Dual-lens Reference Image Super-Resolution Supplemental Material

Jing Zhu
jing.zhu1@samsung.com
Wenbo Li
wenbo.li1@samsung.com
Hongxia Jin
Hongxia.jin@samsung.com

Samsung Research America AI Center

Abstract

In this supplementary, we provide an example to outline the work flow of the feature assemble module, and depict the training architecture for both of our base model and enhanced model. Then, we present the full adaption result comparison on data with different degradation methods. We further visualize more examples to show the effectiveness of our proposed models with comparison to other alternative approaches. Finally, we report the 8x super-resolution results of our proposed models and make the comparison to the state-of-the-art methods.

1 Work Flow of Feature Assemble

Both of our base model and enhanced model consist of a set of feature assemble modules among different source image and reference image pairs. Feature assemble is a crucial component to find and extract the most relevant features from the reference images to the source image super-resolution process. To better understand how our feature assemble works, we take the enhanced model as an example to depict the work flow inside the feature assemble module between source image vs IR-Ref image and IR-Ref image vs HR-Ref image pairs (as shown in Fig. 1). The feature maps are first split into patches, and the features from the most relevant patch (based on the similarity map) will be selected to fulfill the corresponding area in the new assemble feature map. Finally, an upsampled-source-image-feature-size feature map will be assembled from the upsampled IR-Ref features, while an upsampled-IR-Ref-feature-size feature map is assembled by the features from the HR-Ref features.

The mechanism of the features assemble module in the base model is the same as the one in Fig. 1, but it is applied among source image vs IR-Ref and source image vs HR-Ref image pairs. Therefore, there would be two upsampled-source-image-feature-size feature maps assembled from the IR-Ref features and HR-Ref features respectively.

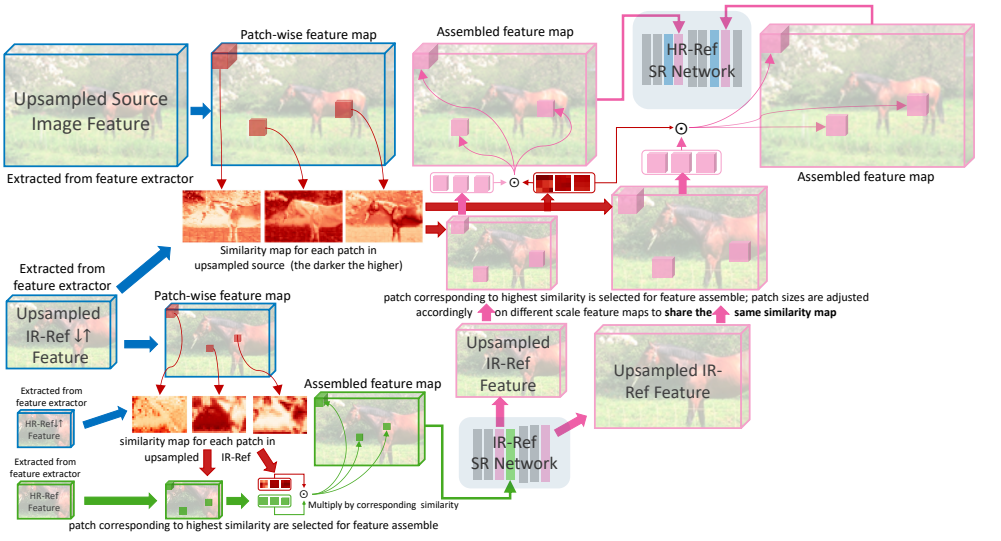

Figure 1: Visualization of the feature assemble module for the enhanced model, where feature assemble module is applied to source image vs IR-Ref image and IR-Ref image vs HR-Ref image pairs. The feature assemble module in the base model works similarly but it is applied to source image vs IR-Ref image and source image vs HR-Ref image pairs. $\odot$ means element-wise multiplication. Please Zoom In for better view to check the details.

2 Architecture of Training Model

As mentioned in the paper, our base model consists of three components, i.e., a feature extractor, a super-resolution (SR) network and a feature assemble module. We calculate the reconstruction loss and adopt dual regression loss at each feature scale level as [1] for better training. As shown in Fig. 2, in order to compute the dual regression loss, an extra convolutional layer ($C_{1\times}$ and $C_{2\times}$) is appended to convert the extracted multi-scale features ($S_{1\times}$ and $S_{2\times}$) from the model into images (denoted as $I_{SR_{1\times}}$ and $I_{SR_{2\times}}$), respectively. Two extra convolutional layers $D_{4\times}$ and $D_{2\times}$ are added to sequentially downscale the generated SR images (denoted as $I_{DR_{1\times}}$, $I_{DR_{2\times}}$). More specifically, $D_{4\times}$ downscales the generated SR ($I_{SR_{4\times}}$) to $I_{DR_{2\times}}$, and then $D_{2\times}$ further downscales the generated $I_{DR_{2\times}}$ to $I_{DR_{1\times}}$.

After we obtain $I_{SR_{4\times}}$, $I_{SR_{2\times}}$, $I_{SR_{1\times}}$, $I_{DR_{2\times}}$, $I_{DR_{1\times}}$, the reconstruction loss and the dual regression loss can be computed between the obtained images and the ground truth images (i.e., $I_{HR_{4\times}}$, $I_{HR_{2\times}}$, I_{LR}). The reconstruction loss ($\mathcal{L}_{rec}$) includes $L1$ errors between all the I_{SR} images and the ground truth images as

$$\mathcal{L}_{rec} = ||I_{SR_{1\times}} - I_{LR}||_1 + ||I_{SR_{2\times}} - I_{HR_{2\times}}||_1 + ||I_{SR_{4\times}} - I_{HR_{4\times}}||_1. \quad (1)$$

For the dual regression loss ($\mathcal{L}_{dual}$), it contains the $L1$ errors between I_{DR} images and the ground truth ones following

$$\mathcal{L}_{dual} = ||I_{DR_{1\times}} - I_{LR}||_1 + ||I_{DR_{2\times}} - I_{HR_{2\times}}||_1. \quad (2)$$

N_{LR}, $N_{HR_{2\times}}$ and $N_{HR_{4\times}}$ represent the sizes of the images at different scales. The entire model is trained with $\mathcal{L} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{dual}$. Note that the extra convolutional layers $C_{1\times}$, $C_{2\times}$, $D_{4\times}$ and $D_{2\times}$ are used during training only, which will be neglected when testing.

