Supplementary Material: Rethinking Image Super Resolution from Long-Tailed Distribution Learning Perspective

Yuanbiao Gou[1], Peng Hu[1], Jiancheng Lv[1], Hongyuan Zhu[2], Xi Peng[1*]
[1] College of Computer Science, Sichuan University, China
[2] Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
{gouyuanbiao, penghu.ml, hongyuanzhu.cn, pengx.gm}@gmail.com; lvjiancheng@scu.edu.cn

1. Experiments

In this supplementary material, we present more experimental results on our FPL including effectiveness on the twin fitting problem, comparisons on different loss functions, evaluations on different color spaces, influences of batch size, convergence curves of training, quantitative results of ablation studies and parameter analyses, and additional qualitative results. Moreover, we present the broader impact statement in the end.

1.1. Analysis Experiments

In this section, we perform more analysis experiments to further prove the effectiveness of our FPL.

Effectiveness on the Twin Fitting Problem. We illustrate the absolute differences between I^{HR}, I^{SR}_{FSRCNN} and I^{SR}_{+FPL} in Fig. 1(b)-(d), which show that FPL significantly improves the fitting degree of high-frequency regions. Besides, Fig. 1(e) shows that "+FPL" obtains more pixels with small differences and fewer pixels with large differences. In other words, SR model with FPL super-resolves the pixels in the high-frequency regions better than without FPL, which verifies the effectiveness of FPL on alleviating the twin fitting problem.

Comparisons on Different Loss Functions. Although FPL looks like the high-order mean error, *e.g.*, mean square or cubic error (MSE/MCE), they are remarkably different in the following aspects. First, FPL is specifically established long-tailed distribution learning solution for SR, while the high-order mean errors cannot effectively learn from the long-tailed pixel distribution, and thus suffering from the twin fitting problem. Second, FPL reformulates mean absolute error (MAE) by re-weighting the pixel-wise contribution and blocking the gradients in the weights. In other words, FPL is essentially a MAE with dynamic weights, and thus enjoys the advantages of MAE to SR tasks. Tab. 1 shows the quantitative results of different loss functions, from which one could see that MSE/MCE generally obtain worse performance than MAE, while FPL significantly outperforms it. The results demonstrate FPL could learn better from the long-tailed pixel distribution so that the twin fitting problem be alleviated.

Table 1. Comparisons of different loss functions w.r.t. FSRCNN on 4x SR task.

Dataset	Set5	Set14	BSD100	Urban100
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
MAE	28.71/0.8500	25.84/0.7389	25.58/0.7122	22.99/0.7184
MSE	28.74/0.8471	25.86/0.7371	25.58/0.7112	23.00/0.7148
MCE	28.54/0.8373	25.73/0.7315	25.48/0.7066	22.89/0.7054
FPL	29.00/0.8565	26.01/0.7446	25.70/0.7178	23.26/0.7313

Evaluations on Different Color Spaces. As many existing SR models are evaluated on YCbCr color space in their official implementations, we conduct this experiments to demonstrate our implementations on RGB color space. Here, we take CARN as an example and show the results in Tab. 2. From the table, one could see that our implementation obtains the similar results as its official implementation which uses more training datasets. As all the SR models have almost the same training and testing settings, our implementations of them are credible. Moreover, although the quantitative results are different in YCbCr and RGB color spaces, the performance gains are almost consistent.

Influences of Batch Size. To investigate the influence of batch size on FPL, we train FSRCNN+FPL with the batch size of 8, 16, 32, 64, and test it on the Set5. The quantitative results are shown in Tab. 3, from which one could see that a small batch size (*e.g.*, 8) degrades the performance, while increasing the batch size to 16 or larger size, the model obtains consistent and significant performance improvements. The possible reason is that a small batch size cannot ensure the significant attendance of the high-frequency regions, because the training patches are cropped from the 2K-resolution images which involve a high proportion of low-frequency regions.

