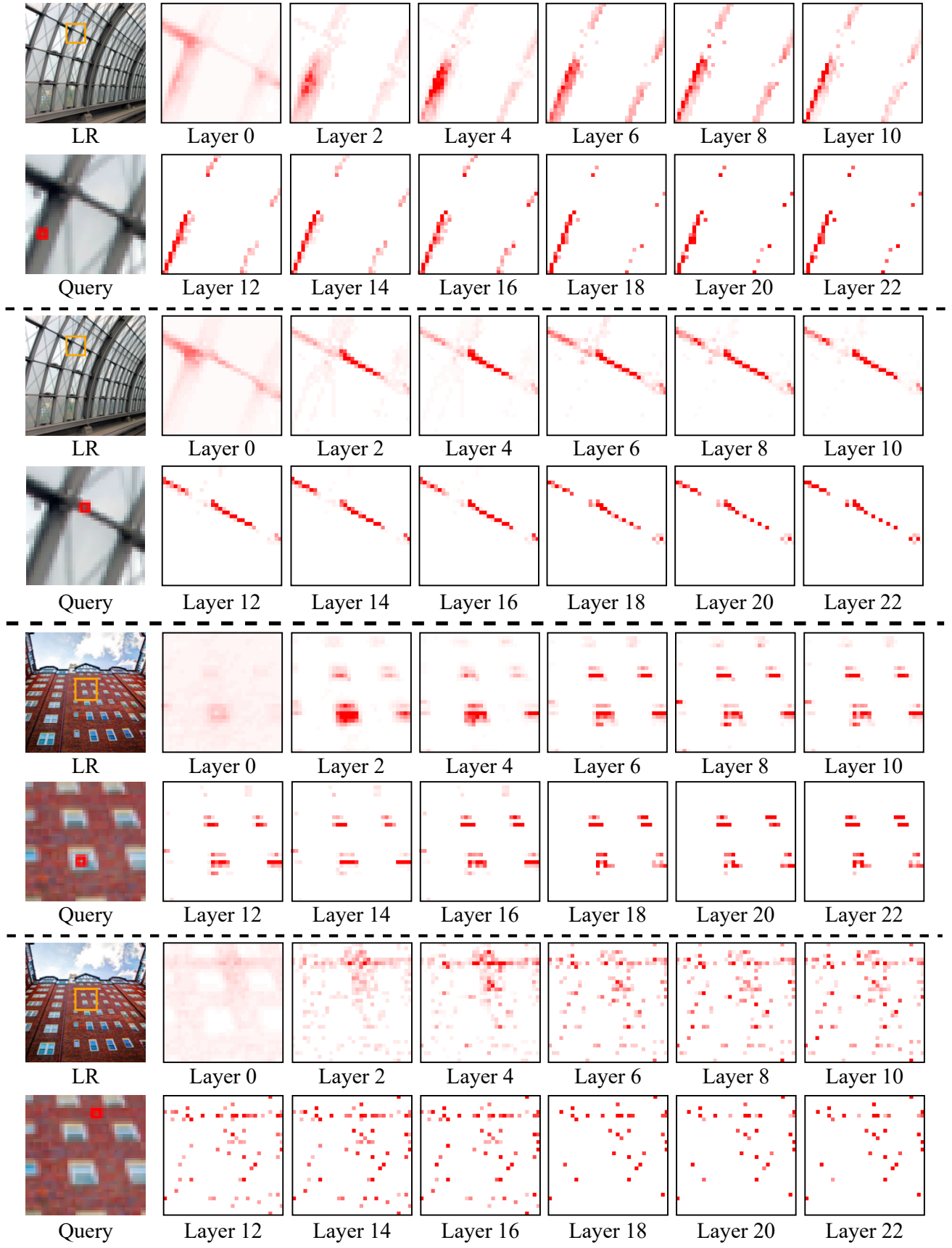

Figure 5. The visualization of attention distributions across different layers of the PFT-light model demonstrates the progressive filtering capability of the PFA module.

