DETAIL LOSS IN SUPER-RESOLUTION MODELS BASED ON THE LAPLACIAN PYRAMID AND REPEATED UPSCALING-DOWNSCALING STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

With advances in artificial intelligence, image processing has also gained significant interest. Image super-resolution, in particular, is a vital technology closely related to real-life applications, as it enhances the quality of existing images. Since enhancing details is important in the super-resolution task, it is often necessary to activate pixels that appear only at high frequencies, distinct from low frequencies. In this paper, we propose a method that generates a detail image separately from the super-resolution image. This approach introduces a loss function designed to enhance detail, allowing the model to generate an upscaled image and a detail image independently, with control over each component. Consequently, the model can focus more effectively on high-frequency data, resulting in an improved super-resolution image. Our loss function utilizes detail images based on the Laplacian Pyramid, which is widely used in image reconstruction. The multi-level property of the Laplacian Pyramid is well-suited for applying upscaling and downscaling repeatedly. Our experiments demonstrate that a structure applying the repetition of upscaling and downscaling integrates effectively with our detail loss control. The results show that this structure efficiently extracts diverse information, enabling the generation of improved super-resolution images from multiple low-resolution features. We conduct two types of experiments. First, we construct a simple CNN-based model incorporating the Laplacian Pyramid-based detail control and a repeated upscaling and downscaling structure. This model achieves a state-of-the-art PSNR value of 38.48 dB, surpassing all currently available CNN-based models and even some attention-based models without additional special techniques. Second, we apply our methods to existing attention-based models on a small scale. In all the experiments, attention-based models using our detail loss show improvements compared to the original models. These experiments demonstrate that our detail control loss effectively enhances performance, regardless of the model's structure in the super-resolution task.

1 INTRODUCTION

In recent years, advances in hardware have enabled the handling of high-resolution (HR) images, making image processing techniques increasingly essential tools. One such technique is the single image super-resolution (SR), a low-level vision task that generates a high-resolution image from a low-resolution (LR) one. Since this classical problem is ill-posed, meaning that multiple HR images can correspond to a single LR image, the single image SR is challenging. However, it attracts significant interest due to its applications in various fields, such as medical imaging (Greenspan, 2009; Isaac & Kulkarni, 2015; Sood et al., 2018), object detection (Na & Fox, 2018; Haris et al., 2021b), and satellite image analysis (Shermeyer & Van Etten, 2019; Cornebise et al., 2022).

Deep learning methods, which have received explosive focus, have been actively used in image processing and are also connected to super-resolution (Dong et al., 2016; Kim et al., 2016a; Wang et al., 2018; Talab et al., 2019; Hui et al., 2021), significantly improving performance. Researchers have explored various approaches, such as developing deeper convolutional neural network (CNN) (Kim et al., 2016b; Lim et al., 2017; Ahn et al., 2018) and designing algorithms (Lai et al., 2017; Liu et al., 2018; 2019a; Sun & Chen, 2020; Haris et al., 2021a; Anwar & Barnes, 2022; Lee & Jin,

2022) that integrate existing image processing techniques. In particular, approaches using attention-based structures (Liu et al., 2019b; Niu et al., 2020; Li et al., 2021; Liang et al., 2021; Zhang et al., 2022; Chen et al., 2023) have recently been proposed. In this circumstance, many deep learning methods focus on enhancing the ability to capture proper features from an LR image and carry them until the end. Thus, in many cases (Liu et al., 2020; Niu et al., 2020; Haris et al., 2021a; Anwar & Barnes, 2022; Chen et al., 2023), the generation of the super-resolution image employs a simple upsampler, and the model is trained using one loss function based on the SR image. However, in SR tasks where refining high-frequency detail is crucial, relying solely on one loss function for SR may provide insufficient guidance for capturing fine details.

In this paper, we propose a detail control loss based on the Laplacian Pyramid (LP) to guide the detail part of SR. Our method leverages the reconstruction concept of the LP, which generates an HR image by adding an upsampled approximation image with a detail image (Burt & Adelson, 1987). It creates a feature map for the detail image from the upsampled features and controls it separately from the SR by introducing an additional loss function. The approach allows the model to activate meaningful pixels for high-frequency details and focus more on generating these fine details. Additionally, we apply a repeated upscaling and downscaling process (RUDP). RUDP repeats downsampling the completed SR feature map and then combining it with the LR image to extract new upsampled approximation and detail features. Our experiments demonstrate that combining RUDP with the LP-based detail control method effectively extracts various information from the LR image.

We conduct two main experiments. These can be broadly classified as follows. First, we construct a simple CNN-based model, Laplacian pyramid-based Upscaling and Downscaling Super-Resolution Network (LaUD), that incorporates the above two methods. This CNN-based model outperforms all currently available state-of-the-art (SOTA) CNN models in the PSNR metric and has also surpassed some attention-based models. Additionally, our ablation study and qualitative analysis demonstrate that our detail control loss and RUDP are effective methods for improving performance. We also confirm that their effectiveness is further enhanced when both methods are used together. Second, we apply our method to existing attention-based models on a small scale. Comparing the results with and without our method, we observe that its application consistently enhances performance across all models. These results show that our method is applicable both with and without attention mechanisms and can also improve the performance of attention-based models.

In summary, our main contributions are the following:

- We propose a new method, the detail control loss based on the LP. This method allows the model to handle the detail image for high-frequency information apart from the SR image. Consequently, the model can focus more on the detail part and supplement information not present in the upsampled image.
- We verify that RUDP effectively integrates with the LP-based detail control. Our experiments demonstrate that RUDP allows the model to capture more diverse information by re-extracting features from the SR features supplemented with details.
- We apply our methods to both CNN-based and attention-based models. As a result, all the models perform effectively, demonstrating that our methods successfully supplement high-frequency information, regardless of the model's structure.

2 Related Works

2.1 Early CNN Models in Super-Resolution

Many studies (Dong et al., 2016; Kim et al., 2016a;b; Zhang et al., 2017) have aimed to deepen models more efficiently in the early days of deep learning for image SR. VDSR (Kim et al., 2016b) is a pioneer in this direction, designing deeper structures using the residual learning. Subsequently, several papers have developed efficient models based on residual networks. EDSR (Lim et al., 2017) enhances performance by constructing a multi-scale structure with residual blocks. CARN (Ahn et al., 2018) introduces a cascade connection between residual blocks, allowing the model to produce SR images efficiently even with fewer parameters. Similarly, in our CNN-based experiments, our LaUD utilizes residual blocks and skip connections to deliver information from the initial to the end. Moreover, RUDP enables LaUD to extract more diverse features for SR within a deep architecture.

2.2 Attention Mechanism

The transformer model has demonstrated excellent feature extraction performance and has been successfully adapted for visual tasks (Dosovitskiy et al., 2020; Liu et al., 2021; Touvron et al., 2021; Tu et al., 2022). Consequently, many studies utilize the attention mechanism in the SR task. Authors in Liu et al. (2020); Niu et al. (2020) enhance performance by using both channel-wise and spatial-wise attention simultaneously. DRLN (Anwar & Barnes, 2022) proposes channel attention with a pyramid concept to capture different sub-frequency-band information. HAT (Chen et al., 2023) and EDT (Li et al., 2021) modify the window shape to improve the connection among windows. Some papers, such as Liang et al. (2021); Zhang et al. (2022); Yang & Wu (2023), apply transformer models (Liu et al., 2021; Tu et al., 2022) that have demonstrated high performance in the visual domain. From the experiments that apply our methods to existing attention-based models, we see that our methods can be adapted to the attention-based model with tiny modifications. Therefore, our LP-based detail control and attention mechanism can result in a synergistic effect.

2.3 Loss Function for the Super-Resolution Task

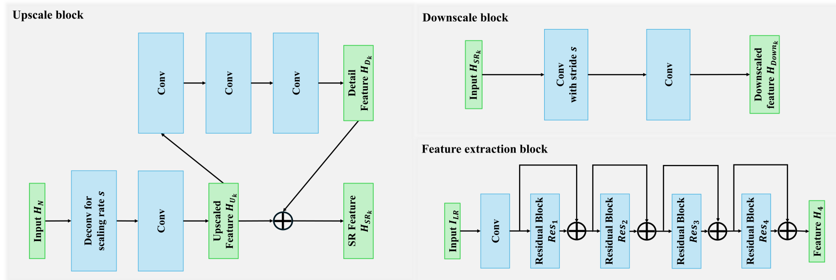
SR problems involve predicting fine details that are not visible in LR images. To address this challenge, many studies have sought to enhance performance by introducing various loss functions beyond traditional ones, such as mean squared error between SR and HR images. In Xu et al. (2017), the model generates multiple SR images and sums their mean squared error losses. While we also compute a weighted sum of multiple SR images when applying RUDP, our method introduces an additional loss specifically for details. Some papers, such as Johnson et al. (2016); Ledig et al. (2017), introduce an additional loss based on the feature maps of a pretrained model. Since a well-trained model captures the style of an image, including texture and patterns, its feature maps help address deficiencies in SR. Although using additional loss beyond SR and HR is similar to our LP-based detail approach, the key difference is that our method uses LP-based detail image to guide the model. In Sims (2020); Fuoli et al. (2021), high-frequency components are supplemented by leveraging frequency-domain information. In Seif & Androutsos (2018); Ge & Dou (2023), the authors extract detailed parts of images for new loss functions through edge detection and gradient extraction convolution. In particular, the method in Seif & Androutsos (2018) is quite similar to our approach. Although this method extracts edge images from HR images, it differs from our detail control in that its edge images are not involved in the reconstruction process of SR images.

2.4 Methods Based on Mathematical Theory

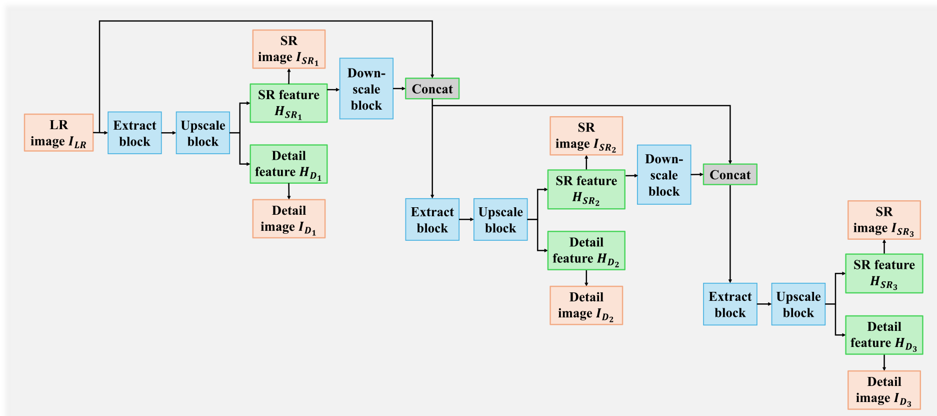
There have been many attempts to combine mathematical theories with deep learning. Given that wavelets can handle multi-resolution images and integrate naturally with a convolution layer, various researches (Huang et al., 2017; Liu et al., 2018; Jeevan et al., 2024) have been conducted. They generate low-frequency and high-frequency images of the same size from the LR input and apply the inverse wavelet transform to produce an SR image. In contrast, we use the LP-based reconstruction. The LP detail image, which is the same size as the HR image, contains more information. Combined with RUDP, this leads to enhanced abundance and diversity in feature extraction. In Lai et al. (2017); Anwar & Barnes (2022); Han et al. (2022), the authors introduce the pyramid structure of LP to their models. The authors of LapSRN (Lai et al., 2017) introduce a pyramidal reconstruction structure in LP. Although the strategies for generating details and the reconstruction process are similar to ours, our approach differs from LapSRN by using detail as the loss function, which guides the model to concentrate high-frequency data. In DRLN (Anwar & Barnes, 2022), the authors propose the Laplacian attention that generates feature maps of different scales similar to the pyramid structure of LP and use them as channel attention. Unlike DRLN, we directly control the detail feature map through the loss function and consider the LP pyramid structure only in the reconstruction process.

3 Method

The LP-based detail control we propose can be applied to various models because it relates to training rather than model structure. Therefore, we categorize the models into CNN-based and attention-based types and compare the effects of our method on each category. In this section, we outline the structure of the models and the loss functions used in each experiment.