*Corresponding author

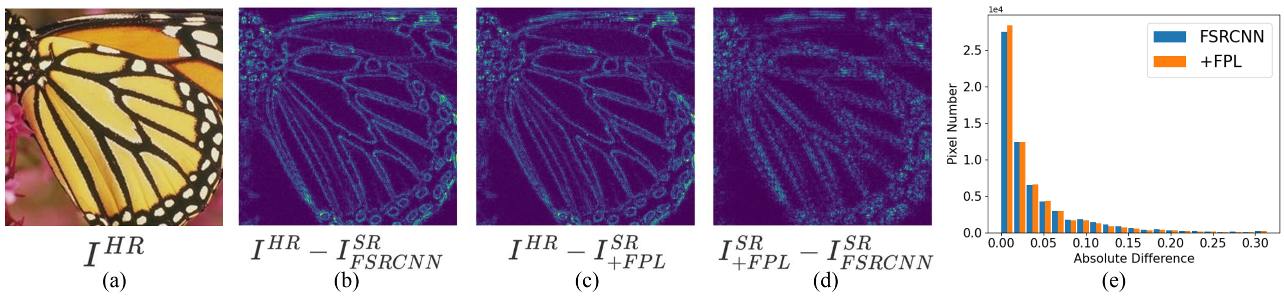

Figure 1. The differences between HR and SR images, where I^{HR} denotes HR image, I^{SR}_{FSRCNN} and I^{SR}_{+FPL} denote SR images from FSRCNN and FSRCNN+FPL, respectively. (e) shows the pixel number of (b) and (c) in the different difference intervals.

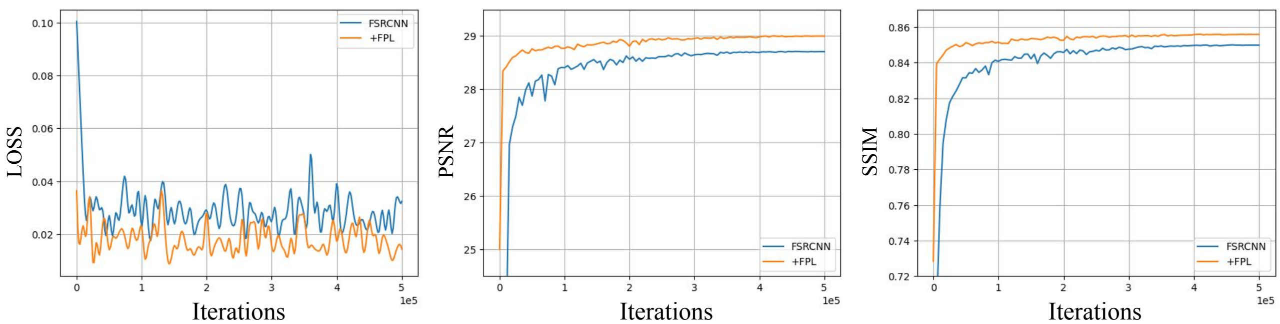

Figure 2. The convergence curves of FSRCNN and FSRCNN+FPL, in which the PSNR and SSIM values are calculated on the Set5.

Table 2. Comparisons of different color spaces w.r.t. CARN on 4x SR task. Taking CARN as an example since the overlaps in testing datasets between its official and our implementations. Note that the training datasets in the official implementation are DIV2K, 291 image set, and BSD500, while that is only DIV2K in ours.

Dataset	Set5	Set14	BSD100	Urban100
YCbCr Color Space				
CARN (Official)	31.92	28.42	27.44	25.62
CARN (Our)	31.85	28.41	27.42	25.61
CARN+FPL (Our)	32.02	28.48	27.51	25.87
Gains	+0.17	+0.07	+0.09	+0.26
RGB Color Space				
CARN (Our)	29.94	26.61	26.07	24.12
CARN+FPL (Our)	30.11	26.69	26.16	24.36
Gains	+0.17	+0.08	+0.09	+0.24

Table 3. Quantitative results of different batch size.

Batch Size	8	16	32	64
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
FSRCNN	28.66/0.8482	28.71/0.8500	28.65/0.8464	28.67/0.8485
+FPL	28.48/0.8422	29.00/0.8565	28.98/0.8559	29.00/0.8565

Convergence curves of training. To investigate training process, we illustrate the convergence curves of FSRCNN and FSRCNN+FPL in Fig. 2. From the figures, one could observe that i) both FSRCNN and FSRCNN+FPL are sufficiently convergent; ii) comparing with FSRCNN, FSRCNN+FPL converges faster and obtains better results; iii) both PSNR and SSIM curves of FSRCNN+FPL are more stable (*i.e.*, less oscillation) than those of FSRCNN. Based on the observations, we could conclude that FPL speeds up and stabilizes the convergence process while improving performance during the training.