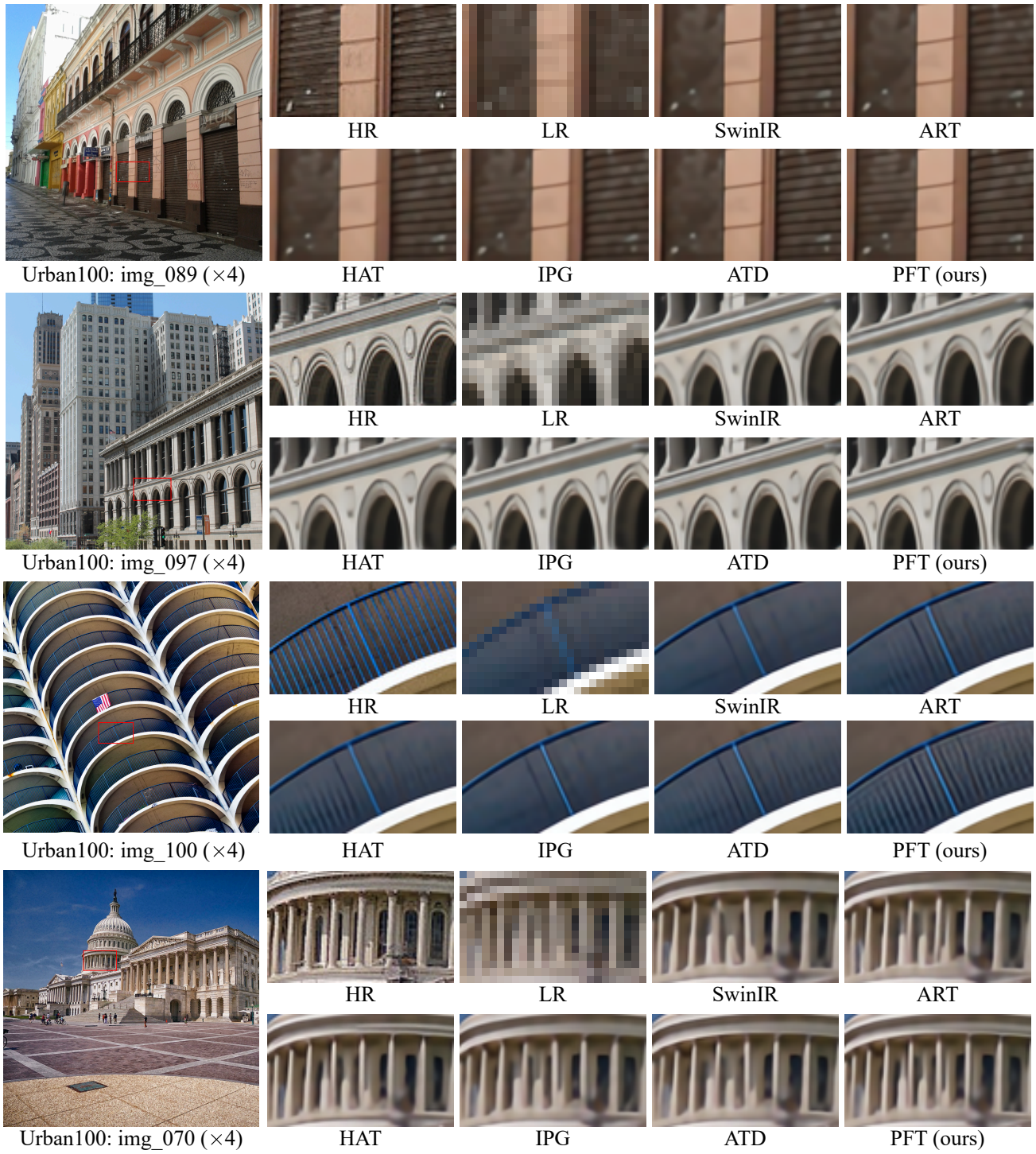

Figure 6. Visual comparison of classical SR reconstruction results.

