Supplementary Materials to “Human Guided Ground-truth Generation for Realistic Image Super-resolution”

Du Chen[1*], Jie Liang[1,2*], Xindong Zhang[1,2], Ming Liu[1,3], Hui Zeng[2] and Lei Zhang[1,2†]
[1]The Hong Kong Polytechnic University, [2]OPPO Research Institute, [3]Harbin Institute of Technology
{csdud.chen, c-ming.liu}@connet.polyu.hk, {liang27jie, cshzeng}@gmail.com
{csxdzhang, cslzhang}@comp.polyu.edu.hk

In this supplementary file, we provide the following materials:

- The two sets of degradation parameters used in our image enhancement model training (please refer to Section 3.2 in the main paper);
- The interface of our annotation software (please refer to Section 3.3 in the main paper);
- More visual samples about the “Positive” and “Negative” GTs (please refer to Section 3.4 in the main paper);
- Examples of the negative and positive GTs and their residual variation maps for illustrating the design of training loss with both positive and negative pairs (please refer to Section 4 in the main paper).
- The degradation parameters used in synthesizing our training and testing data, and the visual samples of the synthesized low-quality images (please refer to Section 5.1 in the main paper).
- More visual examples in our experimental results (please refer to Section 5.2 and Section 5.3 in the main paper).
- Quantitative comparisons of models trained on positive only and both positive and negative pairs (please refer to Section 5.3 in the main paper).

Table 1. The two sets of degradation parameters used in our image enhancement model training. Setting 1 focuses more on processing slightly higher noise, and setting 2 focuses more on dealing with slightly stronger blurs. The differences between the two settings are highlighted in colors red and blue.

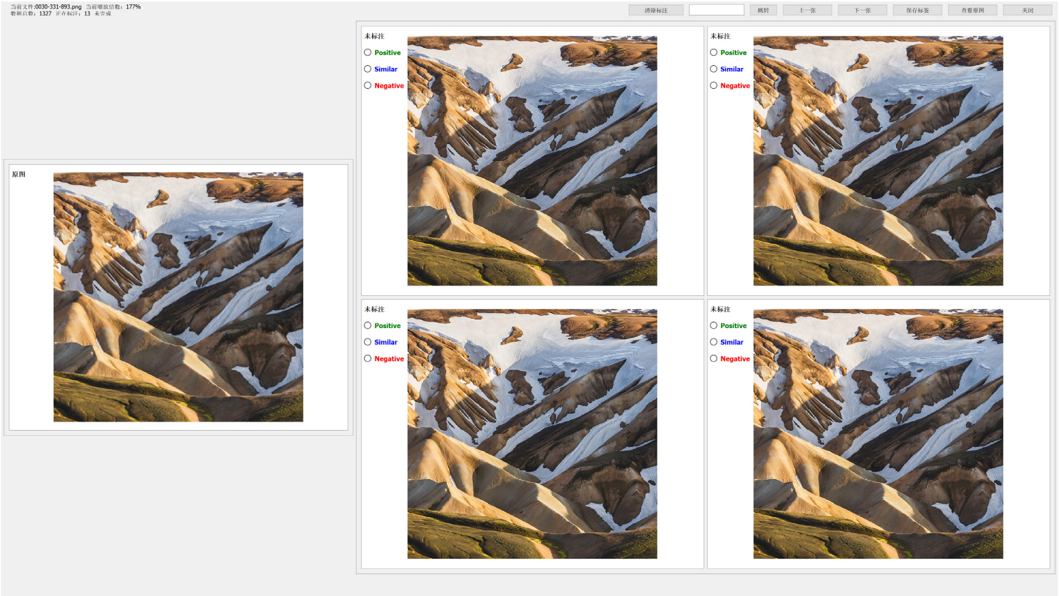
Operation	Parameter	Setting 1	Setting 2
Blur	Kernel size [2m+1]	m $\in$ [1, 3]	m $\in$ [1, 3]
	Kernel list	isotropic, an-isotropic	isotropic, an-isotropic
	Kernel list probability	0.7, 0.3	0.7, 0.3
	Sinc kernel probability	0.1	0.1
	Standard deviation θ	$\theta \in$ [0.1, 0.5]	$\theta \in$ [0.5, 1.0]
Resize	Resize list	downsample,same,upsample	downsample,same,upsample
	Resize list probability	0.85,0.05,0.1	0.85,0.05,0.1
	Resize range ϕ	$\phi \in$ [0.9, 1.1]	$\phi \in$ [0.9, 1.1]
	Resize mode	area, bilinear, bicubic	area, bilinear, bicubic
Noise	Noise list	Gaussian, Poisson	Gaussian, Poisson
	Noise list probability	0.5, 0.5	0.5, 0.5
	Sigma of Gaussian σ	$\sigma \in$ [6.5, 13]	$\sigma \in$ [0.5, 6.5]
	Scale of Poisson γ	$\gamma \in$ [0.45, 0.9]	$\gamma \in$ [0.05, 0.45]
	Gray noise probability	0.1	0.1
JPEG	Quality factor α	$\alpha \in$ [80, 95]	$\alpha \in$ [80, 95]