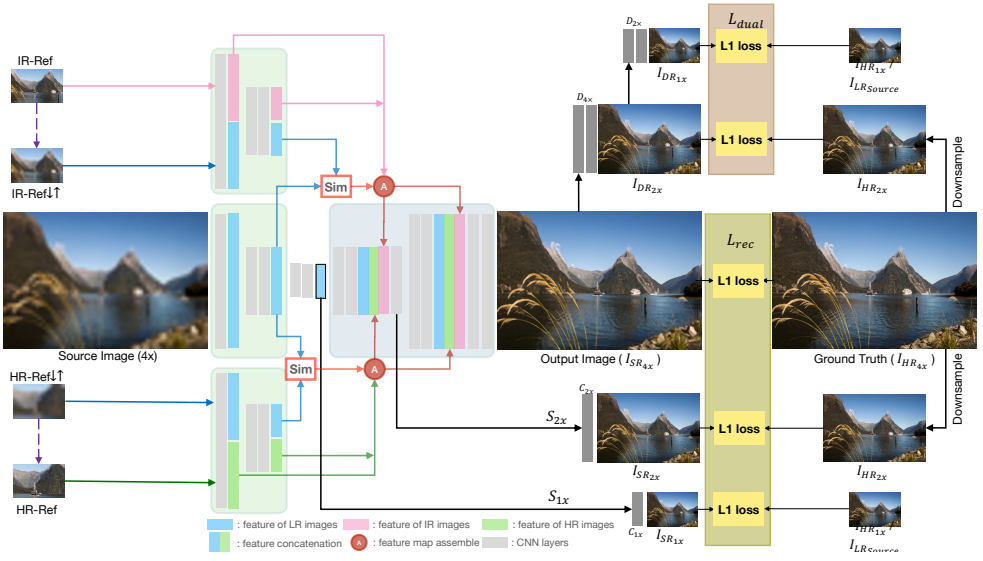

Figure 2: The training framework of our **base model** for (4x) super-resolution. Extra layers ($C_{1\times}$, $C_{2\times}$, $D_{4\times}$, $D_{2\times}$) are added to construct the reconstruction loss and the dual regression loss. The reconstruction loss ($\mathcal{L}_{rec}$) includes $L1$ errors between all the I_{SR} images and ground truth images, while the dual regression loss ($\mathcal{L}_{dual}$) contains the $L1$ errors between I_{DR} images and the ground truth ones.

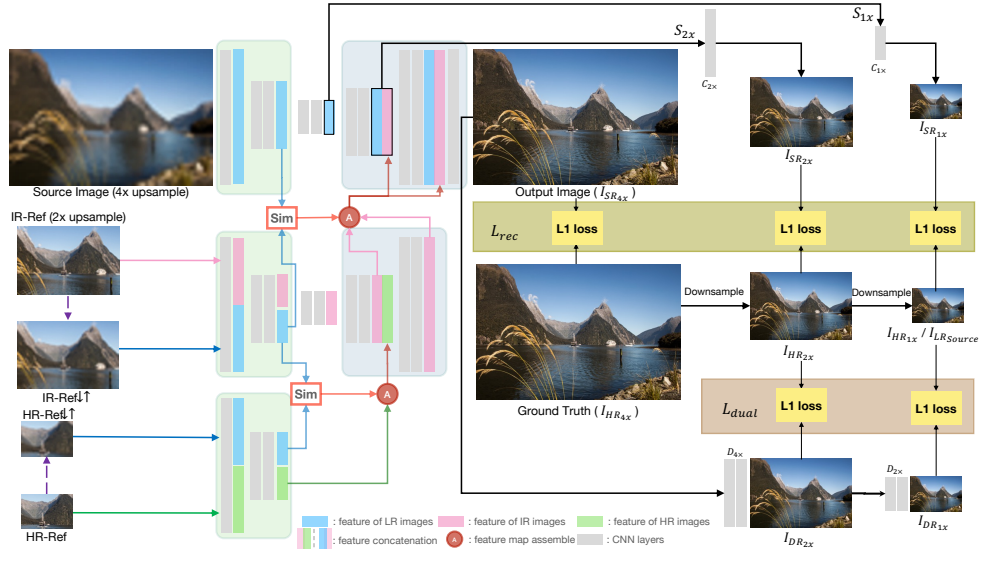

Figure 3: The training framework of our **enhanced model** for (4x) super-resolution. Extra layers ($C_{1\times}$, $C_{2\times}$, $D_{4\times}$, $D_{2\times}$) are added to construct the reconstruction loss and the dual regression loss. The reconstruction loss ($\mathcal{L}_{rec}$) includes $L1$ errors between all the I_{SR} images and ground truth images, while the dual regression loss ($\mathcal{L}_{dual}$) contains the $L1$ errors between I_{DR} images and the ground truth ones.

Since we utilize the same kind of losses to train our enhanced model, the training model architecture of our enhanced model is similar to the base model one, as depicted in Fig. 3.

Table 1: The detailed network structure of our **base model** and **enhanced model** for 4x super-resolution. The numbers for each layer represent *in channels*, *out channels*, *kernel size* and *stride*, respectively. RCAB denotes the residual channel attention block as described in [9].