(a) Sub-components of the model



(b) Entire model structure

Figure 1: The structure of our CNN model, LaUD. In (a), the figure illustrates the sub-components of the model. In (b), the figure shows the overall structure of LaUD.

3.1 CNN-BASED MODEL

For the CNN-based model, we design a new architecture, LaUD, that incorporates LP-based detail loss and RUDP. The model has sufficient depth but remains simple, without incorporating techniques beyond our two methods. This experiment demonstrates the performance of the model in comparison to existing SR models. An ablation study is conducted to further evaluate the impact of each method. Figure 1 shows the overall structure of LaUD. Our model consists of three main blocks: a feature extraction block, an upscale block, and a downscale block.

Feature extraction block. We construct the feature extraction block using only residual blocks and skip connections. For an LR image I_{LR}, the shallow feature H_0 is extracted by a convolution layer,

$$H_0 = Conv(I_{LR}). \tag{1}$$

This convolution layer also helps in uniformly adjusting the number of channels in the feature map before it enters residual blocks during the subsequent RUDP process. Then several residual blocks Res_n with skip connection extracts deeper features,

$$H_n = H_{n-1} + Res_n(H_{n-1}), \quad n = 1, 2, \ldots, N. \tag{2}$$

We choose $N = 4$ and LeakyReLU as the activation function for all processes in our simple model. All convolution layers have a kernel size of 3×3.

Upscale block. The final feature H_N is delivered to the upscale block. The upscale block creates both the upscaled feature H_{U_k} and the detail feature H_{D_k}, where k denotes the order of upscaling within the entire RUDP. Then the two features are added to complete the SR feature H_{SR_k}, similar to the usual construction process of LP: For $k = 1, 2, \ldots, K$, where K is the maximum order,

$$H_{U_k} = Conv(Deconv(H_N)), \tag{3}$$
$$H_{D_k} = Conv(Conv(Conv(H_{U_k}))), \tag{4}$$
$$H_{SR_k} = H_{U_k} + H_{D_k}. \tag{5}$$

Unlike the back-projection in Liu et al. (2019b;a); Haris et al. (2021a), our upscale block generates the SR feature by employing one deconvolution layer and a few convolution layers, thus avoiding complex structure with multiple processes. Our detail loss enables the model to effectively generate the SR feature, even with a simple structure. In this process, there are many ways to generate H_{D_k}, but we choose to derive H_{D_k} from H_{U_k}. As a result, our upscale block returns two feature maps: the detail feature H_{D_k} and the SR feature H_{SR_k}. Each of these feature maps forms a distinct loss.

Downscale block and repetition. For the downsampling process of the SR features in our RUDP structure, we employ one convolution layer with downsampling followed by a convolution layer,

$$H_{Down_k} = Conv(Conv_{\downarrow}(H_{SR_k})), \quad k = 1, 2, \ldots, K-1, \tag{6}$$

where $Conv_{\downarrow}$ indicates convolution layer with downsampling. The generated H_{Down_k} is concatenated with the input LR image I_{LR} and LR feature $H_{Down_{k-1}}$, and then delivered back to the next feature extraction block. Through this process, the feature extraction block extracts more diverse information for the next SR image by referring to the SR features generated in the previous step. We design our LaUD to set $K = 3$. Consequently, LaUD produces detail feature maps $\{H_{D_k}\}_{k=1,2,3}$, SR feature maps $\{H_{SR_k}\}_{k=1,2,3}$, and downscale feature maps $\{H_{Down_k}\}_{k=1,2}$.

Result images and loss function. To ensure delivery without information loss, each block within the model hands over feature maps as they are. Consequently, it is necessary to convert the feature maps to the RGB format at the end. We achieve this conversion with a ToRGB layer using a 1×1 convolution. The entire loss function consists of the loss L_s for the SR image and the L_d for the LP detail. We choose the $L1$ loss function, which effectively reduces the smoothing effect and shows an outstanding ability for image restoration (Zhao et al., 2017). For the SR images, the L_s is the weighted sum of three losses between the HR image I_{HR} and the SR images $\{I_{SR_k}\}$, obtained from $\{H_{SR_k}\}$. For the detail images, we first generate a detail I_D from I_{HR} by the LP process. Then the L_d is defined through the weighted sum of losses between the I_D and the three detail images $\{I_{D_k}\}$, generated by the model from $\{H_{D_k}\}$. The weights used are the same as those used in the L_s. As a result, the final loss is $L = \alpha \cdot L_s + \beta \cdot L_d$, where α and β are the weights, and

$$L_s = \sum_{k=1}^{3} W_k \cdot ||I_{HR} - I_{SR_k}||_1, \quad L_d = \sum_{k=1}^{3} W_k \cdot ||I_D - I_{D_k}||_1. \tag{7}$$

3.2 Attention-based Models

For the attention-based model, we aim to demonstrate that our methodology integrates seamlessly without disrupting the existing attention structure. Therefore, we applied our method to several existing attention-based models and compared the results to those of the original models. This experiment shows that our LP-based detail control is not limited to CNNs but is also effective across various structures. The LP-based detail loss can be implemented with minor modifications to the output part of a model. However, some models require significant structural changes to incorporate our RUDP, which enhances the effectiveness of detail loss. Since these changes may not provide a valid basis for a fair comparison, only the LP-based detail loss is applied to such models.

Choice of base models. We attempted to select SOTA models to examine the results appropriately. However, due to limitations in computing resources, we were only able to conduct experiments on models that require less memory during training. Although we did not test our method on all models, we demonstrated its effectiveness in attention-based models based on the trends observed in the selected models.

We chose the base model according to the following criteria: 1. Models for which the authors provide their code to enable reproduction. 2. Models that can be trained within our resource constraints. 3. Models that demonstrate sufficiently high performance. 4. Models that each use different attention approaches. As a result, three models—ABPN (Liu et al., 2019b), HAN (Niu et al., 2020), and DRLN (Anwar & Barnes, 2022)—were selected. To isolate the effects of our method, we reproduced the original model and compared it with the version to which our method was applied. The reproduction of each model was carried out using the code provided in their respective papers. We primarily used the hyperparameters specified in the papers, and for details not mentioned, we followed the defaults used in their code. When applying our method, all hyperparameters were kept identical to those used in the reproduction.

Application of our methods. To apply our LP-based detail loss, the model must generate detail images separately from the SR image. Hence, the output part of each model requires some modifications. Here, we briefly describe our modification for each model. More details are in the appendix.

ABPN has an iterative up- and down-sample structure similar to our RUDP. We replace only this structure with the upscale and downscale blocks from our LaUD, minimizing modifications to the existing model methodology. Since the attention mechanism in ABPN operates on features after downsampling, our modification enables the model to handle detail features without altering the attention mechanism structure. HAN employs a structure in which layer and channel-spatial attention are applied after feature extraction using residual channel attention blocks. Since incorporating RUDP into HAN would require significant modifications to the model's structure, we apply only LP-based detail control, excluding RUDP. Consequently, we conduct experiments by adding only a block that generates a detail image to the final upsample process. DRLN applies attention within the dense residual Laplacian module, which overlaps several times to form a cascading block. The entire model is composed of several such cascading blocks. Therefore, we integrate RUDP by inserting the upscale and downscale blocks from LaUD between some of these cascading blocks. This enables us to apply detail loss and RUDP while keeping the structure of the original attention mechanism.

4 EXPERIMENTS

In this section, we first compare the performance of LaUD with SOTA models. Next, we provide ablation studies and qualitative analysis on LaUD to validate the effects of LP-based detail control and RUDP. Finally, we demonstrate the impact of our methods when combined with attention-based models. Results on more images can be found in the appendix. Due to page limit, the detailed setup for the training and evaluation of LaUD and attention-based models is provided in the appendix. All our implementation code will be released and made publicly available.

Scale	Methods	Base	Set5		Set14		BSD100		Urban100	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2	EDSR (Lim et al., 2017)	CNN	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351
	MWCNN (Liu et al., 2018)		37.91	0.9600	33.70	0.9182	32.23	0.8999	32.30	0.9296
	D-DBPN (Haris et al., 2021a)		38.09	0.9600	33.85	0.9190	32.27	0.9000	32.55	0.9324
	HBPN (Liu et al., 2019a)		38.13	0.961	33.78	0.921	32.33	0.902	33.12	0.938
	RCAN (Zhang et al., 2018)	Attention	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384
	DRLN (Anwar & Barnes, 2022)		38.27	0.9616	34.28	0.9231	32.44	0.9028	33.37	0.9390
	HAN† (Niu et al., 2020)		38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385
	EDT-B† (Li et al., 2021)		38.63	0.9632	34.80	0.9273	32.62	0.9052	34.27	0.9456
	SwinFIR† (Zhang et al., 2022)		38.65	0.9633	34.93	0.9276	32.64	0.9054	34.57	0.9473
	HAT-L† (Chen et al., 2023)		38.91	0.9646	35.29	0.9293	32.74	0.9066	35.09	0.9505
	LaUD(ours)†	CNN	38.45	0.9625	34.65	0.9256	32.54	0.9042	33.71	0.9507
×4	EDSR (Lim et al., 2017)	CNN	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033
	MWCNN (Liu et al., 2018)		32.12	0.8941	28.41	0.7816	27.62	0.7355	26.27	0.7890
	D-DBPN (Haris et al., 2021a)		32.47	0.8980	28.82	0.7860	27.72	0.7400	26.38	0.7946
	HBPN (Liu et al., 2019a)		32.55	0.900	28.67	0.785	27.77	0.743	27.30	0.818
	RCAN (Zhang et al., 2018)	Attention	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087
	DRLN (Anwar & Barnes, 2022)		32.63	0.9002	28.94	0.7900	27.83	0.7444	26.98	0.8119
	ABPN (Liu et al., 2019b)		32.69	0.900	28.94	0.789	27.82	0.743	27.06	0.811
	HAN† (Niu et al., 2020)		32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094
	EDT-B† (Li et al., 2021)		33.06	0.9055	29.23	0.7971	27.99	0.7510	27.75	0.8317
	SwinFIR† (Zhang et al., 2022)		33.20	0.9068	29.36	0.7993	28.03	0.7520	28.12	0.8393
	HAT-L† (Chen et al., 2023)		33.30	0.9083	29.47	0.8015	28.09	0.7551	28.60	0.8498
	LaUD(ours)†	CNN	32.81	0.9020	29.05	0.7937	27.88	0.7471	27.20	0.8174
×8	EDSR (Lim et al., 2017)	CNN	26.96	0.7762	24.91	0.6420	24.81	0.5985	22.51	0.6221
	D-DBPN (Haris et al., 2021a)		27.21	0.7840	25.13	0.6480	24.88	0.6010	22.73	0.6312
	HBPN (Liu et al., 2019a)		27.17	0.785	24.96	0.642	24.93	0.602	23.04	0.647
	RCAN (Zhang et al., 2018)	Attention	27.31	0.7878	25.23	0.6511	24.98	0.6058	23.00	0.6452
	DRLN (Anwar & Barnes, 2022)		27.36	0.7882	25.34	0.6531	25.01	0.6057	23.06	0.6471
	ABPN (Liu et al., 2019b)		27.25	0.786	25.08	0.638	24.99	0.604	23.04	0.641
	HAN (Niu et al., 2020)		27.33	0.7884	25.24	0.6510	24.98	0.6059	22.98	0.6437
	LaUD(ours)†	CNN	27.51	0.7882	25.34	0.6569	25.04	0.6102	22.07	0.5898

Table 1: Quantitative comparison with state-of-the-art methods on benchmark datasets.

4.1 Performance Analysis of Our Model LaUD

Table 1 presents a quantitative comparison between LaUD and SOTA models. Following the standard conventions in the field, we conduct experiments using four datasets: Set5 (Bevilacqua et al., 2012), Set14 (Zeyde et al., 2012), BSD100 (Martin et al., 2001), and Urban100 (Huang et al., 2015). We evaluate the PSNR and SSIM values for $2\times$, $4\times$, and $8\times$ upscaling. However, since some papers do not report $8\times$ upscaling results, we include only the reported results for $8\times$ upscaling. The PSNR and SSIM values are calculated on the Y channel from the YCbCr space. In the table, "†" is used to indicate models that execute two training sessions: pretraining and fine-tuning.