1. The Degradation Parameter Settings used in Image Enhancement Model Training

As mentioned in Section 3.2 of the main paper, considering the fact that the quality of HR images to be further enhanced is generally not bad, we deliberately control the degradation settings in Eq. (1) of the main paper to ensure that the quality of synthesized low-quality images is closed to the real-world situation. Here we employ two sets of degradation parameters, whose differences lie in the **blur kernel** and **noise level**. As can be seen in Tab. 1, the first degradation setting has weaker blur kernel and higher noise level (highlighted in red color), while the second setting has stronger blur kernel and lower noise level (highlighted in blue color).

2. Annotation Software

We developed an annotation software program for the volunteers to annotate the extracted patches. The interface of the software is shown in Fig. 1. The original HR patch is positioned on the left side of the screen, while the four enhanced versions are located on the right side in random order. Users could zoom in and out to observe the details between the original HR and the enhanced patches. Those patches whose perceptual quality is better than the original one are labeled as "Positive", and the patches with worse perceptual quality are labeled as "Negative". In case the quality of enhanced patch is tied with the original one, the user can label it as "Similar".



当前文件:0030-331-893.png 当前缩放倍数：177%
数据总数：1327 正在标注：13 未完成

Figure 1. The annotation software interface. Users could zoom in and out to observe the details between the original HR and the enhanced patches. Then they could press the "Positive", "Similar" or "Negative" button to label the enhanced patches.

3. Visual Examples of the Positive and Negative GTs

In Fig. 2, we show more examples of the annotated "Positive" and "Negative" GTs, as well as some "Similar" ones. As could be seen in the figures, "Positive" GTs have clearer and richer details, and less noise and artifacts than the original HR images, while "Negative" GTs often have over-sharpening/wrong details and more noise than the original HR images.

4. Examples of Positive and Negative GTs and Their Residual Variation Maps

In Section 4 of the main paper, we design a loss to employ the negative GTs to train the Real-ISR model together with the positive GTs. To this end, we build a residual variation map $M_{i,j}^{Neg}$ to detect the negative areas in a negative GT. Similarly, we build a residual variation map $M_{i,j}^{Pos}$ for a positive GT. At location (i, j), if the negative residual variation is higher than the positive one, we identify this pixel in I^{Neg} as a truly negative pixel, which should be used to update the model. Finally, we obtain an indication map $M_{i,j}^{Ind}$ to penalize the real negative pixels. Fig. 3 shows some negative and positive GTs and their residual variation maps.