Component		**Layer Details**	
		Base Model	Enhanced Model
Feature Extractor		Conv (3, 16, 3, 1)	Conv (3, 16, 3, 1)
		Conv (16, 16, 3, 2), LeakyReLU	Conv (16, 16, 3, 2), LeakyReLU
		Conv (16, 32, 3, 1)	Conv (16, 32, 3, 1)
	Scale 1×	Conv (32, 32, 3, 2), LeakyReLU	Conv (32, 32, 3, 2), LeakyReLU
		Conv (32, 64, 3, 1)	Conv (32, 64, 3, 1)
SR Network	Scale 2×	40 RCABs	40 RCABs
		Conv(64, 256, 3, 1), 2× pixel shuffle	Conv(64, 256, 3, 1), 2× pixel shuffle
		Conv(64, 32, 1, 1)	Conv(64, 32, 1, 1)
		Conv(96, 64, 1, 1)	
	Scale 4×	40 RCABs	40 RCABs
		Conv(64, 256, 3, 1), 2× pixel shuffle	Conv(64, 256, 3, 1), 2× pixel shuffle
		Conv(64, 16, 1, 1)	Conv(64, 16, 1, 1)
		Conv(48, 32, 1, 1)	Conv(32, 3, 3, 1)
		Conv(32, 3, 3, 1)	
Training Auxiliary	$C_{1\times}$	Conv(64, 3, 3, 1)	Conv(64, 3, 3, 1)
	$C_{2\times}$	Conv(64, 3, 3, 1)	Conv(64, 3, 3, 1)
	$D_{4\times}$	Conv(3, 16, 3, 2), LeakyReLU	Conv(3, 16, 3, 2), LeakyReLU
		Conv(16, 3, 3, 1)	Conv(16, 3, 3, 1)
	$D_{2\times}$	Conv(3, 16, 3, 2), LeakyReLU	Conv(3, 16, 3, 2), LeakyReLU
		Conv(16, 3, 3, 1)	Conv(16, 3, 3, 1)

Table 1 provides the detailed structure for each layer in different components of both base model and the enhanced model. The base model and the enhanced model share similar layer structures except the layers right after each feature concatenation operation. The key advances of the enhanced model is the specifically designed progressive feature enrichment mechanism for SR improvement.

3 Full Adaption Results Comparison on Different Degradation

We test the generalization capability of our models on data with different degradation methods. For the compared reference-based super-resolution (RefSR) methods (e.g., SRNTT, TTSR), they can utilize either IR-Ref or HR-Ref as the reference for SR generation, since their model is designed for single reference case. In the paper, due to the page limit, we solely report the results of the compared RefSR methods using the HR-Ref images as their reference images, since their models usually perform better with HR-Ref images comparing to the ones with IR-Ref images. In this supplementary, we present the full comparison of those RefSR methods that includes results from the models worked with IR-Ref images as the references (as shown in Table 2). It can be found that the RefSR methods with HR-Ref images obtain higher PSNR and SSIM than the ones with IR-Ref images, since the HR-Ref images contain high-resolution information while the IR-Ref images only provide intermediate-resolution textures. Our base model achieves better adaptive results than all the other RefSR methods, while our enhanced model performs the best among all the compared approaches.

Table 2: The PSNR/SSIM comparisons of (4x) super-resolution with alternative approaches on images with different degradation methods. The best performance has been **bolded**. Higher score indicate the better SR performance.

Method	Reference	Nearest Neighbor Degradation					Unknown
		Set14	BSDS100	Urban100	Manga109	Real104	DIV2K Val
Nearest	×	22.68 / .629	22.92 / .601	20.06 / .596	21.77 / .738	19.10 / .479	23.07 / .595
DRN	×	22.54 / .628	22.90 / .618	20.01 / .593	22.40 / .731	19.93 / .477	23.12 / .607
SRNTT-*rec*	IR-Ref	21.43 / .622	21.45 / .586	18.62 / .574	19.87 / .698	18.88 / .505	22.90 / .591
SRNTT	IR-Ref	21.06 / .514	20.93 / .480	18.49 / .490	19.85 / .635	18.03 / .379	22.71 / .579
TTSR-*rec*	IR-Ref	20.32 / .596	20.47 / .569	17.90 / .559	20.45 / .737	18.46 / .505	23.45 / .601
TTSR	IR-Ref	18.72 / .440	18.73 / .429	16.74 / .453	19.00 / .627	16.73 / .373	22.56 / .578
MASA-*rec*	IR-Ref	20.79 / .611	20.87 / .585	17.75 / .563	20.58 / .572	17.86 / .495	22.77 / .588
MASA	IR-Ref	17.41 / .301	17.32 / .298	15.49 / .322	17.28 / .221	15.25 / .230	21.89 / .529
C^2-Matching-*rec*	IR-Ref	18.86 / .563	19.24 / .564	16.30 / .530	18.53 / .707	16.73 / .480	22.28 / .565
C^2-Matching	IR-Ref	18.66 / .524	18.91 / .519	15.90 / .495	18.29 / .657	16.27 / .448	21.98 / .547
SRNTT-*rec*	HR-Ref	21.58 / .629	21.50 / .589	18.78 / .580	19.97 / .703	18.91 / .502	22.99 / .598
SRNTT	HR-Ref	21.51 / .544	21.27 / .505	18.71 / .510	19.90 / .653	18.07 / .380	22.76 / .583
TTSR-*rec*	HR-Ref	21.55 / .641	21.69 / .616	18.63 / .598	20.96 / .750	18.50 / .511	23.53 / .607
TTSR	HR-Ref	18.97 / .470	19.16 / .473	16.98 / .481	19.20 / .644	16.76 / .378	22.63 / .584
MASA-*rec*	HR-Ref	20.92 / .614	20.96 / .585	17.93 / .566	20.91 / .578	17.97 / .501	22.83 / .597
MASA	HR-Ref	18.62 / .386	18.52 / .385	16.57 / .409	18.58 / .324	16.22 / .303	21.94 / .536
C^2-Matching-*rec*	HR-Ref	18.94 / .565	19.22 / .567	16.27 / .530	18.58 / .709	16.72 / .484	22.41 / .569
C^2-Matching	HR-Ref	18.92 / .531	19.04 / .521	15.88 / .499	18.31 / .664	16.32 / .447	22.07 / .552
Base Model	IR-Ref + HR-Ref	22.73 / .633	23.05 / .621	20.17 / .601	22.40 / .744	20.27 / .514	24.57 / .610
Enhanced Model	IR-Ref + HR-Ref	**23.28 / .649**	**23.41 / .632**	**20.47 / .609**	**22.59 / .753**	**20.44 / .522**	**24.96 / .621**

Figure 4: Some examples from the simulated datasets (left) and the phone-captured real dataset (right). The datasets cover both indoor and outdoor scenes.

4 Visualize More Examples

As discussed in the paper, in order to validate our models to address the newly defined multi-lens reference image super-resolution problem, we construct a number of simulated datasets based on publicly available SISR datasets, where each image is extended with two simulated multi-lens reference images. We even use an IPhone 12 Pro to capture a set of images by the embedded multiple cameras simultaneously to build a real phone-captured dataset. The images are collected from a range of indoor and outdoor scenarios, from single object to scenery. Fig. 4 displays some samples from our simulated datasets and real phone-captured dataset.