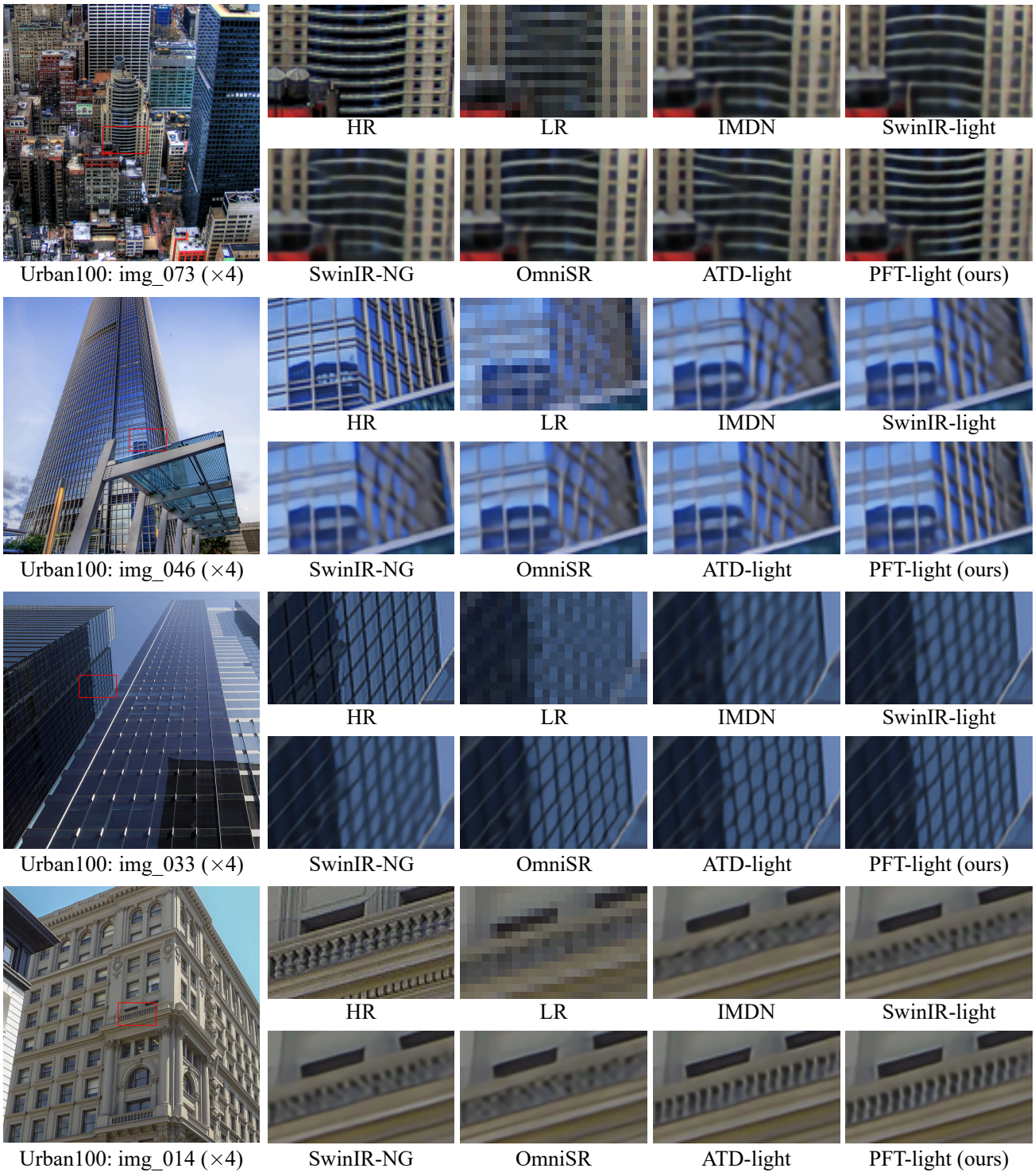

Figure 7. Visual comparison of lightweight SR reconstruction results.