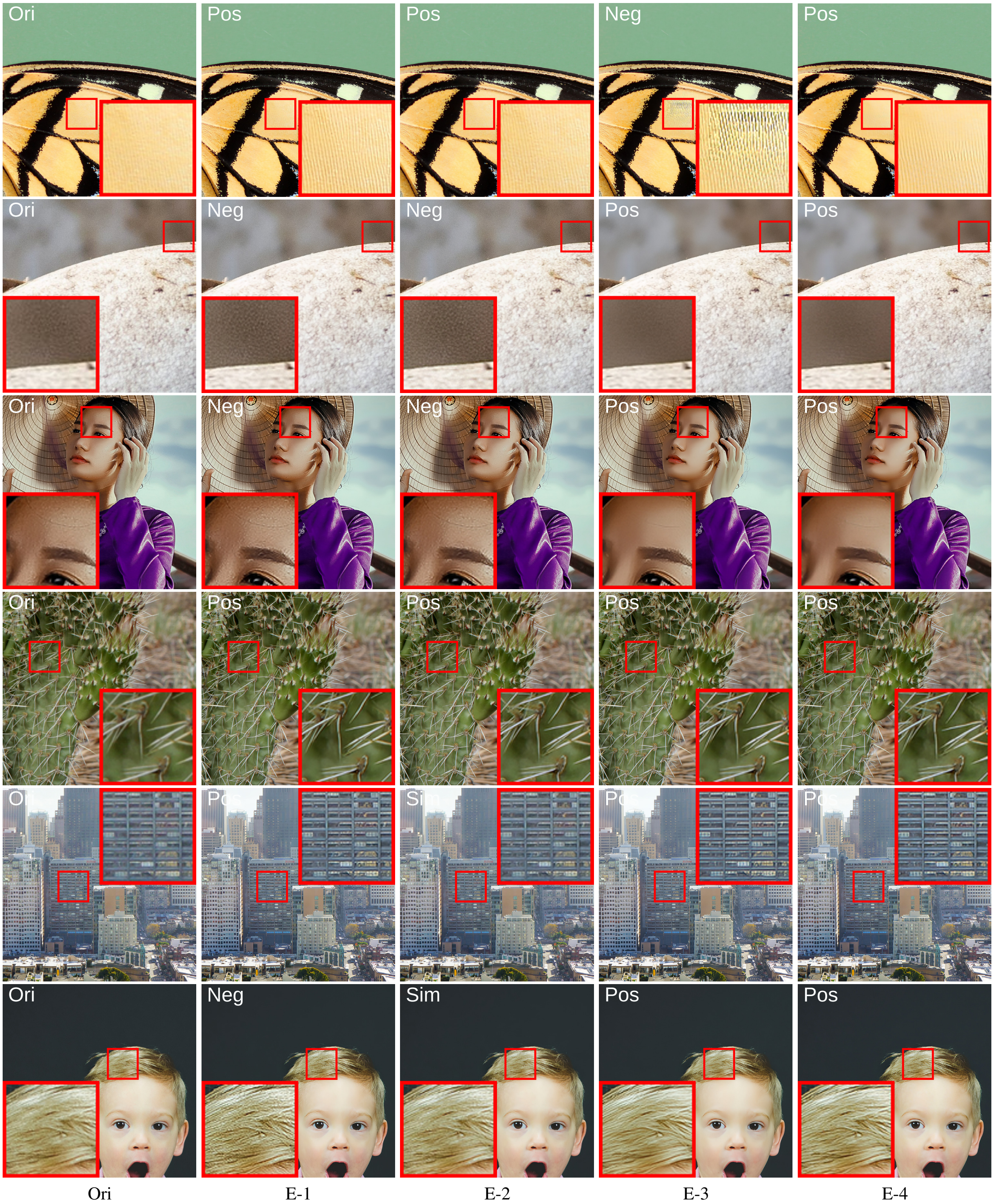


Figure 2. From left to right: the original HR image (Ori), and the enhanced images (E-1, E-2, E-3, E-4) with their annotations displayed at the top-left corner. “Pos”, “Sim” and “Neg” represent “Positive”, “Similar” and “Negative”, respectively. **Please zoom in for better observation**.