In the paper, we have visualized a few 4x generated SR images using our proposed models with comparison to the ground truth and the results generated by alternative approaches (e.g., DRN, SRNTT, TTSR, MASA, C^2-Matching). In this supplementary, we present more examples for further qualitative comparison in Fig. 5, Fig. 6, Fig. 7 and Fig. 8. It can be seen that our models are able to produce high-quality SR images with clearer shapes and finer details when comparing to other methods. We can also easily recognize the improvement of results from the enhanced model over those from the base model, demonstrating the effectiveness of the proposed progressive feature fusion schema. Moreover, we visualize the entire generated images instead of regional selection in Fig. 9 and Fig. 10.

5 8× Super-Resolution Comparison

For comprehensive comparison, we also test our proposed methods for 8× super-resolution using the models in Fig. 11 and 12. We collect and report the PNSR/SSIM for the 8× super-resolution images from each dataset with compared methods in Table 3. Since existing RefSR methods (e.g., SRNTT, TTSR) mainly work on 4× super-resolution task, they could not work directly on the 8× super-resolution. Therefore, we provide the results of state-of-the-arts SISR methods here for comparison. As we can see from the table, our base model achieves comparable performance with other SISR methods, while our enhanced model outperforms all the compared methods with highest scores among all the compared datasets.

Table 3: The PSNR/SSIM comparisons of (8x) super-resolution with alternative approaches on five common validation datasets. The best performance has been **bolded**. Note that, most of the RefSR methods only be applied on 4x super-resolution, so we report the results of SISR methods for comparison.

Method	Reference	Set14	BSDS100	Urban100	Manga109	Real104
Bicubic	SISR	23.19 / .568	23.67 / .547	20.74 / .515	21.47 / .649	19.63 / .453
ESPCN [5]		23.45 / .598	23.92 / .574	21.20 / .554	22.04 / .683	20.55 / .464
SRResNet [9]		24.55 / .624	24.65 / .587	22.05 / .589	23.88 / .748	20.87 / .489
SRGAN [8]		21.57 / .495	21.78 / .442	19.64 / .468	20.42 / .625	18.54 / .434
LapSRN [20]		24.35 / .620	24.54 / .585	21.81 / .580	23.39 / .734	20.84 / .473
EDSR [4]		25.05 / .641	24.80 / .595	22.55 / .618	24.54 / .775	20.94 / .479
RCAN [6]		25.23 / .651	24.96 / .605	22.97 / .643	25.23 / .802	21.11 / .490
DRN [10]		25.25 / .652	24.98 / .605	22.96 / .641	25.30 / .805	21.17 / .491
Base Model	IR-Ref + HR-Ref	25.33 / .657	25.12 / .608	23.05 / .644	25.36 / .809	21.43 / .495
Enhanced Model	IR-Ref + HR-Ref	**25.39 / .662**	**25.22 / .616**	**23.11 / .653**	**25.50 / .814**	**21.53 / .501**

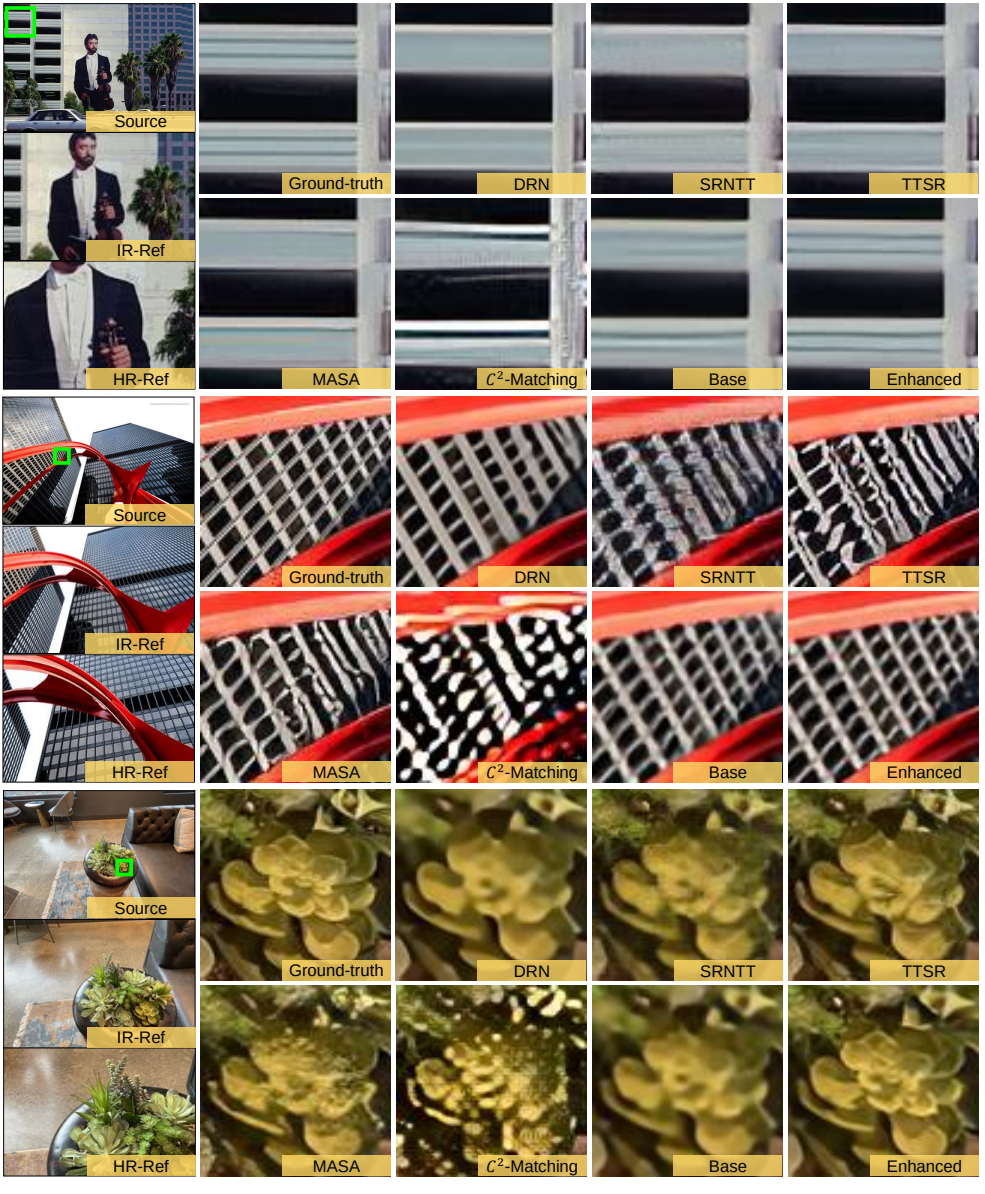

Figure 5: Examples of the (4x) SR results using our proposed base model (Base) and enhanced model (Enhanced) with comparison to those generated by other methods. Our models can produce high-quality SR images with fine details.

