





# Accelerating Image Super-Resolution Networks with Pixel-Level Classification

Jinho Jeong<sup>1</sup>, Jinwoo Kim<sup>1</sup>, Younghyun Jo<sup>2</sup>, and Seon Joo Kim<sup>1</sup>

<sup>1</sup> Yonsei University

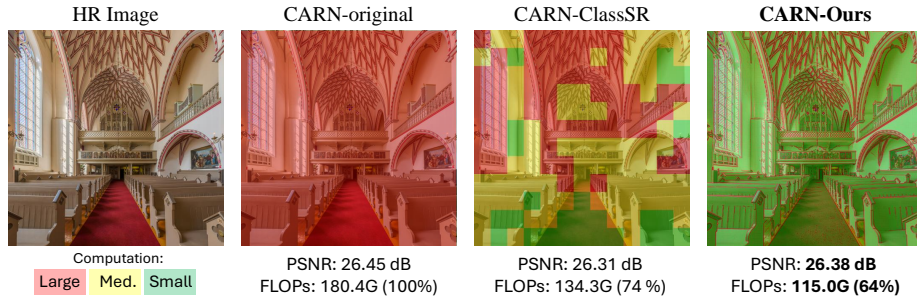
<sup>2</sup> Samsung Advanced Institute of Technology

**Abstract.** In recent times, the need for effective super-resolution (SR) techniques has surged, especially for large-scale images ranging 2K to 8K resolutions. For DNN-based SISR, decomposing images into overlapping patches is typically necessary due to computational constraints. In such patch-decomposing scheme, one can allocate computational resources differently based on each patch’s difficulty to further improve efficiency while maintaining SR performance. However, this approach has a limitation: computational resources is uniformly allocated within a patch, leading to lower efficiency when the patch contain pixels with varying levels of restoration difficulty. To address the issue, we propose the Pixel-level Classifier for Single Image Super-Resolution (PCSR), a novel method designed to distribute computational resources adaptively at the pixel level. A PCSR model comprises a backbone, a pixel-level classifier, and a set of pixel-level upsamplers with varying capacities. The pixel-level classifier assigns each pixel to an appropriate upsampler based on its restoration difficulty, thereby optimizing computational resource usage. Our method allows for performance and computational cost balance during inference without re-training. Our experiments demonstrate PCSR’s advantage over existing patch-distributing methods in PSNR-FLOP trade-offs across different backbone models and benchmarks. The code will be available at <https://github.com/3587jjh/PCSR>.

## 1 Introduction

Single Image Super-Resolution (SISR) is a task focused on restoring a high-resolution (HR) image from its low-resolution (LR) counterpart. The task has wide real-life applications across diverse fields, including but not limited to digital photography, medical imaging, surveillance, and security. In line with these significant demands, SISR has advanced in last decades, especially with Deep Neural Networks (DNNs) [6, 12, 14, 16, 23, 24].

However, as the new SISR models come out, both capacity and computational cost tend to go up, making it hard to apply the models in real-world applications or devices with limited resources. Therefore, it has led to a shift towards designing simpler, efficient lightweight models [2, 7, 8, 15, 19, 25] that consider a balance between performance and computational cost. In addition,

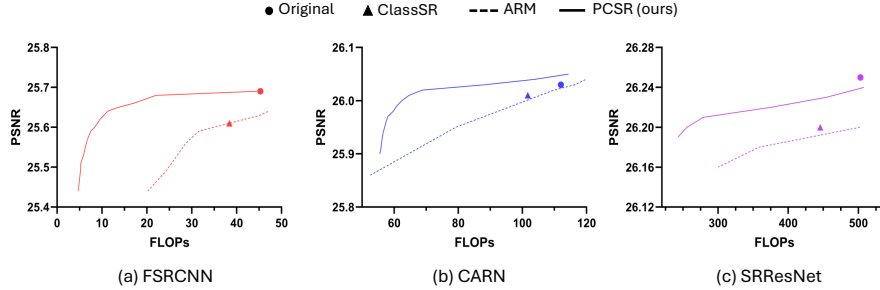


**Fig. 1:** The SR result on the image “1228” (Test2K),  $\times 4$ . By adaptively distributing computational resources in a pixel-wise manner, our method can reduce the overall computational costs in terms of FLOPs compared to the patch-distributing method, while also achieving a better PSNR score.

extensive researches [4, 10, 13, 17, 20, 21] have been developed to reduce the parameter size and/or the number of floating-point operations (FLOPs) of existing models without compromising their performance.