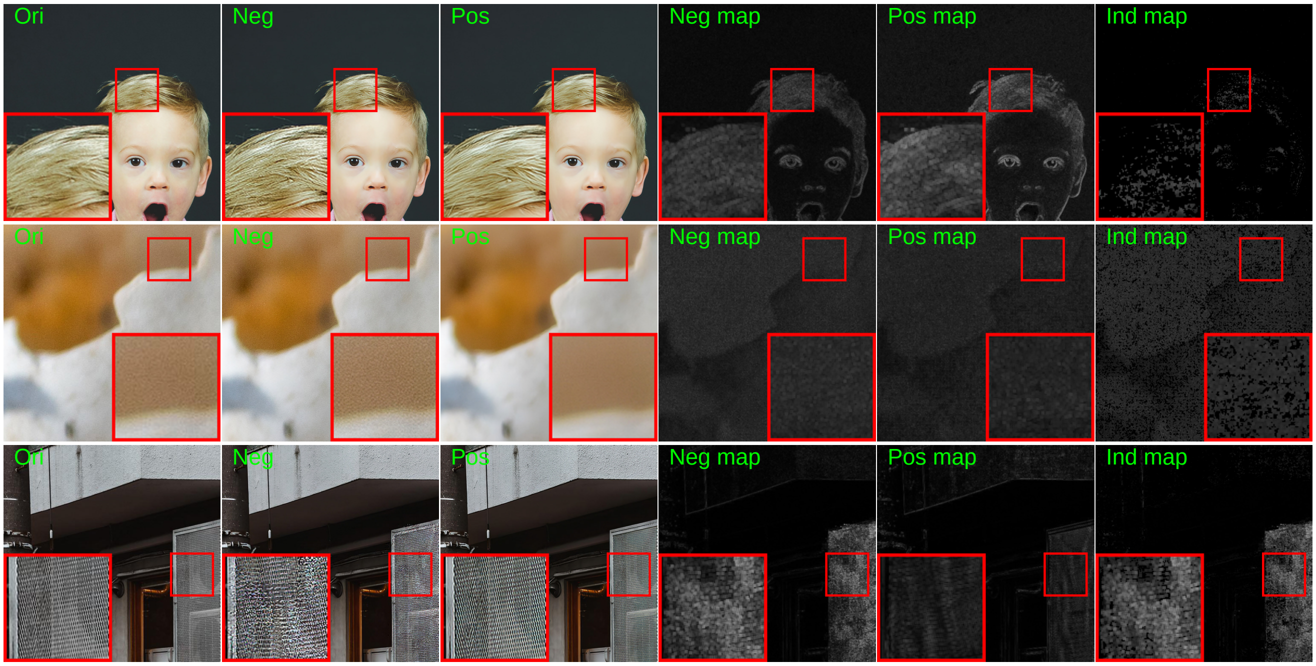


Figure 3. Some representative negative GTs and thier residual variation maps. The negative GT in the first row shows over-sharpening details, the one in row 2 shows strong noise, and the one in row 3 shows wrong details. Columns 4 and 5 visualize the residual variation maps M^{Neg} (for negative GT) and M^{Pos} (for positive GT), and column 6 shows the indication map M^{Ind}, which indicates the true negative pixels that could be used to update the Real-ISR model. **Please zoom in for better observation.**

5. The Degradation Parameters for Synthesizing Training and Testing Data

With the annotated dataset, we synthesize low-quality images using the degradation model in Eq. 1 of the main paper for Real-ISR model training and testing. The degradation parameters are listed in Tab. 2. Some examples of our synthesized low-quality images could be seen in Fig. 4.

Table 2. The degradation parameters used to synthesizing the LR images for Real-ISR model training and testing.

Operation	Parameter	Setting
Blur	Kernel size [2m+1] Kernel list Kernel list probability Sinc kernel probability Standard deviation θ	m $\in$ [1, 4] isotropic, an-isotropic 0.7, 0.3 0.01 $\theta \in [0.1, 1.0]$
Resize	Resize list Resize list probability Resize range ϕ Resize mode	downsample,same,upsample 0.85, 0.05, 0.1 $\phi \in [0.8, 1.2]$ area, bilinear, bicubic
Noise	Noise list Noise list probability Sigma of Gaussian σ Scale of Poisson γ Gray noise probability	Gaussian, Poisson 0.5, 0.5 $\sigma \in [1, 12]$ $\gamma \in [0.05, 0.8]$ 0.1
JPEG	Quality factor	[75, 95]

6. More Visual Comparisons in Our Experimental Results

In Section 5.2 of the main paper, we compared the Real-ISR models trained on DF2K-OST and the positive GTs in our proposed HGGT dataset. More visual comparisons are shown in Fig. 5 and Fig. 6. In Section 5.3 of the main paper, we validated the effectiveness of negative GTs. More visual comparisons are shown in Fig. 7.