Quantitative results of ablation studies and parameter analyses. In addition to the illustrations in the main body of the paper, we present more quantitative results in Tabs. 4 and 5. To be specific, Tab. 4 shows the quantitative results of changing γ_{sp}, γ_{lp}, where $\gamma_{sp} = 0$ or $\gamma_{lp} = 0$ corresponds to remove one weighting function, and $\gamma_{sp} = 0, \gamma_{lp} = 0$ simultaneously disables the two weighting functions and only BI is remained to under-sample the pixels. These results demonstrate the indispensable roles of the two structure priors and BI under-sampling. Moreover, Tab. 5 shows the quantitative results of changing α_{sp}, α_{lp}, where are no very obvious performance changes among the different parameter settings.

1.2. Qualitative Results.

In addition to the qualitative results from BSD100 in the main body of the paper, we present more results on the

Table 4. Quantitative results of changing γ_{sp}, γ_{lp}.

Dataset	Set5		Manga109	
	PSNR	SSIM	PSNR	SSIM
L1	28.71	0.8500	25.78	0.8346
$\gamma_{sp} = 0, \gamma_{lp} = 0$	28.91	0.8543	26.18	0.8420
$\gamma_{sp} = 0, \gamma_{lp} = 1$	28.94	0.8551	26.25	0.8444
$\gamma_{sp} = 0, \gamma_{lp} = 2$	28.94	0.8553	26.32	0.8454
$\gamma_{sp} = 1, \gamma_{lp} = 0$	28.98	0.8561	26.35	0.8462
$\gamma_{sp} = 1, \gamma_{lp} = 1$	29.00	0.8565	26.41	0.8473
$\gamma_{sp} = 1, \gamma_{lp} = 2$	28.96	0.8556	26.33	0.8441
$\gamma_{sp} = 2, \gamma_{lp} = 0$	28.89	0.8552	26.29	0.8442
$\gamma_{sp} = 2, \gamma_{lp} = 1$	28.91	0.8552	26.33	0.8451
$\gamma_{sp} = 2, \gamma_{lp} = 2$	28.94	0.8547	26.31	0.8431

Table 5. Quantitative results of changing α_{sp}, α_{lp}.

Dataset	Set5		Manga109	
	PSNR	SSIM	PSNR	SSIM
L1	28.71	0.8500	25.78	0.8346
$\alpha_{sp} = 0.1, \alpha_{lp} = 1.0$	28.96	0.8556	26.33	0.8454
$\alpha_{sp} = 0.5, \alpha_{lp} = 1.0$	29.00	0.8565	26.41	0.8473
$\alpha_{sp} = 1.0, \alpha_{lp} = 1.0$	28.98	0.8552	26.33	0.8453
$\alpha_{sp} = 2.0, \alpha_{lp} = 1.0$	28.96	0.8556	26.39	0.8468
$\alpha_{sp} = 0.5, \alpha_{lp} = 0.1$	28.91	0.8549	26.31	0.8453
$\alpha_{sp} = 0.5, \alpha_{lp} = 0.5$	28.95	0.8557	26.39	0.8468
$\alpha_{sp} = 0.5, \alpha_{lp} = 2.0$	28.96	0.8557	26.34	0.8452

other datasets, *i.e.*, Set5 (Fig. 7), Set14 (Fig. 6), Manga109 (Fig. 4), Urban100 (Fig. 3) and Test2K (Fig. 5). From the figures, one could see that, compared with the original SR models, FPL enables them to produce more faithful image structures and clearer image details, while obtaining higher PSNR and SSIM values. These results demonstrate the superiority of FPL in learning from the long-tailed pixel distribution so that the twin fitting problem be alleviated. Some areas are highlighted by color rectangles, and zooming in is recommended for better visualization.

1.3. Broader Impact Statement

This work reveals and solves the twin fitting problem in SR tasks caused by the long-tailed pixel distribution in natural images. There are many benefits to solve the twin fitting problem, *e.g.*, obtaining fine image details, speeding up models' convergence, and achieving performance gains. Despite the benefits, it should pay attention to the potential negative impacts including but not limited to i) as an image enhancement technology, it has the potential of prejudicing the rights of others with improper use; ii) although FPL enjoys high interpretability, it doesn't change the black-box nature of deep SR models and result in tremendous security risks when applied to some critical fields such as autopilot and medical; iii) a lot of energy will be consumed to train the SR model, and thus causing massive CO2 emissions.