Considering the overall PSNR results, LaUD outperforms all CNN-based models, except in the $8\times$ upscaling on Urban100. Among CNN-based models, DBPN and HBPN use the back-projection method, which is similar to our RUDP. While these models perform well within CNN-based architectures, LaUD achieves better results and highlights the effectiveness of LP-based detail loss. In addition, our LaUD demonstrates performance comparable to attention-based models, despite being a CNN-based architecture. It outperforms RCAN, DRLN, ABPN, and HAN across all datasets, except for Urban100 at the $8\times$ scale. Surpassing models that employ attention mechanisms, which excel at feature extraction, clearly show that our model effectively extracts and utilizes features through the LP-based detail control and RUDP.

In detail, for the $2\times$ upscaling, LaUD improves by 0.18 dB on Set5 and 0.37 dB on Set14 compared to DRLN, which also aims to utilize the concept of the LP. Compared to DBPN and HBPN, which employ iterative upsampling and downsampling through back-projection structures similar to our RUDP, LaUD demonstrates significant performance improvements over both DBPN and HBPN, achieving gains of at least 0.8 dB on Set14 and 0.59 dB on Urban100. This indicates that our model, which incorporates the LP-based detail loss and RUDP, more effectively restores high-frequency data. The influence of detail control is maintained even as the scaling increases. For instance, with an $8\times$ scaling factor, LaUD achieves the PSNR values of 27.51 dB on Set5, 25.34 dB on Set14, and 25.04 dB on BSD100, outperforming all other models on these datasets. Since an LR image contains significantly less information compared to an $8\times$ SR image, it is challenging to generate an appropriate feature map from the LR image solely through the SR loss. In this context, catching the missing information from the $8\times$ upscale features through detail loss plays an important role.

4.2 Ablation studies on LaUD

No.	RUDP	Weighted SUM	Detail Control	Set5		Set14		BSD100		Urban100	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
1	X	X	X	38.1887	0.9615	34.0821	0.9212	32.3554	0.9023	32.7903	0.9422
2	O	X	X	38.1741	0.9613	34.0184	0.9206	32.3052	0.9013	32.5221	0.9393
3	O	O	X	38.3154	0.9620	34.6050	0.9250	32.4888	0.9037	33.6879	0.9497
4	X	X	O	38.2841	0.9619	34.2761	0.9224	32.4013	0.9029	33.0743	0.9448
5	O	X	O	38.2511	0.9618	34.2763	0.9224	32.4133	0.9030	33.1355	0.9453
6	O	O	O	38.4237	0.9625	34.7677	0.9256	32.5504	0.9045	34.0834	0.9529

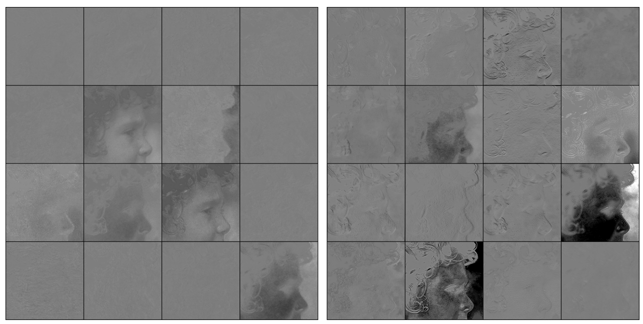
Table 2: Ablation for the LP detail control and RUDP (for $\times 2$). The term "Weighted Sum" refers to whether the loss is defined as the weighted sum of losses using each image generated during RUDP.

In this section, we conduct ablation studies to assess the impact of our LP-based detail control and RUDP. In Section 3.1, we define the total loss for LaUD using the weighted sum of the losses for each SR image. However, the total loss can also be determined using only the final SR image. Therefore, we investigate the effect of the weighted sum as well. Eventually, we examine six scenarios considering RUDP, the weighted sum of the losses, and LP detail control. We train each model for the six scenarios only once on ImageNet. To minimize the influence of model complexity, we adjust the architectures to have a similar number of parameters across models. The experiment focuses solely on $2\times$ scaling. The results are presented in Table 2.

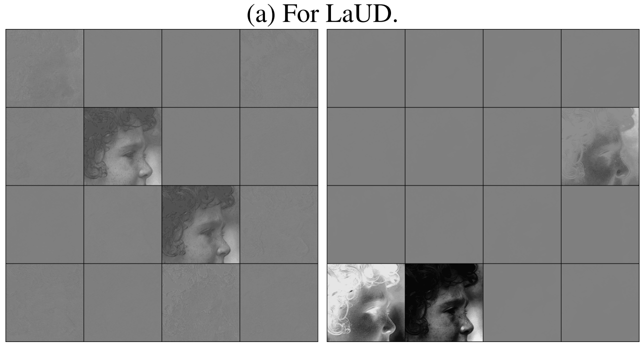
The results in Table 2 demonstrate the impact of the LP-based detail loss and the synergistic effect when RUDP is applied simultaneously. When comparing models No. 1 and 4, 2 and 5, and 3 and 6, the models incorporating our detail control consistently outperform the others. This indicates that guiding the model with detail loss is an effective approach, especially in enhancing the high-frequency components of the SR image. When RUDP is applied alone, it seems to interfere with the model's training. For instance of Set5, the model No. 1, which employs no additional methods, achieves 38.1887 dB, whereas the model No. 2, using only RUDP, achieves 38.1741 dB. However, the model No. 3, which incorporates a weighted sum of losses using intermediate SR images, improves performance to 38.3154 dB. By using images from the intermediate layer as loss, the model generates accurate SR images at that stage. This approach helps guide the model to extract more appropriate features in RUDP and progressively refine the SR image in subsequent steps.

Notably, the model that applies all methods achieves the highest performance, with a score of 38.4237 dB on Set5. This model No. 6 shows a significant improvement of approximately 0.23 dB on Set5 over the model without any methods. This substantial difference demonstrates that detail control and RUDP with a weighted sum complement each other, resulting in a synergistic effect. While detail control guides the model to focus on high-frequency components, deficiencies are compensated by re-extracting features from the SR image of the previous step through RUDP.

4.3 Qualitative Analysis of LaUD



(a) For LaUD.



(b) For LaUD without detail loss.

Figure 2: Parts of the SR feature map: (a) for LaUD, and (b) for LaUD without detail loss. In each set, the left image shows the first upscaling, while the right image shows the last upscaling in RUDP.

We aim to analyze the outputs of our model. Figure 2 illustrates the SR feature maps of LaUD. Before the final ToRGB layers, the feature maps consist of 256 channels; however, in the figure, we zoom in on the last 16 channels to highlight the changes more clearly. The full and different images of Set5 can be found in the appendix.

The figure shows two feature maps: the first and last SR features in RUDP. In both cases (a) and (b), the final SR features contain more activated channels. Additionally, the contrast between the channels in the final feature map is more clearly distinguished. This is because the model enhances information for SR through the processes of upscaling and downscaling. When comparing the feature maps between (a) and (b), we see an obvious fact that detail loss affects the diversity of feature maps. As shown in the feature map of (b), if there is no guidance for the model to capture high-frequency information, RUDP amplifies only a few prominent channels, keeping the values in most channels close to zero. Conversely, in (a), even in the first SR feature map, shapes containing texture are revealed in many channels. Notably, in the final feature map, this texture is further enhanced. As a result, each channel conveys distinct and clear texture information. This difference highlights the importance of guiding high-frequency information by the detail loss in SR tasks. Furthermore, it demonstrates that RUDP, when combined with LP detail control, provides significant benefits by efficiently extracting diverse information for high-frequency components.

We also provide the visual comparison in Figure 3. We compare the SR images produced by LaUD with those generated by other SOTA models on Urban100. Urban100 consists of images where structural information is crucial, such as buildings with numerous windows or spiral staircases converging to a point. This comparison allows us to evaluate whether the model can accurately identify and reproduce repetitive structures down to the fine details in the SR process.

Figure 3 presents the results of D-DBPN, DRLN, SwinFIR, and our LaUD across two images. We report additional examples in the appendix. In the first image, a closer inspection of the patches reveals differences in the wall's detailed texture. Our LaUD achieves the highest performance, with 34.5614 dB. It produces an image closer to the HR by generating a texture that resembles dust along the line that separates windows at the bottom of the patches. In contrast, DBPN and SwinFIR create images with clean lines in that area but fail to capture finer details. We think that our detail loss allows the model to focus more effectively on these intricate textures. The image "image072" features a pattern of circular lines. The enlarged red box highlights the area where straight lines intersect. All models render these lines without distortion. However, when inspecting the diagonal line from the top left to the bottom right, our model more clearly distinguishes the boundary between the black and white lines compared to other models. This is because our model enhances the contrast in this area, similar to the level of the HR image. This result is achieved by capturing the boundary with LP-based detail loss and enhancing pixels through RUDP. Our PSNR is the best here as well.

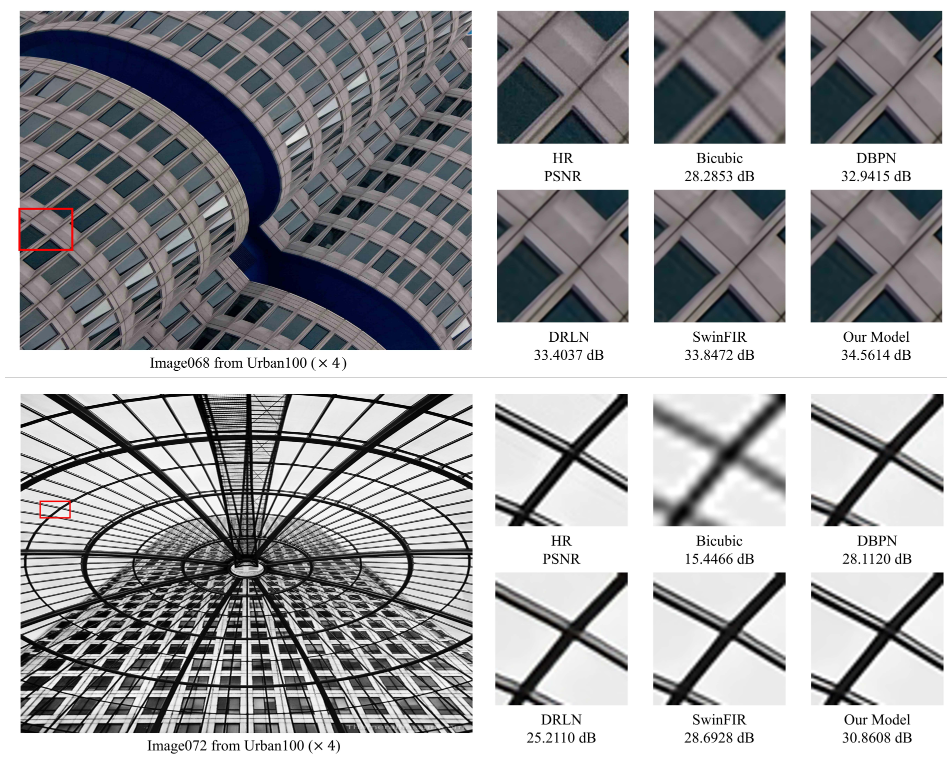


Figure 3: Visual comparison for $\times 4$ SR on Urban100. The patches for comparison are marked with red boxes in the original images. The PSNR values are calculated based on the patches.

4.4 Application to Attention-based Models

This section presents the results of applying our LP-based detail loss to attention-based models. As outlined in Section 3.2, we selected three models: DRLN, HAN, and ABPN. Depending on the structure of each model, we applied either the detail loss alone or together with RUDP.

Table 3 presents our experimental results. We reproduced all the original models using the code provided in each paper. However, the results did not match the values reported in the respective papers. Despite this discrepancy, a valid comparison is still possible, since the original models and those incorporating our method were trained in the same environment and under identical conditions.