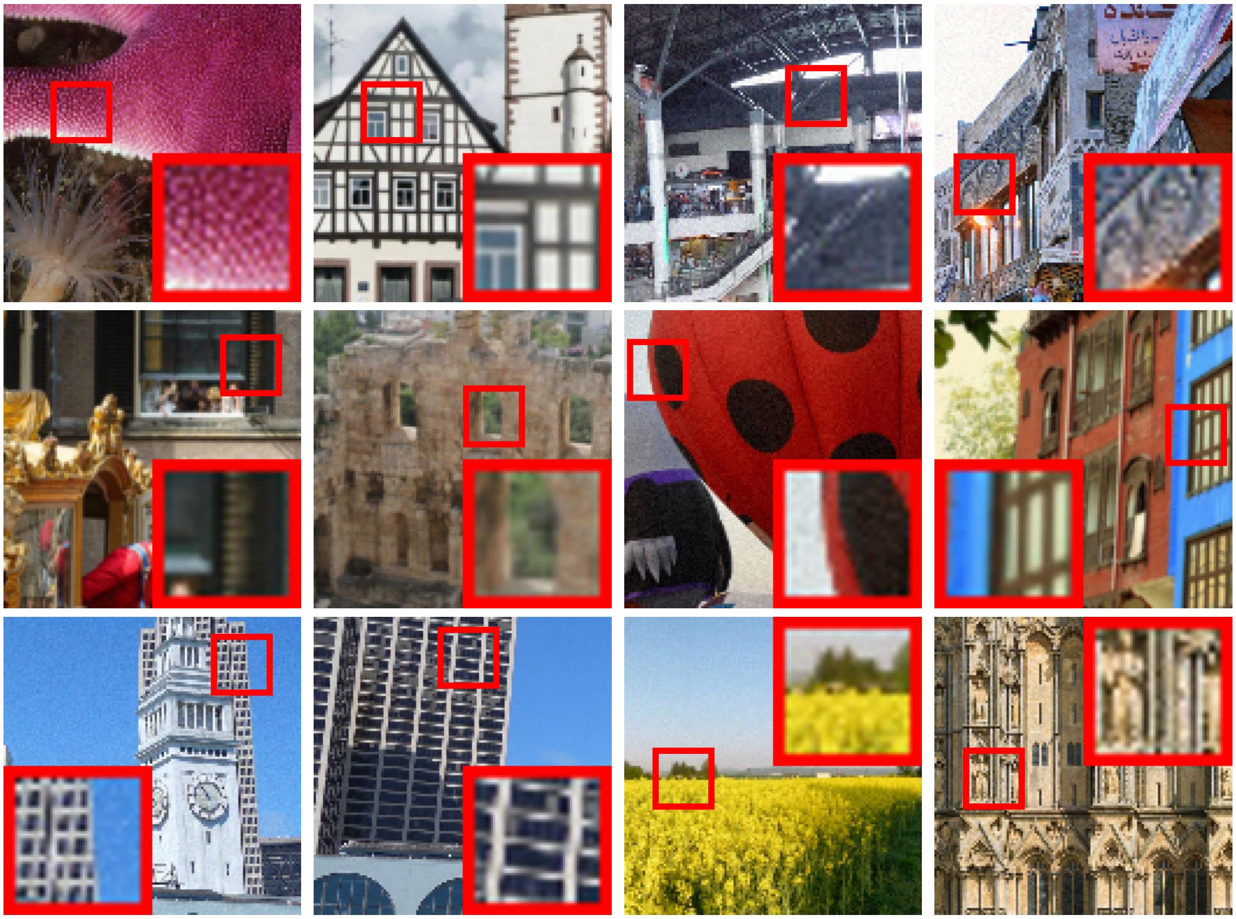


Figure 4. Some examples of our synthesized low-quality images.

7. Quantitative Comparisons of the Trained Models on Positive and Negative Pairs

As shown in Table 3, training with the original HR images leads to the worst LPIPS scores under both RRDB and SwinIR backbones, while training with our positive GTs demonstrates a clear improvement in LPIPS/DISTS (about 24.54%/23.03% and 25.64%/26.53% on RRDB and SwinIR, respectively) while sacrificing certain pixel-wise fidelity. This is consistent with the observations we made in Section 5.2 of the main paper. Introducing the negative GTs into training brings further performance improvement in LPIPS. It is reasonable that the improvement is not significant in numbers because most of the artifacts are of high-frequency and they occupy only a minority of pixels.

Table 3. The quantitative results of RRDB-GAN and SwinIR-GAN models trained on the original HR patches, positive GTs only, and both positive and negative GTs (Pos+Neg GT).