References

[1] Yong Guo, Jian Chen, Jingdong Wang, Qi Chen, Jiezhang Cao, Zeshuai Deng, Yanwu Xu, and Mingkui Tan. Closed-loop matters: Dual regression networks for single image super-resolution. In *CVPR*, 2020.

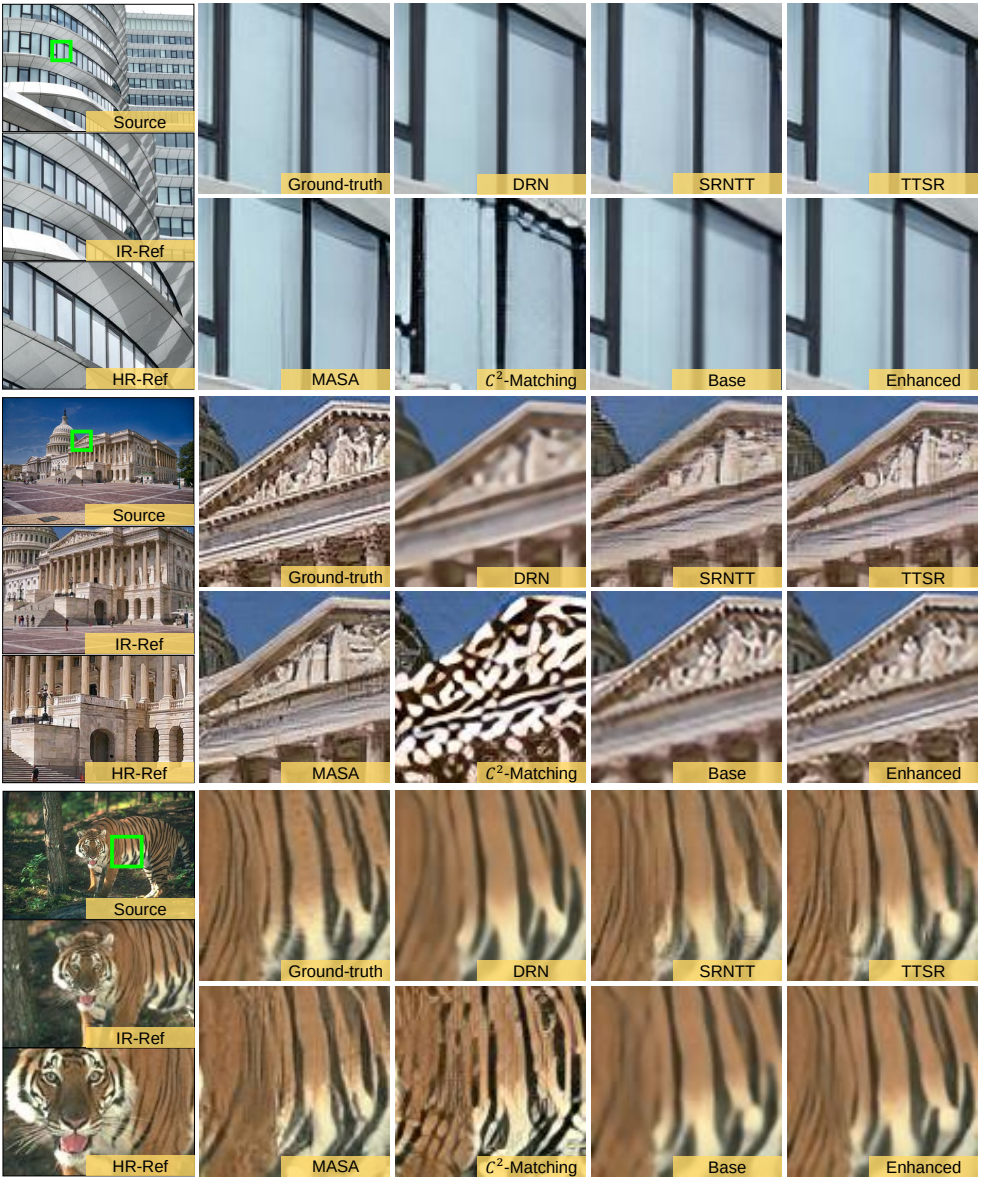

Figure 6: Examples of the (4x) SR results using our proposed base model (Base) and enhanced model (Enhanced) with comparison to those generated by other methods. Our models can produce high-quality SR images with fine details.

[2] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017.