In parallel, there has been a growing demand for efficient SR, particularly with the rise of platforms that provide large-scale images for users such as advanced smartphones, high-definition televisions, or professional-grade monitors that support resolutions ranging from 2K to 8K. Nevertheless, SR on a large image is challenging; a large image cannot be processed in a single pass (*i.e.*, *per-image processing*) due to the limitation in computational resources. Instead, a common approach for large image SR involves dividing a given LR image into overlapping patches, applying an SR model to each patch independently, and then merging the outputs to obtain a super-resolved image. Several studies [4, 13, 20] have explored the approach, namely *per-patch processing* approach, with the aim of enhancing the efficiency of existing models while preserving their performance. These studies share the observations that each patch varies in restoration difficulty, thus allocating different computational resources to each patch.

While adaptively distributing computational resources at the patch-level achieves remarkable improvements of efficiency, it has two limitations that may prevent it from fully leveraging the potential for higher efficiency: 1) Since SR is a low-level vision task, even a single patch can contain pixels with varying degrees of restoration difficulty. That is, when allocating large computational resources to a patch that includes easy pixels, it can lead to a waste of computational effort. Conversely, if a patch with a smaller allocation of computational resources contains hard pixels, it would negatively impact performance. 2) These so-called *patch-distributing* methods become less efficient with larger patch sizes, as they are more likely to contain a balanced mix of easy and hard pixels. It introduces a dilemma: we may want to use larger patches since it not only minimizes redundant operations from overlapping but also enhances performance by leveraging more contextual information.



**Fig. 2:** Visual comparison of PSNR and FLOPs between ClassSR, ARM, and PCSR (ours) on Test2K at scale  $\times 4$ .

In this paper, our primary goal is to enhance the efficiency of existing SISR models, especially for larger images. To overcome the aforementioned limitations from patch-distributing methods, we propose a novel approach named Pixel-level Classifier for Single Image Super-Resolution (PCSR), which is specifically designed to adaptively distribute computational resources at the pixel-level. The model based on our method consists of three main parts: a backbone, a pixel-level classifier, and a set of pixel-level upsamplers with varying capacity. The model operates as follows: 1) The backbone takes an LR input and generates an LR feature map. 2) For each pixel in the HR space, the pixel-level classifier predicts the probability of assigning it to the specific upsampler using the LR feature map and the relative position of that pixel. 3) Accordingly, each pixel is assigned adaptively to a properly sized pixel-level upsampler to predict its RGB value. 4) Finally, super-resolved output is obtained by aggregating the RGB values of every pixels.

To the best of our knowledge, our method is the first to apply a pixel-wise distributing method in the context of efficient SR for large images. By cutting down redundant computations in a pixel-wise manner, we can further improve the efficiency of the patch-distributing approach, as illustrated in Fig. 1. During the inference phase, we offer users tunability to traverse the trade-off between performance and computational cost without the need for re-training. While our method enables users to manage the trade-off, we also provide an additional functionality that automatically assigns pixels based on the K-means clustering algorithm which can simplify the user experience. Lastly, we introduce a post-processing technique that effectively eliminates artifacts which can arise from the distribution of computation on a pixel-wise basis. Experiments show that our method outperforms existing patch-distributing approaches [4, 13] in terms of the PSNR-FLOP trade-off across various SISR models [7, 14, 25] on several benchmarks, including Test2K/4K/8K [13] and Urban100 [11]. We also compare our method with the per-image processing-based method [10], which process images in their entirety rather than decomposing them into patches.

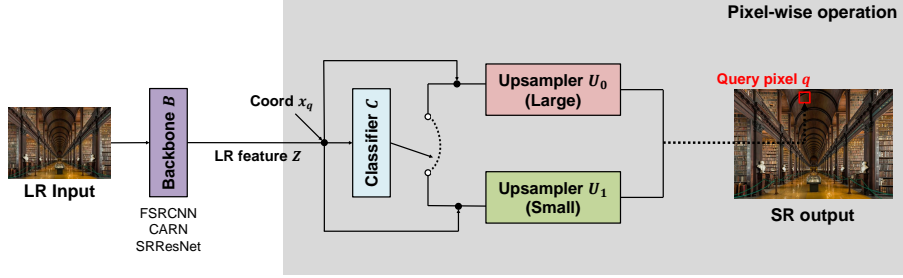
## 2 Related Works

**CNN-based SISR.** The evolution of deep learning in SISR begins with SRCNN [6], which introduces convolutional neural networks. VDSR [12] deepens this approach with residual learning. SRResNet [14] further expands the architecture using residual blocks, while EDSR [16] streamlines it, removing batch normalization for improved performance. RCAN [23] and RDN [24] advance feature extraction through channel attention and dense connections, respectively. These developments have greatly improved image quality but have also raised capacity and computational costs, posing challenges for real-world applications.