Method	Train Dataset	PSNR/SSIM/LPIPS/DISTS
RRDB-GAN	Original HR	22.6388/0.6454/0.2408/0.1676
	Positive GT	22.2378/0.6400/0.1817/0.1290
	Pos+Neg GT	22.2453/0.6405/0.1806/0.1290
SwinIR-GAN	Original HR	22.7147/0.6516/0.2274/0.1620
	Positive GT	22.3027/0.6474/0.1691/0.1227
	Pos+Neg GT	22.2733/0.6476/0.1688/0.1227

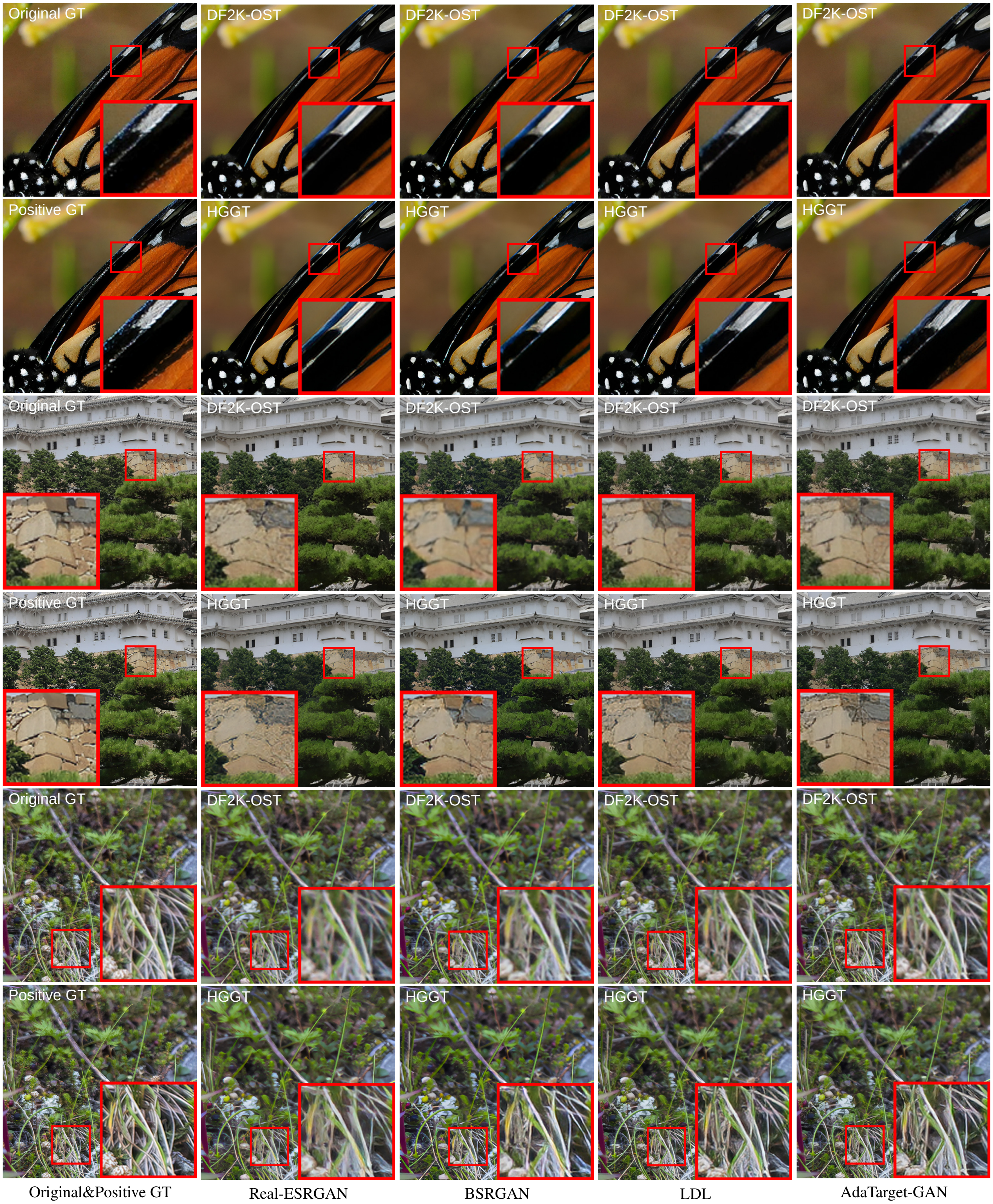


Figure 5. Visual comparison of state-of-the-art models trained on the DF2K-OST and our proposed HGGT datasets. The 1st, 3rd and 5th row show the results of models trained on DF2K-OST, while the 2nd, 4th and 6th row show the results of models trained on ours positive GTs. The left column shows the original GT and the positive GT in our dataset. **Please zoom in for better observation**.