Scale	Model	Set5		Set14		BSD100		Urban100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2	DRLN	38.0971	0.9610	33.7461	0.9188	32.2428	0.9006	32.0521	0.9353
	DRLN + Detail loss + RUDP	38.2657	0.9618	34.1755	0.9223	32.4246	0.9029	33.0870	0.9450
	HAN	38.2759	0.9616	34.1278	0.9218	32.3898	0.9027	32.9821	0.9443
	HAN + Detail loss	38.2941	0.9617	34.1416	0.9219	32.4123	0.9030	33.0814	0.9451
×4	ABPN	32.2792	0.8955	28.6666	0.7828	27.6110	0.7379	25.3536	0.7646
	ABPN + Detail loss	32.4739	0.8980	28.7962	0.7861	27.6948	0.7416	25.7796	0.7797

Table 3: The performance of attention-based models applying our detail loss or RUDP.

We evaluated all models using various test datasets, and the results exhibited consistent tendencies. Therefore, a detailed examination of the results only for Set5 is as follows. First, DRLN significantly improves performance by applying both our LP-based detail loss and RUDP with a weighted sum loss using SR and detail images. The reproduced DRLN achieves 38.0971 dB, while the model with our methods records 38.2657 dB, showing an improvement of approximately 0.17 dB. DRLN is well-suited for introducing RUDP with the weighted sum loss due to its multi-block structure. This further enhances the effect of our detail loss.

We only apply the LP-based detail loss to HAN, as adding RUDP poses a risk of significantly altering the structure. This results in a slight improvement from 38.2759 dB to 38.2941 dB. While the improvement is small, the detail loss still has an impact on the model. Since HAN uses RCAN as a pretrained model, the sub-pixel convolution for the SR image is also pretrained. However, the part responsible for generating the detail image must be trained with a small learning rate without pretraining. This likely explains why the PSNR value does not show a more significant difference.

Performance improvement is also observed in ABPN. ABPN has a structure similar to RUDP but generates an SR image by collecting all SR features produced during the mid-process. As a result, we are unable to introduce RUDP and instead integrate only our detail loss. With the addition of our detail loss, PSNR improves by approximately 0.2 dB, and SSIM increases from 0.8955 to 0.8980.

In summary, across all three models and all datasets, combining the attention-based model with our detail loss leads to performance improvements. The result demonstrates that our LP-based detail loss is not limited to CNN structures but can be effectively integrated with attention mechanisms to enhance a model.

5 Conclusions

In this paper, we proposed a novel detail loss based on the LP and a RUDP for the SR task. The LP-based detail loss can be used with CNN models and transformers, as it is independent of the model's architecture. In addition, when combined with RUDP, the LP-based detail loss produces a synergistic effect, significantly improving the performance. Qualitative analysis shows that the detail loss helps the model capture high-frequency information, resulting in many channels in the SR feature map conveying different texture information (cf. Figure 2). In our experiments, we constructed a CNN-based model incorporating both the LP detail loss and RUDP. Through ablation studies, we confirmed the effectiveness of each technique (cf. Table 2). Additionally, we evaluated our CNN model on four datasets using PSNR and SSIM metrics (cf. Table 1). The model outperformed all other CNN models and performs better than several attention-based models. Moreover, we integrated our method into several existing attention-based models, resulting in improved performance across all of them (cf. Table 3). This demonstrates that the LP-based detail loss is effective with attention mechanisms and applicable regardless of model structure.

Acknowledgments

REFERENCES

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Saeed Anwar and Nick Barnes. Densely residual Laplacian super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1192–1204, 2022. doi: 10.1109/TPAMI.2020.3021088.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the 23rd British Machine Vision Conference (BMVC)*, pp. 135.1–135.10, July 2012. ISBN 1-901725-46-4.

Peter J. Burt and Edward H. Adelson. The Laplacian pyramid as a compact image code. *Readings in Computer Vision*, pp. 671–679, 1987.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22367–22377, June 2023.

Julien Cornebise, Ivan Oršolić, and Freddie Kalaitzis. Open high-resolution satellite imagery: The WorldStrat dataset –with application to super-resolution. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25979–25991. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a6fe99561d9eb9c90b322afe664587fd-Paper-Datasets_and_Benchmarks.pdf.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. doi: 10.1109/TPAMI.2015.2439281.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2360–2369, October 2021.

Lei Ge and Lei Dou. G-Loss: A loss function with gradient information for super-resolution. *Optik*, 280:170750, 2023. ISSN 0030-4026. doi: https://doi.org/10.1016/j.ijleo.2023.170750.

Hayit Greenspan. Super-resolution in medical imaging. *The Computer Journal*, 52(1):43–63, 2009. doi: 10.1093/comjnl/bxm075.

Sangjun Han, Taeil Hur, and Youngmi Hur. Laplacian Pyramid-like Autoencoder. In Kohei Arai (ed.), *Intelligent Computing*, pp. 59–78, Cham, 2022. Springer International Publishing. ISBN 978-3-031-10464-0.

Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for single image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4323–4337, 2021a. doi: 10.1109/TPAMI.2020.3002836.

Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-driven super resolution: Object detection in low-resolution images. In Teddy Mantoro, Minho Lee, Media Anugerah Ayu, Kok Wai Wong, and Achmad Nizar Hidayanto (eds.), *Neural Information Processing*, pp. 387–395, Cham, 2021b. Springer International Publishing.

Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-SRNet: A wavelet-based CNN for multi-scale face super resolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Zheng Hui, Jie Li, Xinbo Gao, and Xiumei Wang. Progressive perception-oriented network for single image super-resolution. *Information Sciences*, 546:769–786, 2021. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2020.08.114. URL https://www.sciencedirect.com/science/article/pii/S002002552030880X.

Jithin Saji Isaac and Ramesh Kulkarni. Super resolution techniques for medical image processing. In *2015 International Conference on Technologies for Sustainable Development (ICTSD)*, pp. 1–6, 2015. doi: 10.1109/ICTSD.2015.7095900.

Pranav Jeevan, Akella Srinidhi, Pasunuri Prathiba, and Amit Sethi. WaveMixSR: Resource-efficient neural network for image super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5884–5892, January 2024.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Proceedings of the European Conference on Computer Vision (ECCV) 2016*, pp. 694–711. Springer International Publishing, 2016. ISBN 978-3-319-46475-6.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016a.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016b.

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep Laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1929–1938, June 2022.

Wenbo Li, Xin Lu, Shengju Qian, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer-based image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1833–1844, October 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-CNN for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

Yuqing Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. Progressive multi-scale residual network for single image super-resolution. *arXiv preprint arXiv:2007.09552*, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, October 2021.

Zhi-Song Liu, Li-Wen Wang, Chu-Tak Li, and Wan-Chi Siu. Hierarchical back projection network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019a.

Zhi-Song Liu, Li-Wen Wang, Chu-Tak Li, Wan-Chi Siu, and Yui-Lam Chan. Image super-resolution via attention based back projection networks. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3517–3525, 2019b. doi: 10.1109/ICCVW.2019.00436.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pp. 416–423 vol.2, 2001. doi: 10.1109/ICCV.2001.937655.

Bokyoon Na and Geoffrey C Fox. Object detection by a super-resolution method and a convolutional neural networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2263–2269, 2018. doi: 10.1109/BigData.2018.8622135.

Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 191–207, Cham, 2020. Springer International Publishing.

George Seif and Dimitrios Androutsos. Edge-based loss function for single image super-resolution. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1468–1472, 2018. doi: 10.1109/ICASSP.2018.8461664.

Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Shane D. Sims. Frequency domain-based perceptual loss for super resolution. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2020. doi: 10.1109/MLSP49062.2020.9231718.

Rewa Sood, Binit Topiwala, Karthik Choutagunta, Rohit Sood, and Mirabela Rusu. An application of generative adversarial networks for super resolution medical imaging. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 326–331, 2018. doi: 10.1109/ICMLA.2018.00055.

Wanjie Sun and Zhenzhong Chen. Learned image downscaling for upscaling using content adaptive resampler. *IEEE Transactions on Image Processing*, 29:4027–4040, 2020. doi: 10.1109/TIP.2020.2970248.

Mohammed Ahmed Talab, Suryanti Awang, and Saif Al-din M. Najim. Super-low resolution face recognition using integrated efficient sub-pixel convolutional neural network (ESPCN) and convolutional neural network (CNN). In *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, pp. 331–335, 2019. doi: 10.1109/I2CACIS.2019.8825083.

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. NTIRE 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/touvron21a.html`.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxViT: Multi-axis vision transformer. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 459–479, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20053-3.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

Jinchang Xu, Yu Zhao, Yuan Dong, and Hongliang Bai. Fast and accurate image super-resolution using a combined loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

Bincheng Yang and Gangshan Wu. MaxSR: Image super-resolution using improved MaxViT. *arXiv preprint arXiv:2307.07240*, 2023.

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In Jean-Daniel Boissonnat, Patrick Chenin, Albert Cohen, Christian Gout, Tom Lyche, Marie-Laurence Mazure, and Larry Schumaker (eds.), *Curves and Surfaces*, pp. 711–730, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-27413-8.

Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. SwinFIR: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*, 2022.

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017. doi: 10.1109/TCI.2016.2644865.

A Appendix

A.1 Laplacian Pyramid

The Laplacian Pyramid (Burt & Adelson, 1987) is an image representation consisting of multi-scale high-frequency images and one low-frequency image of the smallest scale. This representation is similar to the Gaussian Pyramid presented in the same paper, but differs in that the LP comprises residual images except for the last level.

Figure 4 illustrates the overall process for constructing the LP as used in our paper. First, we obtain a downscaled image I_1 by low-pass filtering and downsampling an original image I_0. Then, I_1 is expanded to a re-upscaled image $I_{1\uparrow}$ through interpolation with the same size as I_0. We get a residual image ΔI_1 by subtracting the re-upscaled image $I_{1\uparrow}$ from the original image I_0. Consequently, the original image is decomposed into the approximation image I_1 and the detail image ΔI_1, forming the first level of the LP. Repeating this process to the approximation image, we create a pyramid composed of k multi-scale high-frequency images $\Delta I_1, \Delta I_2, \ldots, \Delta I_k$ and one low-frequency image I_k of the smallest size, after k steps. Since the construction process involves subtraction, the

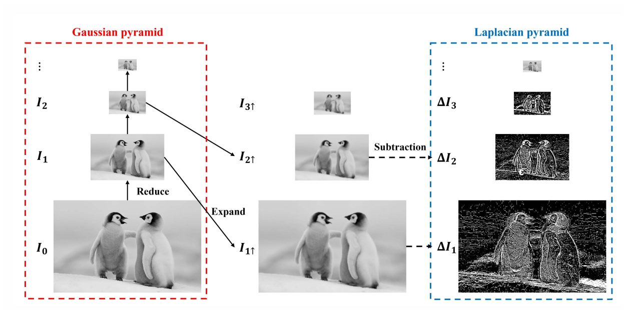


Figure 4: The construction process of the Laplacian Pyramid.

LP can completely reconstruct the HR image by adding a detail image and an upsampled approximation image of the same level. Therefore, the LP is a useful technique for image compression and reconstruction.

We consider the high-frequency image of LP appropriate for refining the detail part of the SR image. If a model generates an elaborate detail image of LP, the perfect reconstruction property of LP is operated efficiently to enhance the SR. From this point of view, we develop the LP-based detail control. It optimizes the model to generate a feature map containing high-frequency information through supervised learning with the LP detail of the HR image.

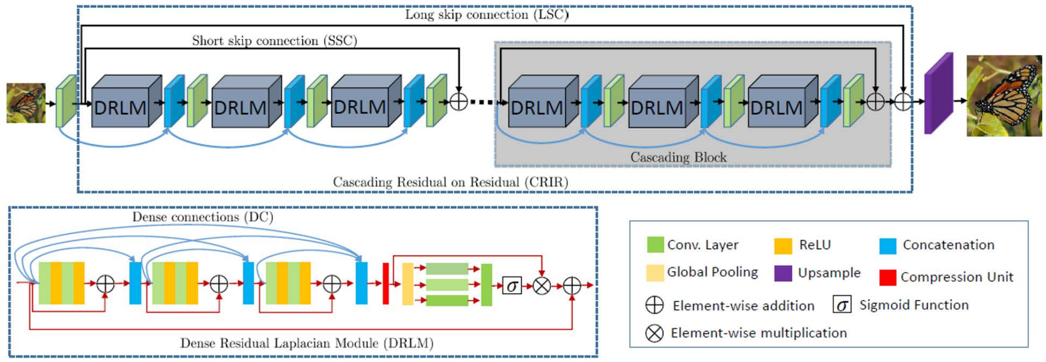
The ground truth detail image used in our LaUD is identical to the largest detail image in the LP process. Specifically, by selecting the ground truth HR image as I_0 in Figure 4, ΔI_1 is generated through the LP process. This ΔI_1 corresponds to I_D in our loss function, as defined in Equation (7). Since the LP process supports multi-scale analysis, a stepwise upscaling approach can be applied to tackle higher-scale SR problems. For example, in a $4\times$ upscaling problem, the process could be divided into two stages: first performing $2\times$ upscaling as an intermediate step, followed by another $2\times$ upscaling in the final step. While various alternative approaches exist, we opted to perform the entire $4\times$ upscaling in a single step to simplify the model. As a result, whether addressing a $4\times$ or $8\times$ SR problem, the I_D in Equation (7) remains equivalent to ΔI_1.