[3] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In

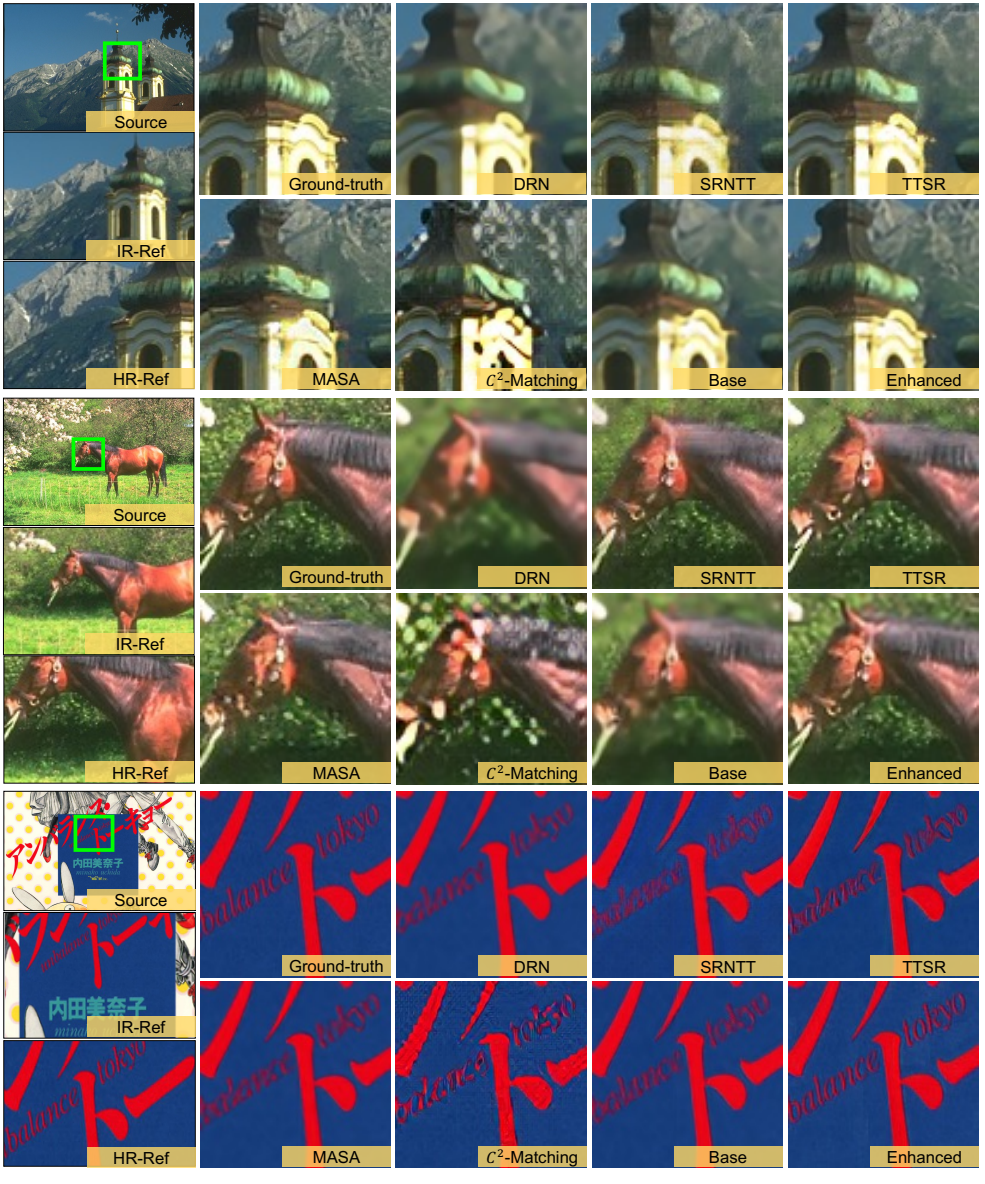

Figure 7: Examples of the (4x) SR results using our proposed base model (Base) and enhanced model (Enhanced) with comparison to those generated by other methods. Our models can produce high-quality SR images with fine details.

CVPR, 2017.

[4] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017.

[5] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob

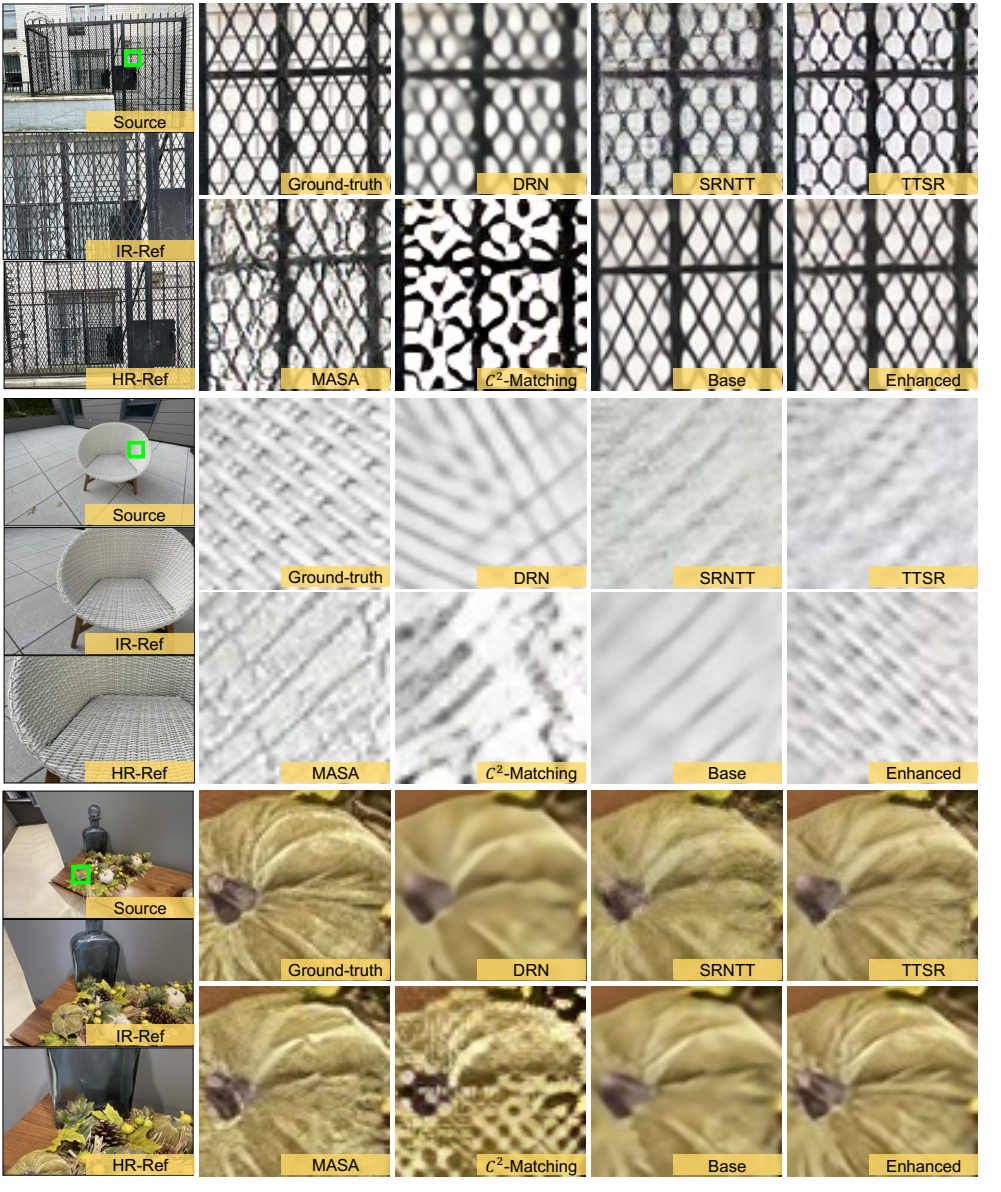

Figure 8: Examples of the (4x) SR results using our proposed base model (Base) and enhanced model (Enhanced) with comparison to those generated by other methods. Our models can produce high-quality SR images with fine details.

Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.

[6] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018.

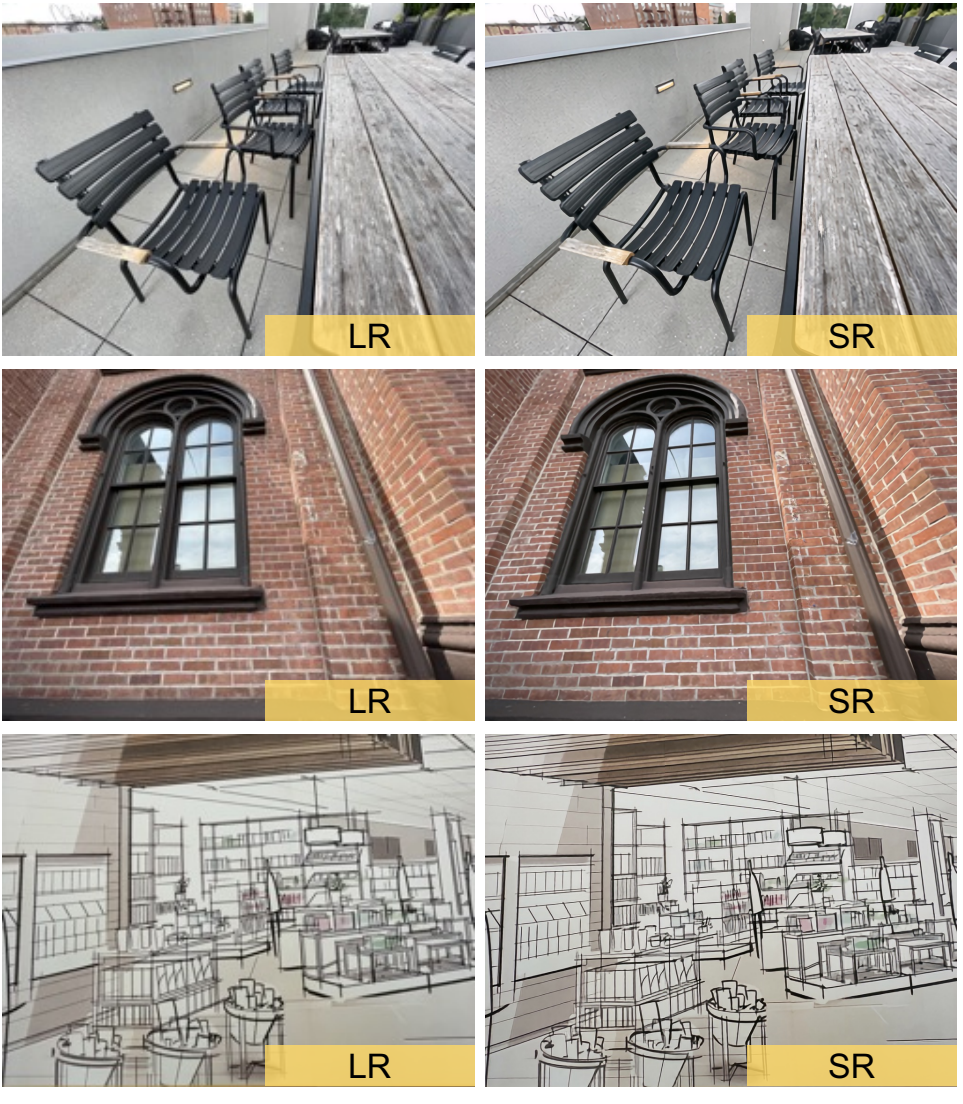

Figure 9: Examples of the (4x) SR results using our proposed enhanced model. LR denotes the low-resolution input while SR is the super-resolution images from our model.