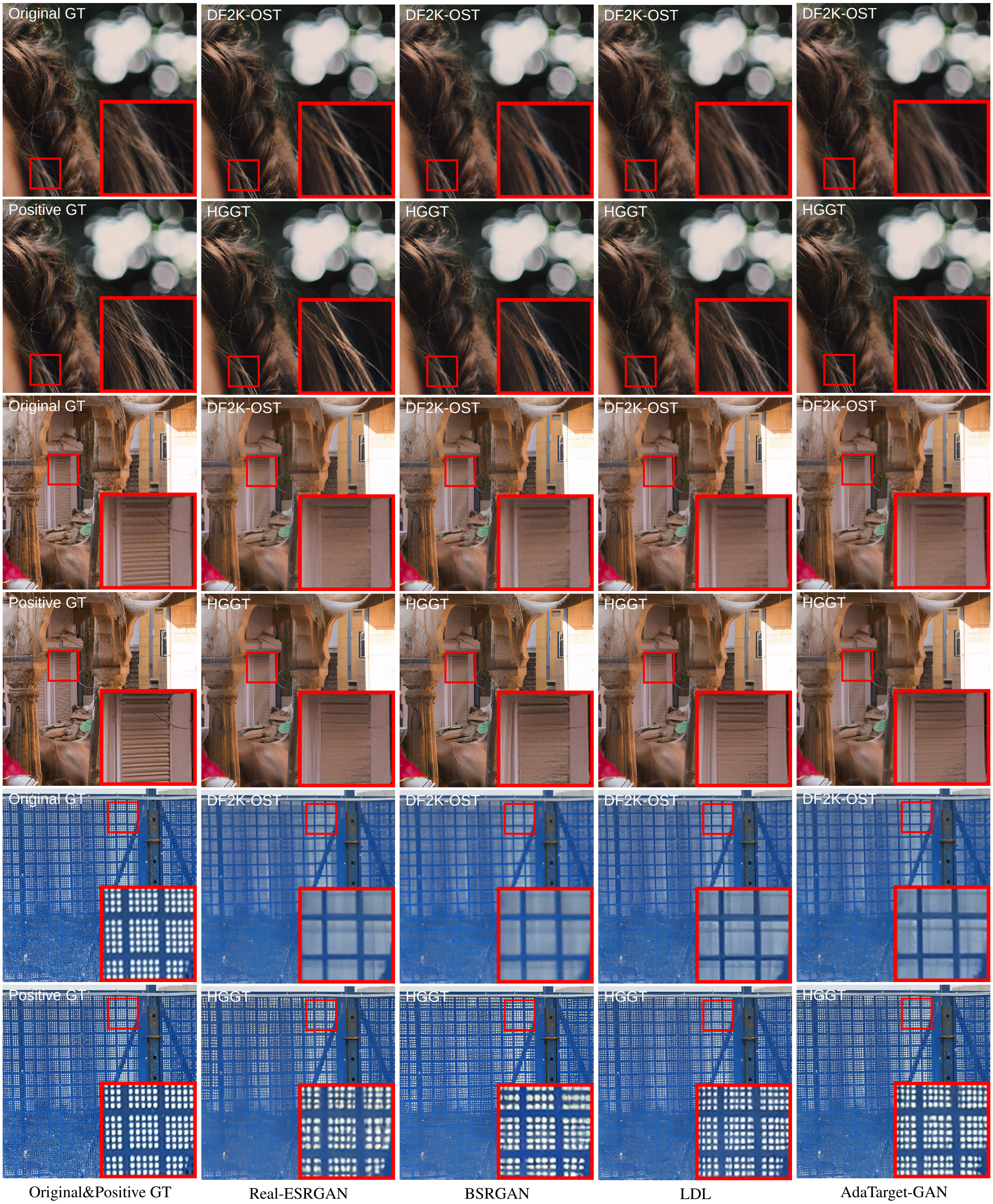


Figure 6. Visual comparison of state-of-the-art models trained on the DF2K-OST and our proposed HGGT datasets. The 1st, 3rd and 5th row show the results of models trained on DF2K-OST, while the 2nd, 4th and 6th row show the results of models trained on ours positive GTs. The left column shows the original GT and the positive GT in our dataset. **Please zoom in for better observation**.