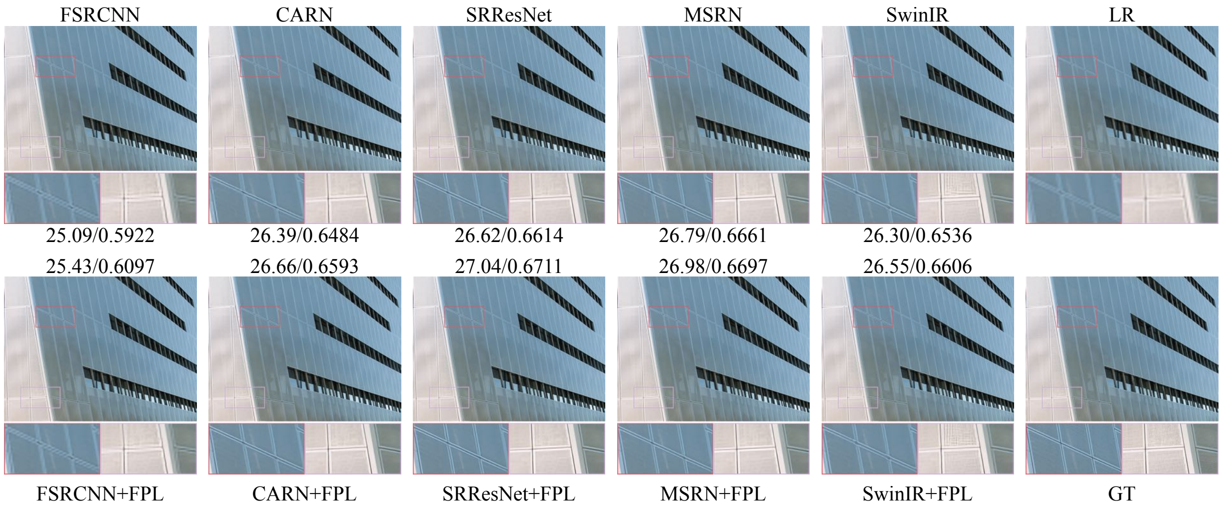

Figure 3. Qualitative comparisons on Urban100 dataset for $4\times$ SR. The top row is the results of the models trained through L1 loss and the LR image, while the bottom row is the results of the models trained through FPL and the GT image.

Figure 4. Qualitative comparisons on Manga109 dataset for $4\times$ SR. The top row is the results of the models trained through L1 loss and the LR image, while the bottom row is the results of the models trained through FPL and the GT image.