A.2 Ready For Applying Our Methods to Attention-based Models

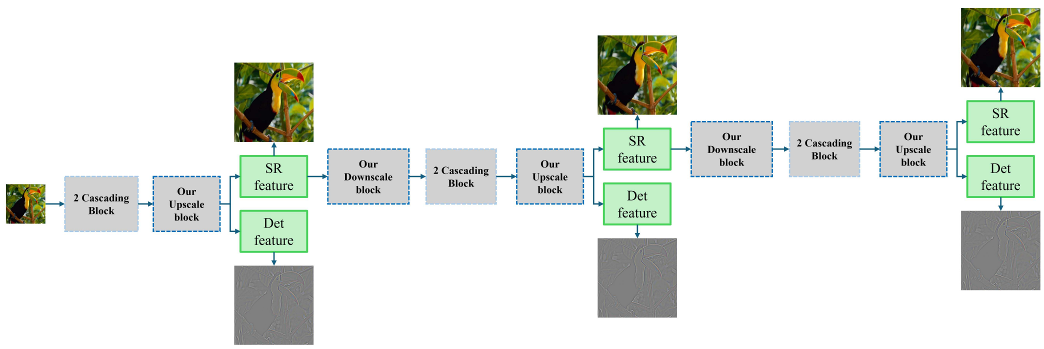
As previously mentioned in the main context, small modifications are required to adapt our method to the three selected attention-based models: DRLN, HAN, and ABPN.

Figure 5 shows the original DRLN structure and modified version for applying our methods. The image for the original DRLN is taken from the paper Anwar & Barnes (2022). DRLN consists of cascading blocks, each containing multiple Dense Residual Laplacian Module (DRLM). According to the author's code, the original DRLN structure passes through a total of three cascading blocks, with a short skip connection after each block. After each cascading block and short skip connection, we apply the LP-based detail loss and RUDP by incorporating our upscale and downscale blocks. As a result, we modify the model to the structure shown in (b), without altering the attention mechanism within the DRLM. With the introduction of three upscale blocks, we generate three SR images and three detail images, similar to LaUD. The total loss is calculated as a weighted sum using these images.

Figure 6 shows the original HAN structure and modified version for applying our methods. The image for the original HAN is taken from the paper Niu et al. (2020). HAN extracts features through residual groups and then applies layer attention and channel-spatial attention to these features. Since channel attention is also present within the residual groups, the attention mechanism would need to be disrupted to introduce RUDP in the feature extraction phase. Consequently, we choose not to apply RUDP and instead train the model using only the LP-based detail loss. Upon reviewing the code, we confirm that HAN uses RCAN as its pretrained base. This causes insufficient training when the upscale block of LaUD is used instead of the original sub-pixel convolution. Therefore, we continue using sub-pixel convolution for upsampling and add an additional sub-pixel convolution layer to generate the detail image. The resulting modified version is shown in (b).



(a) The original DRLN taken from Anwar & Barnes (2022).



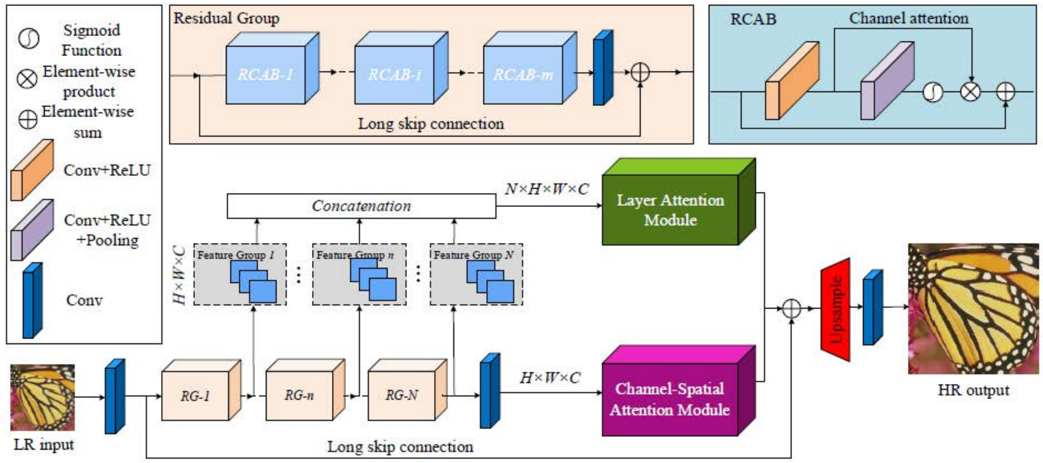
(b) The modified DRLN.

Figure 5: The original DRLN structure and modified version for applying our methods.

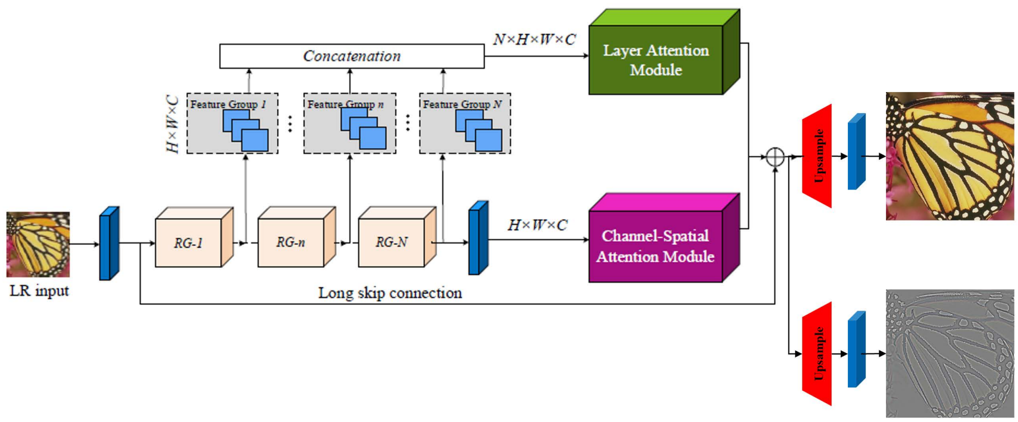
Figure 7 shows the original ABPN structure and modified version for applying our methods. The image for the original ABPN is taken from the paper Liu et al. (2019b). ABPN has a structure that repeatedly performs upsampling and downsampling, with the attention mechanism applied after the downsampling back-projection block. Therefore, we replace the original upsampling and downsampling blocks with the upscale and downscale blocks from LaUD. Since our upscale block generates both detail and SR features, we can define the detail loss naturally. However, the original ABPN follows a complex process to produce an SR image, where concatenated SR and LR features are convolved and then added to the bicubic-upsampled LR. Considering whether to apply this process to the generation of a detail image, we determine that it could lead to incorrect changes, such as requiring a downsampled version of the detail. Therefore, we opt for a simpler structure that gathers detail features and generates a detail image through convolution, as shown in (b). Unfortunately, due to the process of image generation, while RUDP is used, multiple images are not generated. As a result, it is not possible to construct a weighted sum loss using multiple images. Based on the ablation results of LaUD, applying only RUDP without the weighted sum loss tends to interfere with the model. We think that the performance of the modified model may be limited.

A.3 Experimental Setup for LaUD

In this section, we specify the training setting of LaUD. The repeated upscaling and downscaling structure requires setting several hyperparameters. As mentioned earlier, we apply three upscaling steps. The weights in L_s and L_d, $\{W_k\}_{k=1,2,3}$, are set to 1, 3, and 10, respectively, for progressive advancement. The weights between L_s and L_d, α and β, are each set to 1 to ensure the model focuses sufficiently on the detail image. Our choice for $\{W_k\}_{k=1,2,3}$ has not been completely optimized through a systematic process. However, we selected the weight that demonstrated the highest performance among the several comparative experiments we conducted.



(a) The original HAN taken from Niu et al. (2020).



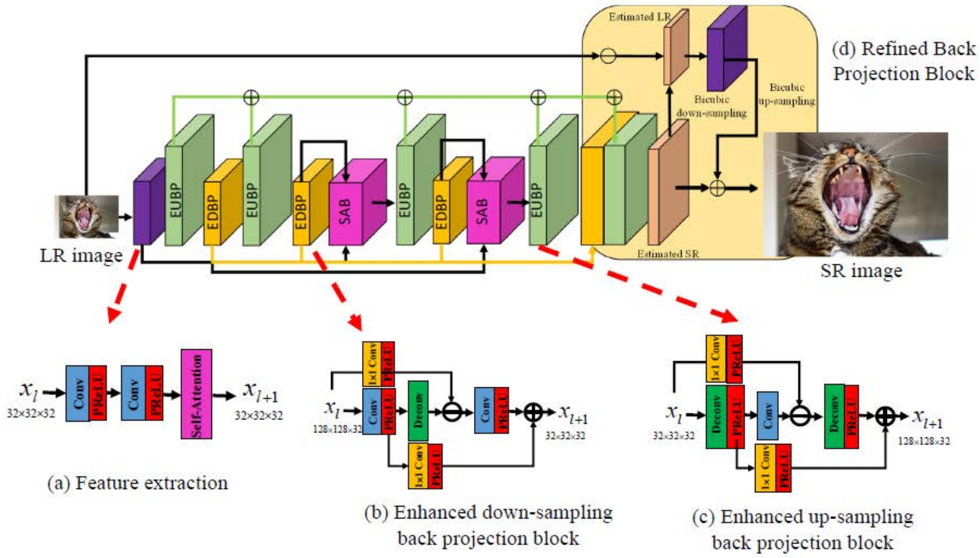
(b) The modified HAN.

Figure 6: The original HAN structure and modified version for applying our methods.

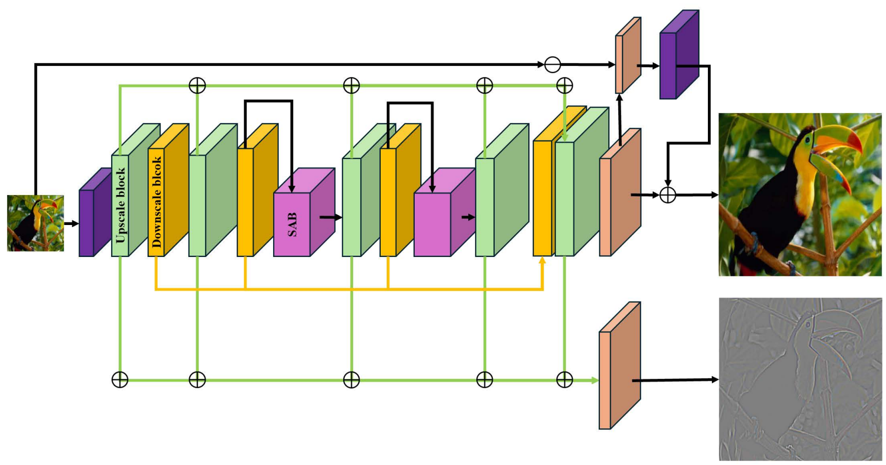
We execute two training sessions, as shown in HAT (Chen et al., 2023), SwinFIR (Zhang et al., 2022), and EDT (Li et al., 2021): pretraining on ImageNet 2012 (Deng et al., 2009) and fine-tuning on DIV2K (Agustsson & Timofte, 2017) + Flickr2K (Timofte et al., 2017). For both training sessions, the number of training epochs, initial learning rate, and learning rate schedules are based on previous studies. The model shows significant performance with just pretraining, but we improve slightly by fine-tuning with higher-resolution images.