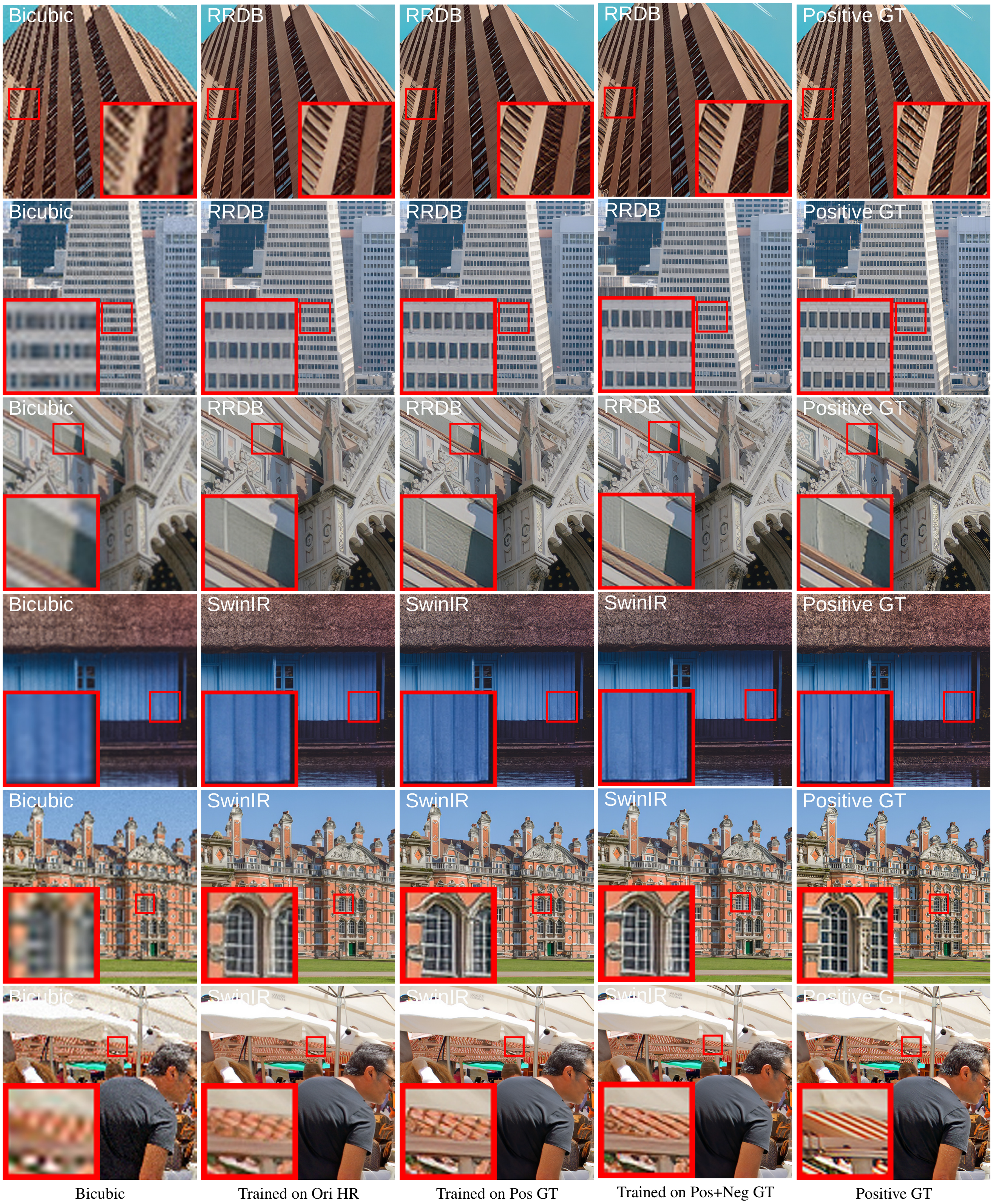


Figure 7. Visualizations of RRDB-GAN and SwinIR-GAN models trained on the original HR (Ori HR) patches, positive GTs (Pos GT) only, and both positive and negative GTs (Pos+Neg GT). The 1st, 2nd, 3rd rows show the results of RRDB-GAN, and the 4th, 5th, 6th rows show the results of SwinIR-GAN. From left to right are the results of bicubic interpolation and the models trained on the Ori HR, Pos GT, Pos+Neg GT, respectively. **Please zoom in for better observation**.