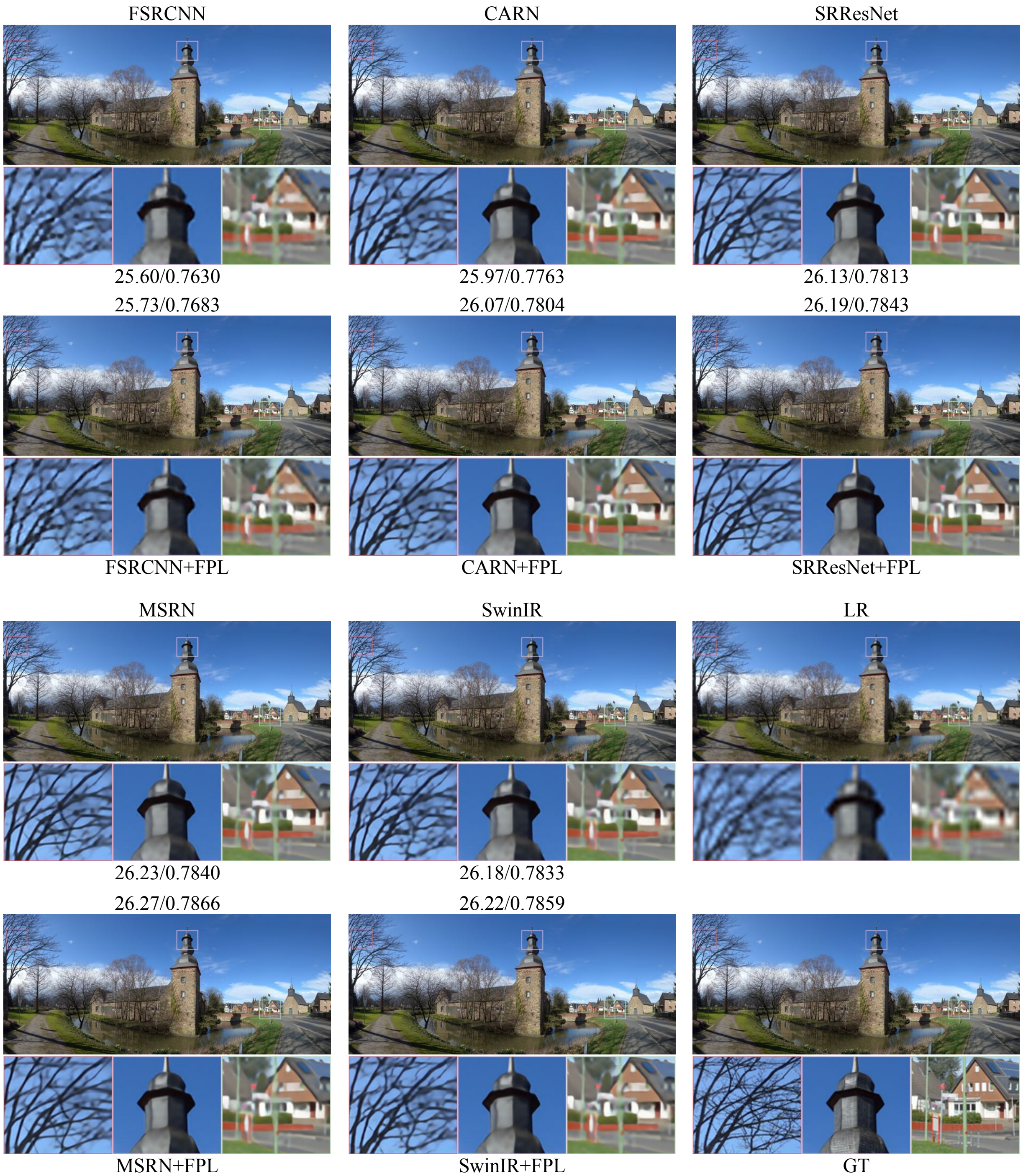

Figure 5. Qualitative comparisons on Test2K dataset for $4\times$ SR. The first and third rows are the results of the models trained through L1 loss and the LR image, while the second and fourth rows are the results of the models trained through FPL and the GT image.