The hyperparameters used in the two training sessions are similar to those in previous studies, such as Liu et al. (2019a); Anwar & Barnes (2022); Hui et al. (2021); Chen et al. (2023); Zhang et al. (2022); Li et al. (2021) In the pretraining stage, we resize the images in ImageNet to 224×224 and randomly crop them to 128×128. The augmented images are the HR we must fit, and LR images are produced with size $128/s \times 128/s$ through the bicubic interpolation according to the scaling rate s. We set the initial learning rate to $2 \cdot 10^{-4}$ and train models for 25 epochs. The learning rate is halved at 50%, 80%, 90%, and 96% of the total epochs.

For fine-tuning, we combine DIV2K and Flickr2K as training data. Since these images are huge, resizing significantly compromises their quality. Therefore, unlike in pretraining, we only perform random cropping. However, the crop size of 128×128, as used in pretraining, often makes images contain no objects. Such images negatively affect the model's performance after fine-tuning. To



(a) The original ABPN taken from Liu et al. (2019b).



(b) The modified ABPN.

Figure 7: The original ABPN structure and modified version for applying our methods.

address this, we increase the crop size to 512×512 until maximum of our resources. Fine-tuning is conducted during 1000 epochs, with an initial learning rate of 10^{-5}, halved at 50%, 80%, 90%, and 95% of the training progress.

A.4 Experimental Setup for Modified Attention-based Model

When training the modified attention-based models using our method, most settings follow the original paper or the default configurations from the original code. As these settings are the same when reproducing the original model, we briefly summarize them in here and recommend referring to the original paper for details.

DRLN was trained for 3000 epochs with a batch size of 16 on a dataset combining DIV2K and Flickr2K. During training, images were randomly cropped to 48×48 for LR and $(48 * s) \times (48 * s)$

for HR, where s is the scaling factor. Random horizontal flips, vertical flips, and 90-degree rotations were applied as data augmentation. The initial learning rate was set to 10^{-4} and halved every 200 epochs. When our LP-based detail loss and RUDP were applied to DRLN, the weights for the weighted sum of the losses were set exactly as in LaUD.

For HAN, training was conducted using images in the 0-255 range from the DIV2K dataset. Since RCAN was used as a pretrained model, only 400 epochs of training were performed with a batch size of 16. The learning rate setup, the cropped LR and HR sizes, and the data augmentation were identical to those used in DRLN. When our method was applied, RUDP could not be used, so we only needed to set the weights between the SR loss and detail loss, which were kept at a 1:1 ratio, the same as in LaUD.

Finally, ABPN differs slightly from the previous two models because the smallest scaling factor is 4. The training data is DIV2K + Flickr2K, and the model is trained for 5000 epochs with a batch size of 16. However, the HR image size is set to 160×160, and the LR size is 40×40. Only random horizontal and vertical flips are applied as augmentation. The initial learning rate is set to 10^{-4}, the same as in the previous two models, but it is halved only once at 2500 epochs. Unfortunately, when we applied our method, we were unable to incorporate the weight sum connected to RUDP. As a result, only the weight for SR loss and detail loss were set as a 1:1 ratio. However, by replacing the existing upsample and downsample back projection blocks with LaUD's upscale and downscale blocks, the number of model parameters is reduced by half. To minimize the impact of the model size, we compensated by increasing the number of feature maps generated in the intermediate layers. Consequently, in our experiment, the ABPN and ABPN with our methods had nearly the same number of parameters.

A.5 Additional Ablation Studies

In this section, we present additional ablation studies for our two methods: the LP-based detail loss and RUDP. These experiments were conducted using our LaUD model.

A.5.1 Ablation on Upscaling and Downscaling Repetitions

Number of RUDP	Set5 PSNR	Set5 SSIM	Set14 PSNR	Set14 SSIM	BSD100 PSNR	BSD100 SSIM	Urban100 PSNR	Urban100 SSIM
1	38.1402	0.9613	34.0151	0.9209	32.3038	0.9015	32.5285	0.9396
2	38.2802	0.9619	34.4278	0.9238	32.4483	0.9035	33.4187	0.9476
3	38.4237	0.9625	34.7677	0.9256	32.5504	0.9045	34.0834	0.9529

Table 4: The performance variations with different numbers of RUDP. All models are designed for the $2\times$ super-resolution problem.

The number of upscaling and downscaling repetitions should be treated as a hyperparameter. In LaUD, this hyperparameter was set to three upscaling repetitions, which were determined through extensive experimentation. Table 4 presents the results of the model based on the Number of RUDP. The Number of RUDP refers to the number of upscaling steps when applying RUDP. We evaluated performance by increasing the number of upscaling steps to 1, 2, and 3 in LaUD. All models used in the experiment were trained once on the ImageNet dataset with our LP-based detail loss and weighted sum loss.

In summary, increasing the number of RUDP consistently resulted in higher PSNR and SSIM values across all datasets. For PSNR, an improvement of at least 0.1 dB was observed in every case as the number of RUDPs increased. Notably, in Urban100, increasing the number of RUDP from 1 to 2 led to a significant improvement of nearly 1 dB. Although the trained model was lost and could not be recorded, performance improvements became minimal when the number of RUDP exceeded 4. In some instances, the model even demonstrated inferior performance. Furthermore, increasing the number of RUDP significantly raised the time and memory required for training. Based on these experimental results, we aimed to determine the number of RUDP that could achieve fine performance within the limitations of our resources. Consequently, our LaUD described in the main text was configured to proceed with three upscaling processes.

A.5.2 Comparison of Loss Functions

Loss SR	Loss Det	Set5 PSNR	Set5 SSIM	Set14 PSNR	Set14 SSIM	BSD100 PSNR	BSD100 SSIM	Urban100 PSNR	Urban100 SSIM
L_1		38.3154	0.9620	34.6050	0.9250	32.4888	0.9037	33.6879	0.9497
L_1	L_2	38.3899	0.9627	34.5614	0.9252	32.5058	0.9041	33.8258	0.9514
L_1	L_1	38.4237	0.9625	34.7677	0.9256	32.5504	0.9045	34.0834	0.9529

Table 5: The performance across different loss functions. All models are designed for the $2\times$ super-resolution problem.

Although we did not directly compare variations of the loss function in the main text, the ablation study (cf. Table 2) provides valuable insight into the differences between the standard L_1 loss for SR, commonly employed in SR tasks, and our proposed loss function, which combines the L_1 loss for SR with our LP-based detail loss. As shown in the comparisons between No. 1 and No. 4, No. 2 and No. 5, as well as No. 3 and No. 6 in Table 2, the models employing the combined loss function consistently achieved higher performance. For clarity, we report again the results of our LaUD without the LP-based detail loss and with the LP-based detail loss in the first and third rows of Table 5, respectively.

In addition, we conducted an additional experiment with a slight modification to our combined loss function, as shown in the second row of Table 5. Specifically, this experiment involved a model that retained the L_1 loss for SR but replaced the L_1 loss for detail with an L_2 loss. As demonstrated in Table 5, using the L_1 loss for detail resulted in higher performance compared to the L_2 loss, except for one SSIM value on Set5.

A.5.3 Comparison of Model Complexity

Model	Number of Parameters (M)	Memory Usage (MiB)	Training Time (s)	Training Time per Iteration (s)	Inference Time (ms)
EDSR	40.73	1348.00	45.0799	0.0451	6.940
D-DBPN	5.95	1350.00	27.4748	0.0275	9.934
RCAN	15.44	1714.00	174.1959	0.1742	43.491
DRLN	34.43	1952.00	63.2655	0.0633	18.058
HAN	15.92	1856.00	172.0624	0.1721	44.417
EDT-B	11.48	8070.00	383.7992	0.3838	84.582
SwinFIR	14.35	5742.00	161.3138	0.1613	66.072
HAT-L	40.70	13146.00	331.3637	0.3314	90.863
LaUD	29.33	1982.00	53.5079	0.0535	11.951

Table 6: A comparison of model size, memory usage, training time, and inference time between LaUD and the other state-of-the-art models. For measurement units, M represents a million and MiB denotes a mebibyte.

LaUD was a model designed without the intention of introducing particularly complex techniques, aside from the LP-based detail loss and RUDP. However, increasing the number of RUDP naturally raised the model's complexity due to the repeated upscaling process. To assess this, we aimed to compare the complexity of LaUD with existing state-of-the-art models. Table 6 presents the model size, memory usage, training time, and inference time for LaUD and SOTA models. From the models listed in Table 1, we selected those that required no modifications, as their configuration and model construction code were publicly available. All experiments were conducted under consistent conditions, and the code used for these experiments will be released on GitHub at a later date.

Memory usage, training time, and training time per iteration were measured using an input image with a size of $2 \times 3 \times 64 \times 64$. Generally, larger batch sizes are used during training, so a high batch size was initially considered for measurement. However, for models such as EDT-B, SwinFIR, and HAT-L, the memory requirements exceeded our resource limits. Consequently, the batch size was standardized to 2 for all models. Inference time was measured using an input image with a size of $1 \times 3 \times 64 \times 64$. Before all measurements for time, 100 warm-up iterations were performed. For training time, the duration of 1,000 training iterations was measured.

When comparing model sizes, LaUD has 29.33 million parameters, ranking fourth after EDSR, DRLN, and HAT-L. Excluding the top three attention-based models—EDT-B, SwinFIR, and HAT-L—which require substantial memory, LaUD occupies a middle position. Furthermore, considering that LaUD achieves the best performance among models outside the top three, its size can be regarded as relatively reasonable.

For memory usage, the top three models demand an overwhelmingly large amount of memory. In contrast, other models, including LaUD, operate within memory constraints that are not a concern. LaUD occupied 1982 MiB to process an input image of size $2 \times 3 \times 64 \times 64$, which is comparable to models such as RCAN, DRLN, and HAN.

In terms of training time, EDT-B and HAT-L had the longest durations, averaging around 350 seconds. RCAN, HAN, and SwinFIR followed, taking approximately half that time. LaUD, however, demonstrated significantly faster training at just 53.51 seconds, emphasizing the simplicity of the model. The overall trend is similar for inference time. Notably, LaUD required only 11.951 milliseconds, comparable to DBPN, which has a much smaller model size.

The results in Table 6 highlight that LaUD is a simple model. We believe its ability to outperform all but the top three models while maintaining a relatively small size indirectly demonstrates the effectiveness of the LP-based detail loss and RUDP.

A.6 Additional Image Results

A.6.1 Results for LaUD on Diverse Dataset

Figure 8 shows two additional results of LaUD. In all cases, the PSNR value increases along the progress of RUDP.

A.6.2 Analysis of the Role of Upscaled Feature and Detail Feature in LaUD

As explained in the main text, when designing LaUD, we derived a detail feature H_{D_k} from the upscaled feature H_{U_k} and combined them to form the SR feature H_{SR_k}. However, H_{D_k} exhibits distinct characteristics compared to H_{U_k}, since both H_{SR_k} and H_{D_k} are guided respectively by L_1 loss and detail loss. Through our comparison and analysis of feature maps, we observed that H_{D_k} frequently captures information about boundaries and textures. From this perspective, when combined with H_{U_k}, H_{D_k} enhances the information in H_{U_k} and helps adjust overly flat or overly emphasized values.

To illustrate our analysis, we present an image from the Set5 dataset as an example. Figure 9 displays the feature maps generated during the $2\times$ super-resolution process of LaUD. The upper-left corresponds to H_{U_3}, the upper-right corresponds to H_{D_3}, and the lower-left corresponds to H_{SR_3}. Examining this figure, the upscaled feature map H_{U_3} predominantly retains low-frequency information, such as complete object structures, and consists of features with varying contrasts. On the other hand, the detail feature map H_{D_3}, which generally has smaller values, tends to exhibit relatively flat distributions. Nevertheless, it often highlights distinct boundaries or textures that are absent in the upscaled features.

Figure 9 provides an overview of the changes that occur during the creation of the SR feature by combining upscaled and detailed features. Specifically, we now focus on a detailed comparison of the two cases highlighted by the red and green boxes. Figure 10 presents an enlarged view of the features within the red and green boxes from Figure 9. Each row, from left to right, corresponds to the upscaled feature, detail feature, and SR feature, respectively.