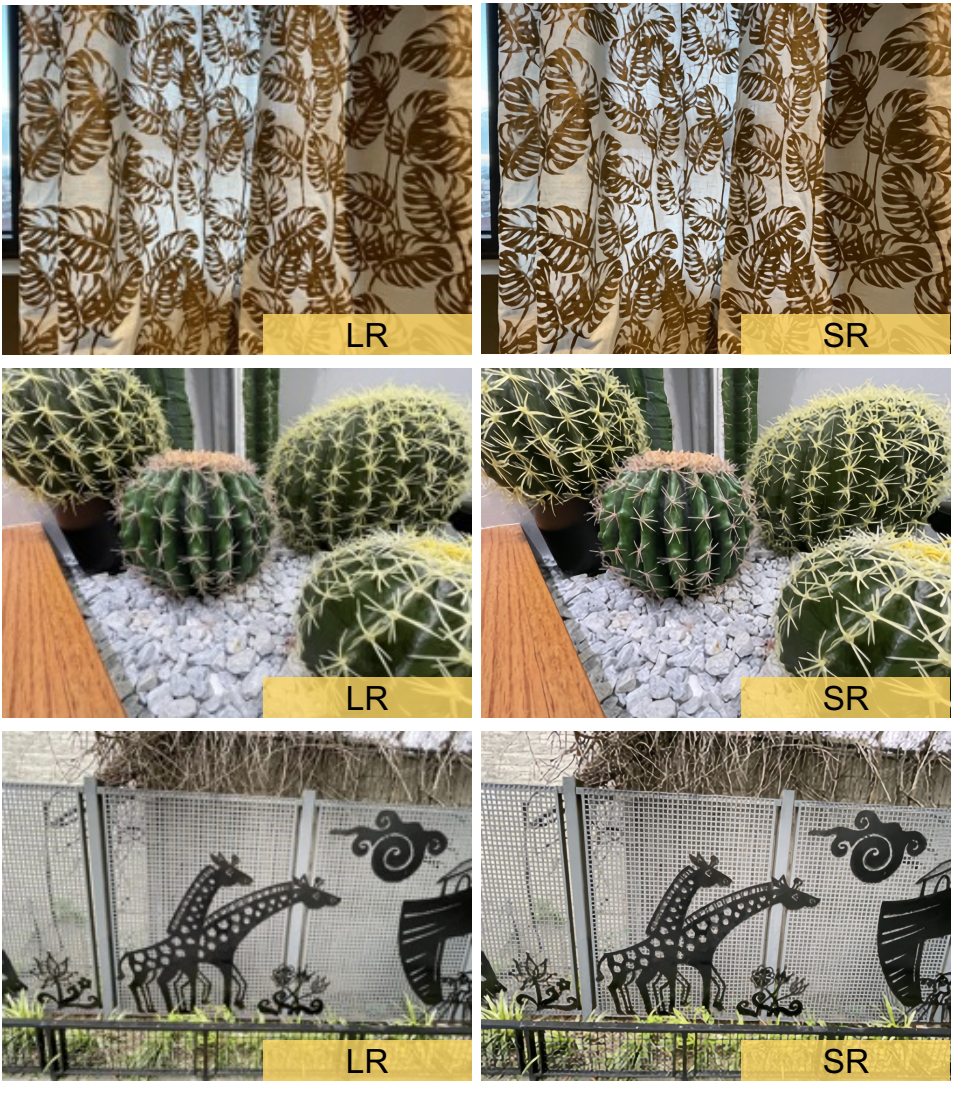

Figure 10: Examples of the (4x) SR results using our proposed enhanced model. LR denotes the low-resolution input while SR is the super-resolution images from our model.

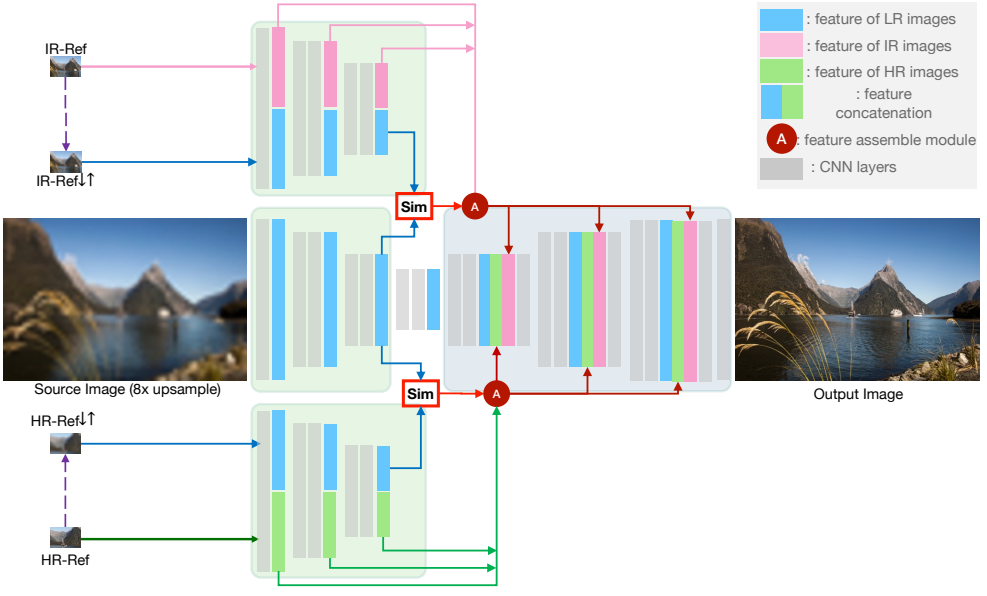

Figure 11: The framework of our **base model** for (8×) super-resolution.

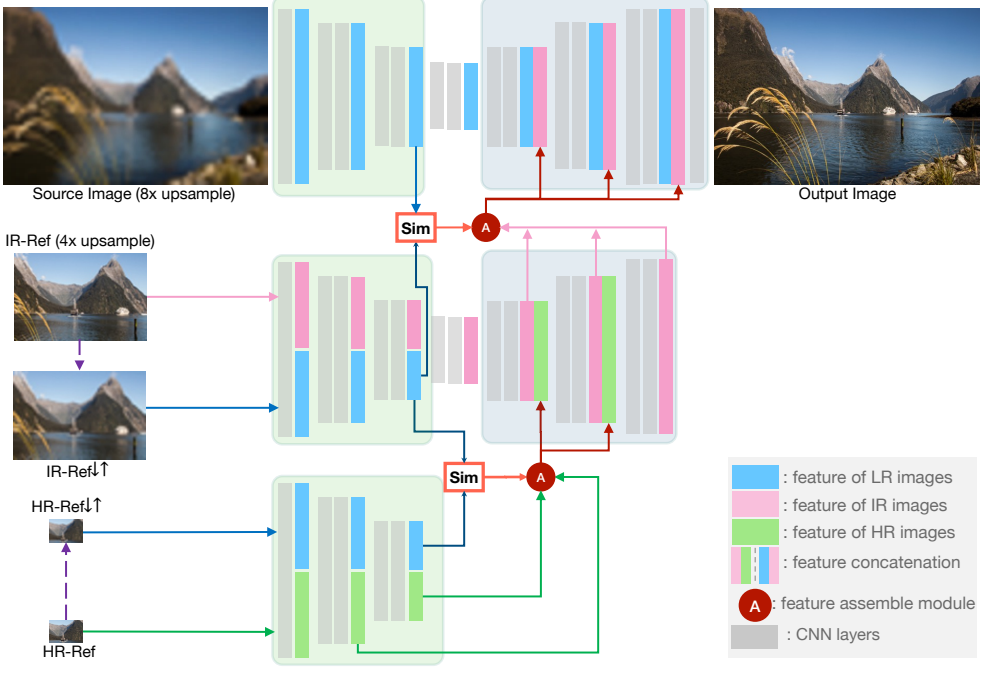

Figure 12: The framework of our **enhanced model** for (8×) super-resolution.