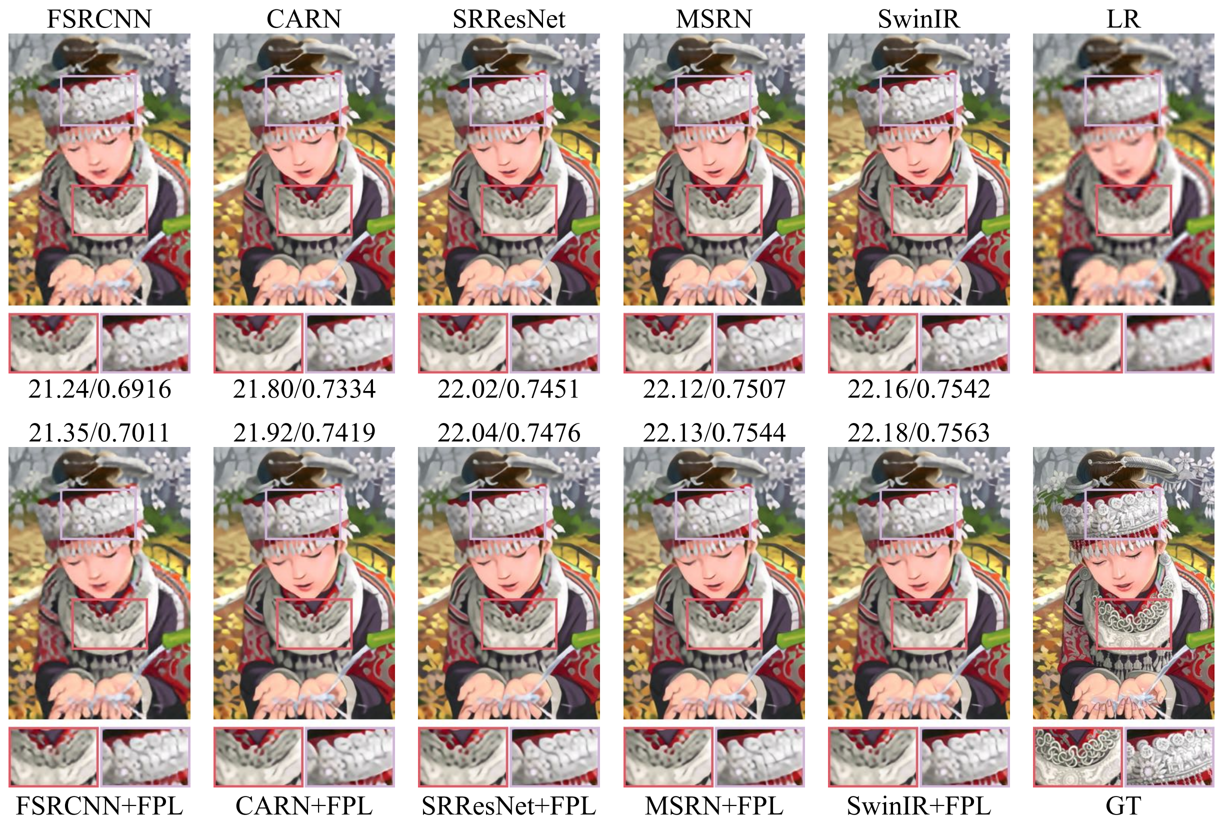

Figure 6. Qualitative comparisons on Set14 dataset for $4\times$ SR. The top row is the results of the models trained through L1 loss and the LR image, while the bottom row is the results of the models trained through FPL and the GT image.

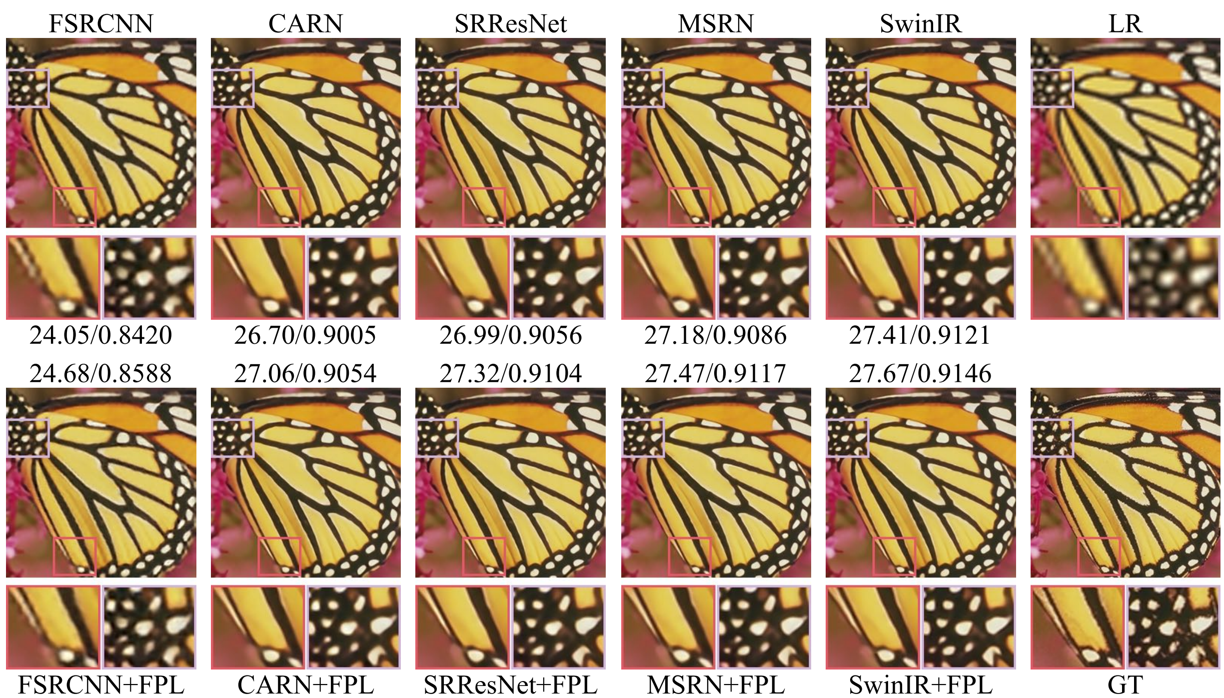

Figure 7. Qualitative comparisons on Set5 dataset for $4\times$ SR. The top row is the results of the models trained through L1 loss and the LR image, while the bottom row is the results of the models trained through FPL and the GT image.