In the case of the red box (top row), the detail feature reveals prominent boundaries and textures. As a result, when the SR feature is formed by combining the detail feature with the upscaled feature, the insufficient high-frequency information in the upscaled feature is effectively reinforced. For example, compared to the upscaled features, the SR feature exhibits more distinct facial lines, stronger emphasis around the eyes and forehead, and newly introduced textures in the temple and along the

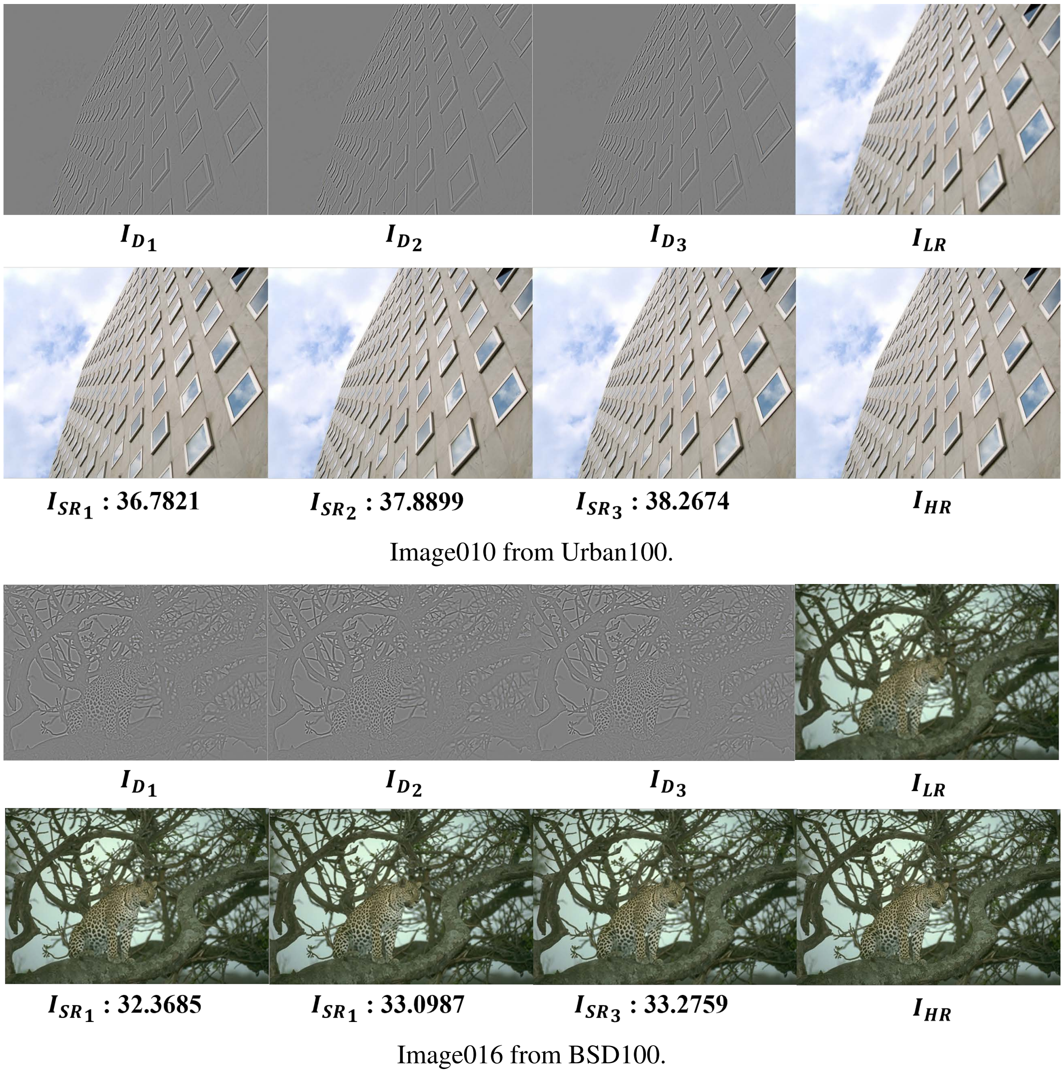


Figure 8: Two additional examples of LaUD. Each example is taken from Urban100 and BSD100, respectively. In the first row, there are three detail images by RUDP and a low-resolution image. In the second row, there are SR images by RUDP and a high-resolution image.

sides of the nose. Next, in the case of the green box (bottom row), the detailed feature serves a distinct role, unlike in the red box. In the upscaled feature, the values are generally flat, producing a hazy image. However, this flatness is somewhat corrected by incorporating the detailed feature. As a result, the SR feature demonstrates greater value curvature and seems to capture a more dynamic and lively appearance.

A.6.3 Additional Images for Feature Map Analysis of LaUD

In Figure 2 of the main text, we present a part of the SR feature map, highlighting the impact of LP-based detail loss. Figure 11 and Figure 12 below expand on this analysis by displaying not only the part but all 256 channels using examples on Set5.

The tendency is consistent with what is described in the main text. In both cases (a) and (b), more apparent feature map is generated as the upscaling process is repeated by RUDP. In particular, in case (a), where LP-based detail loss is applied, more channels are activated at the first upscaling compared to case (b). When comparing the right SR feature map (the third upscaled SR feature in RUDP) between (a) and (b), significantly more features remain active in (a) without fading. This

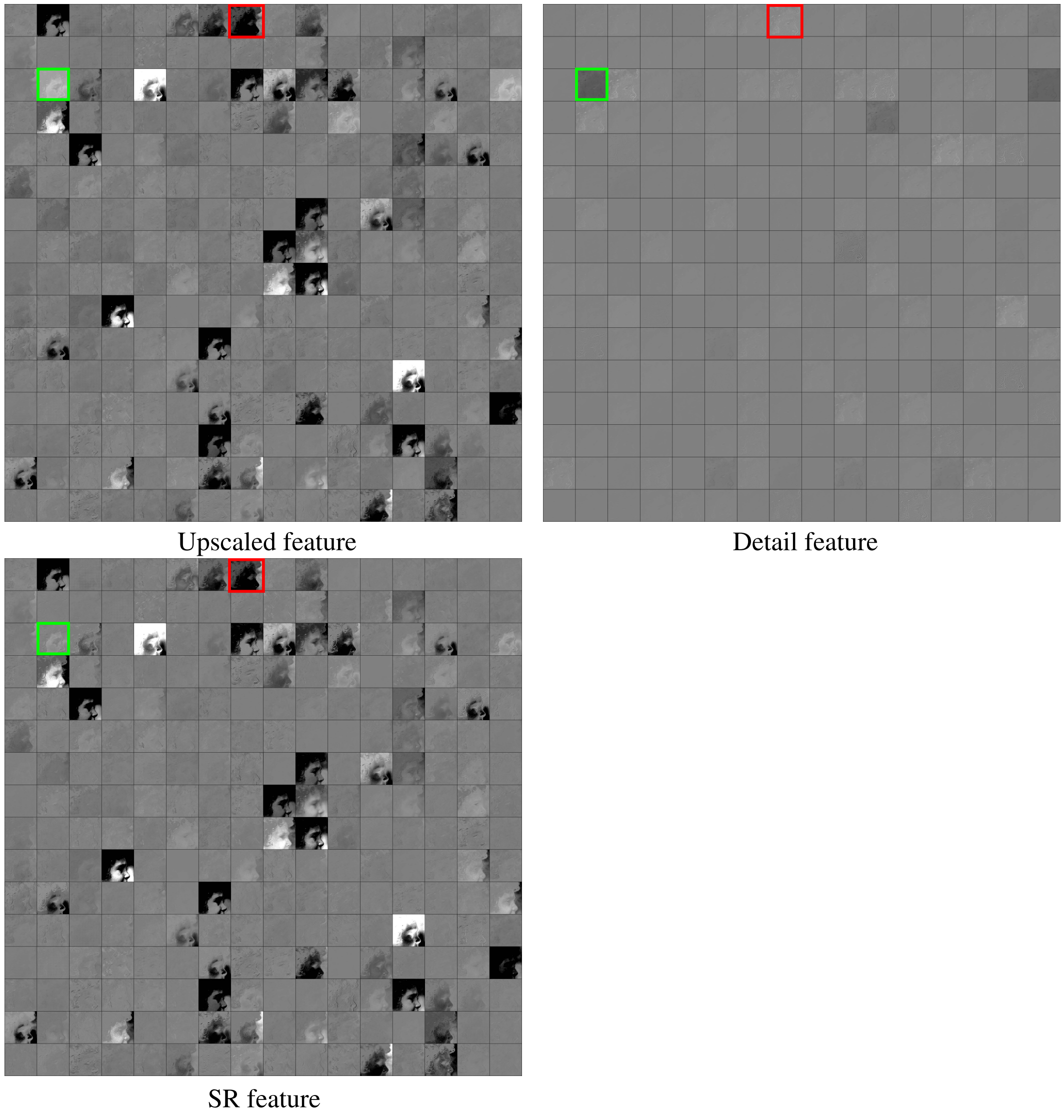


Figure 9: Three feature maps for an example image in Set5, generated during the $2\times$ super-resolution process of LaUD.

demonstrates the effect of LP-based detail loss, which helps the model extract more diverse high-frequency information.

A.6.4 Additional Visual Comparison for $\times 4$ SR on Urban100

"Image016" shows a building with a vertical line pattern. In the red-boxed area, although the wall's texture becomes visible, all models can not render this detail. A closer inspection of the window frame at the bottom reveals clear differences. DBPN produces a blurred result, showing the lowest performance. In the right part of the frame, compared to DRLN and SwinFIR, our model more effectively highlights white pixels reflecting light. Similar to the HR image, our model captures high-frequency details, maintaining bright pixels across about half of the frame.

The image, "image045," features a repeating vertical straight-line pattern. In the region highlighted by the red box, our LaUD demonstrates the second-highest performance, following DRLN. Overall, ours produces an image with fewer blurs in the middle section compared to DBPN and SwinFIR. Additionally, in the upper-left part, our image shows a more pronounced contrast, resembling light reflection.