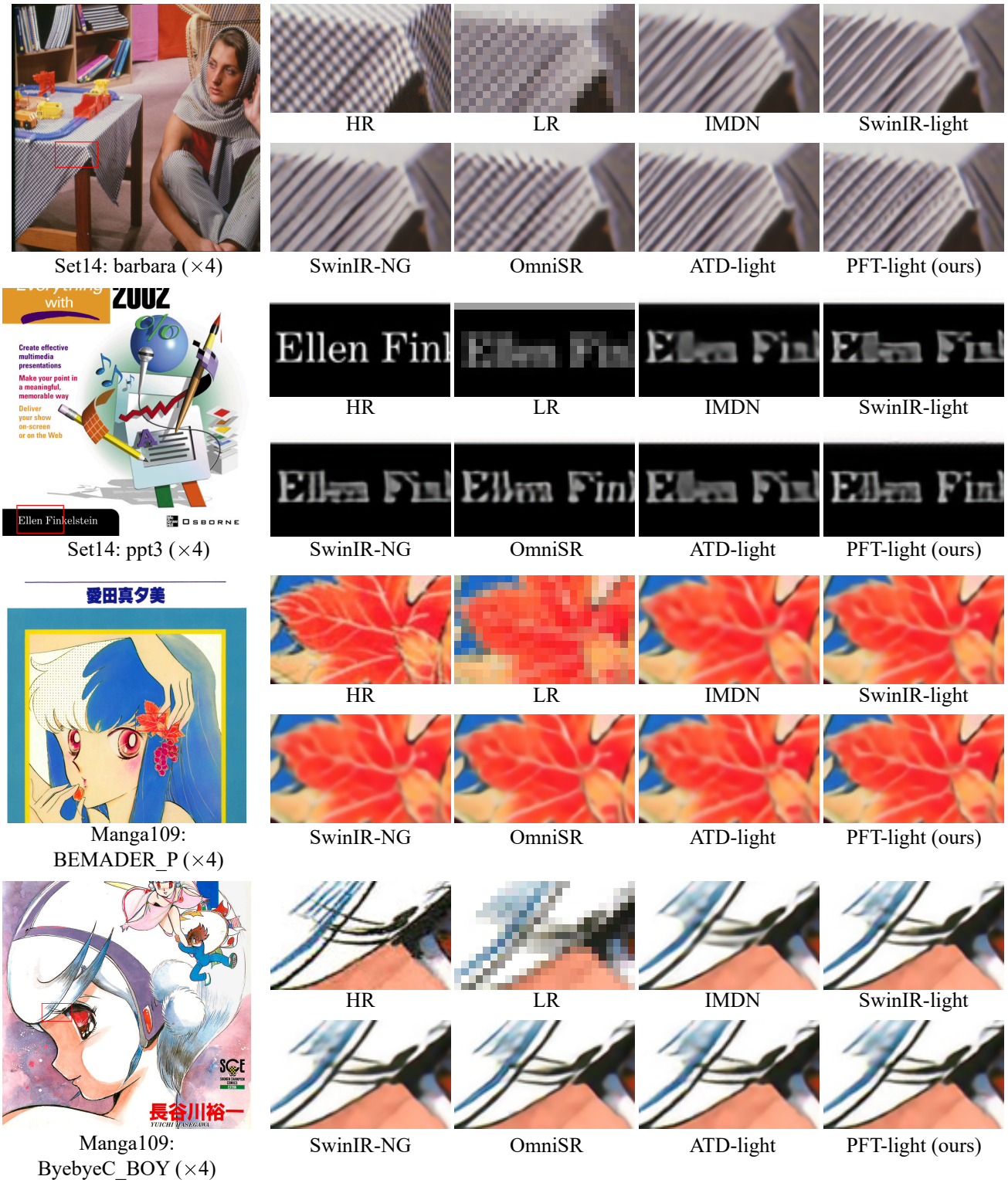

Figure 8. Visual comparison of lightweight SR reconstruction results.