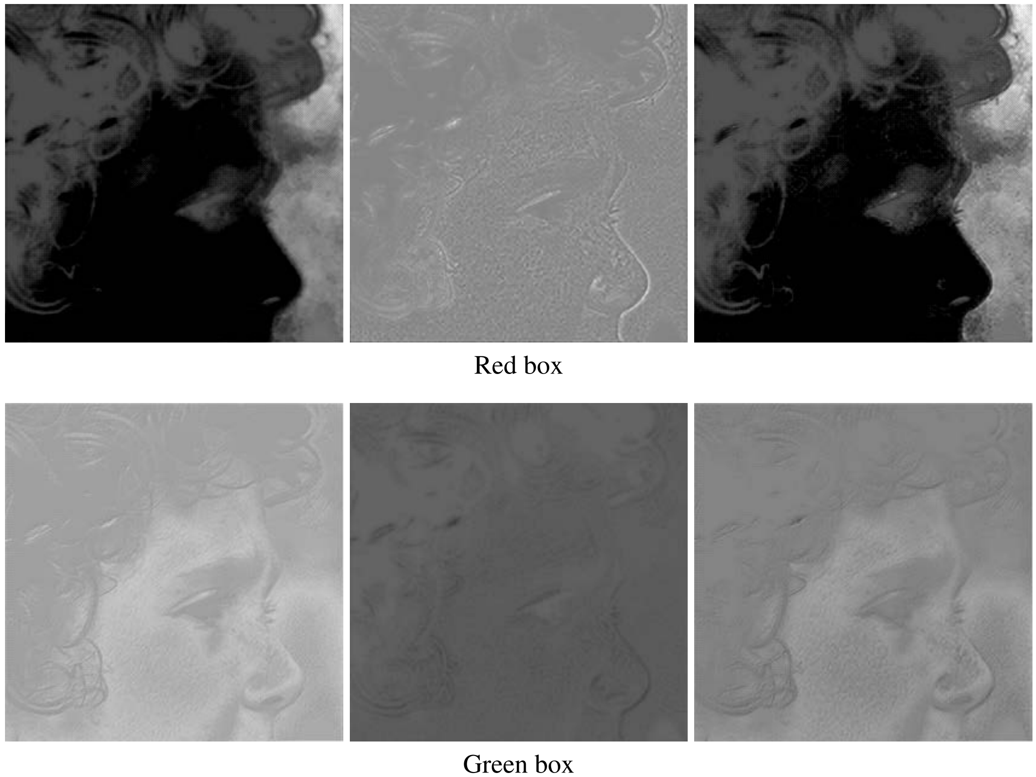
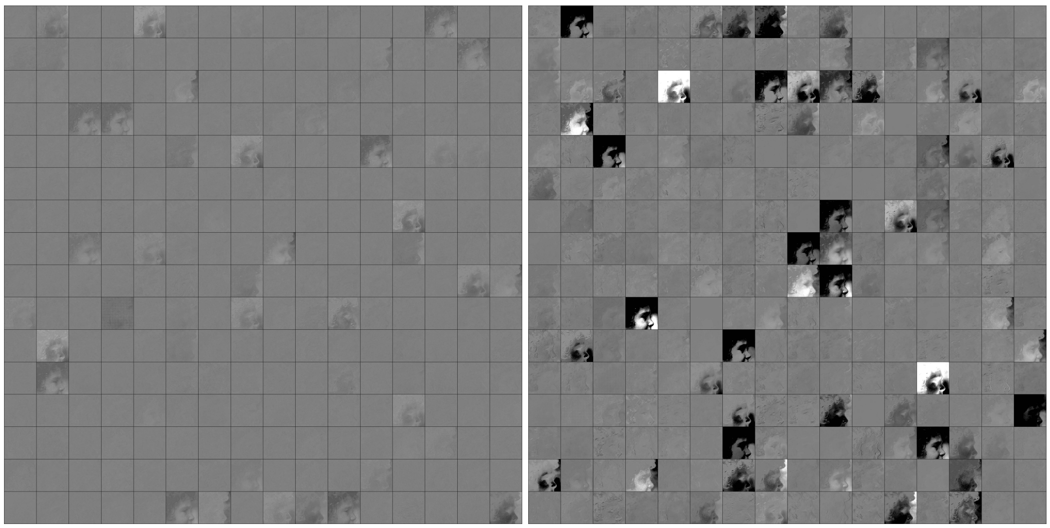
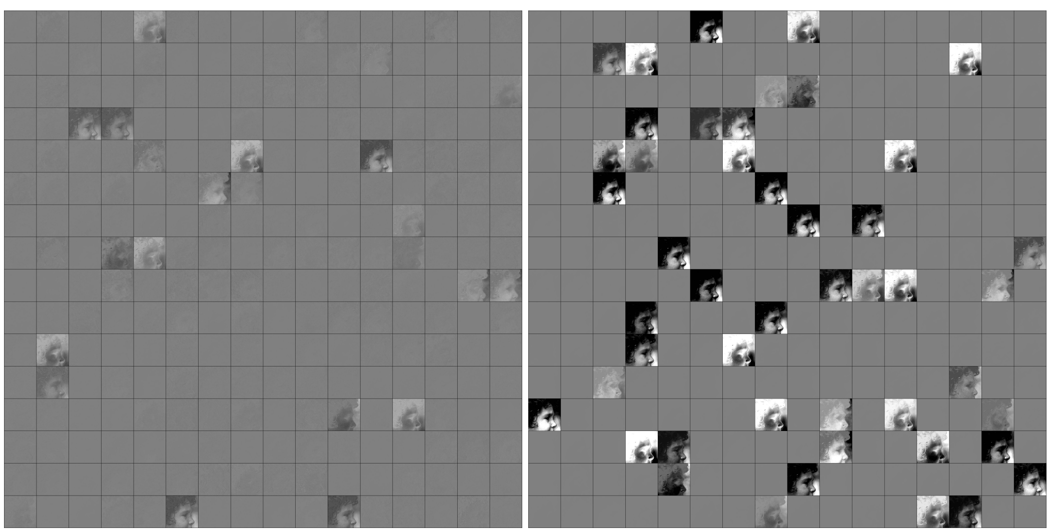


Figure 10: Enlarged views of the red (top row) and green (bottom row) boxes from Figure 9. Each row, from left to right, corresponds to the upscaled feature, detail feature, and SR feature, respectively.

The red box in the last image highlights a section with a repeating horizontal pattern of alternating bright and dark pixels. DRLN struggles to create a straight horizontal line at the bottom of the building, resulting in a low performance of 26.3223 dB. In contrast, DBPN and SwinFIR successfully generate well-formed repeating straight-line patterns, improving their performance to the 27 dB range. When comparing the repetition of the yellow and black horizontal lines in LaUD and SwinFIR, LaUD completes the pattern and enhances the contrast between the dark and bright lines, achieving the highest performance of 28.6791 dB.

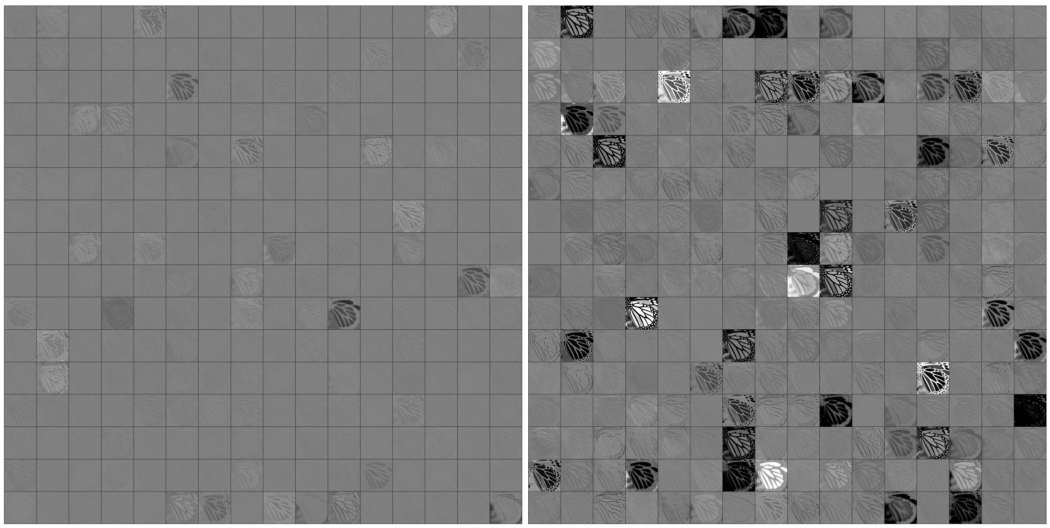


(a) For LaUD.

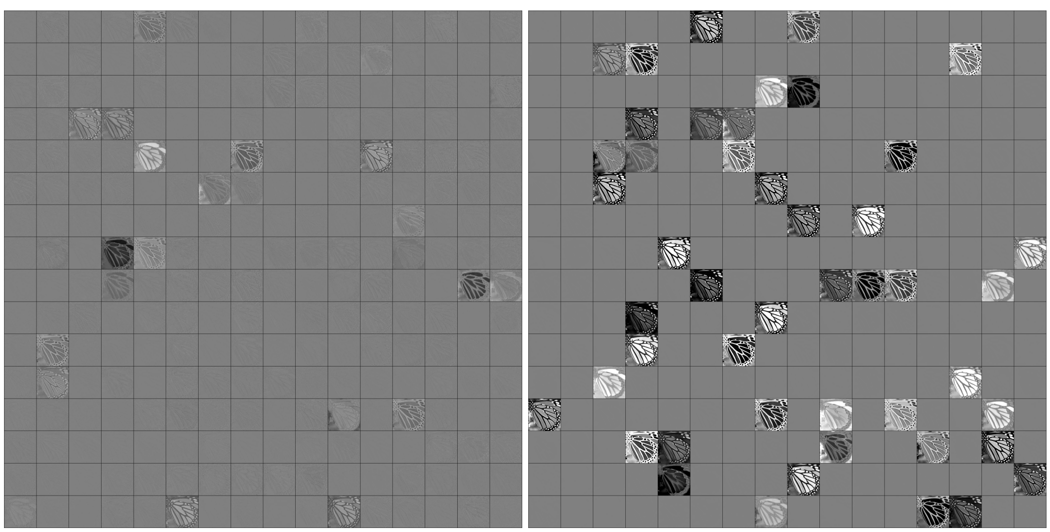


(b) For LaUD without detail loss.

Figure 11: Full SR feature map of same image in Figure 2. (a) for LaUD, and (b) for LaUD without detail loss. For each set, the left image shows the first upscaling, and the right image shows the last upscaling.



(a) For LaUD.



(b) For LaUD without detail loss.

Figure 12: Full SR feature map of another example on Set5. (a) for LaUD, and (b) for LaUD without detail loss. For each set, the left image shows the first upscaling, and the right image shows the last upscaling.

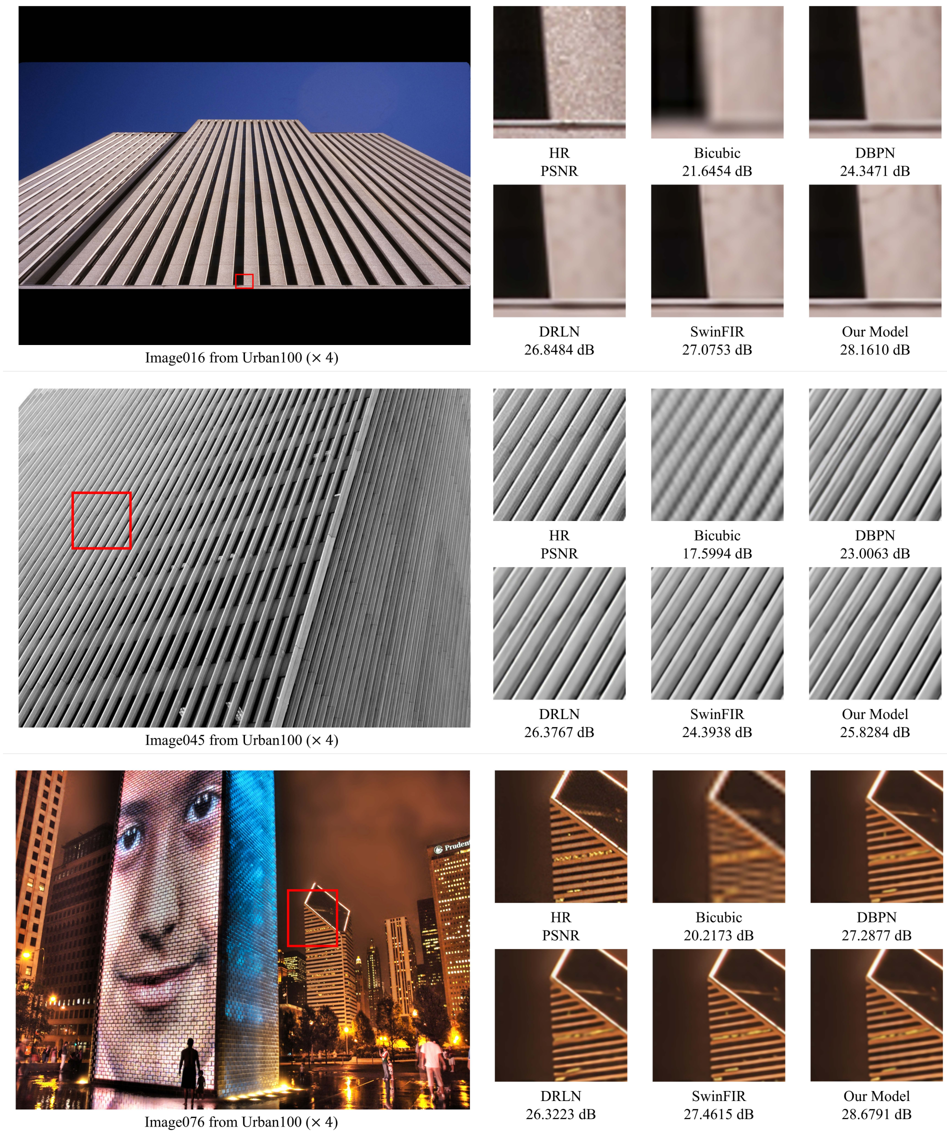


Figure 13: Additional visual comparison for $\times 4$ SR on Urban100. The patches for comparison are marked with red boxes in the original images. The PSNR values below the patches are calculated based on the patches.