**Lightweight SISR.** The evolution of lightweight SISR models emphasizes efficiency in enhancing image quality. FSRCNN [7] starts with directly working on LR images for speed. MemNet [19] built upon this by introducing a memory mechanism for deeper detail restoration, while CARN [2] balances efficiency and accuracy using cascading designs. PAN [25] adds pixel attention for detail enhancement without heavy computational costs. LBNet [8] merges CNNs with transformers for high-quality SR on resource-constrained devices, and BSRN [15] progress with a scalable approach using separable convolutions.

**Region-aware SISR.** Region-aware SISR leverages the insight that high-frequency regions in an image are more challenging to restore than low-frequency ones. This approach aims to enhance efficiency by reducing redundant computation in low-frequency regions. AdaDSR [17] tailors its processing depth to the image’s complexity, optimizing efficiency. FAD [21] adjusts its focus based on the input’s frequency characteristics, enhancing detail in critical regions while conserving effort on smoother parts. MGA [10] initially applies a global restoration to the entire image and then refines specific regions locally, guided on a predicted mask.

Alongside, various studies have emerged focusing on efficiency in large-scale image SR. These studies decompose images into several patches and aim to enhance efficiency by dynamically allocating computational resources according to the restoration difficulty of each patch. ClassSR [13] is the first work of this area of research: it utilizes a classifier to categorize patches into simple, medium, or hard type, and assigns them to subnets with different capacities to reduce FLOPs. However, since ClassSR employs independent subnets, it leads to a significant increase in parameter count. ARM [4] resolves the limitation by decomposing the original network into subnets that share parameters, thus no additional parameters are introduced. On the other hand, APE [20] uses a regressor that predicts the incremental capacity at each layer for each patch, reducing FLOPs by early patch exiting while forwarding through network layers. In this line of study, moving away from the existing patch-distributing methods, we aim to distribute computational resources on a pixel-wise, seeking additional efficiency improvements through finer granularity.



**Fig. 3:** The architecture of the proposed PCSR model when the number of classes  $M$  is 2. We denote  $q$  as a single query pixel in the HR space and  $x_q$  for its coordinate. Pixel-level probabilities obtained from the classifier are used to allocate each query pixel to a suitably-sized upsampler for the prediction of its RGB value.

### 3 Method

#### 3.1 Preliminary

Single Image Super-Resolution (SISR) is a task aimed at generating a high-resolution (HR) image from a single low-resolution (LR) input image. Within the framework of neural networks, the SISR model aims to discover a mapping function  $F$  that converts a given LR image  $I^{LR}$  into an HR image  $I^{HR}$ . It can be represented by the equation:

$$I^{HR} = F(I^{LR}; \theta), \quad (1)$$

where  $\theta$  is a set of model parameters. Typical models [2, 7, 8, 14–16, 23–25] can be decomposed into two main components: 1) a backbone  $B$  that extracts features from  $I^{LR}$ , and 2) an upsampler  $U$  that utilizes the features to reconstruct  $I^{HR}$ . Thus, the process can further be represented as follows:

$$Z = B(I^{LR}; \theta_B), \quad I^{HR} = U(Z; \theta_U). \quad (2)$$

Here,  $\theta_B$  and  $\theta_U$  are the parameters of the backbone and the upsampler respectively, and  $Z$  is the extracted feature. In a convolutional neural network-based (*i.e.*, CNN-based) upsampler, diverse operations are employed along with convolution layers to increase the resolution of the image being processed. These range from simple interpolation to more complex methods like deconvolution or sub-pixel convolution [18]. Instead of using a CNN-based upsampler, one can employ a multilayer perceptron-based (*i.e.*, MLP-based) upsampler to operate in a pixel-wise manner, which will be further described in the following section.

#### 3.2 Network Architecture

The overview of PCSR is shown in Fig. 3. Based on our prior discussion, a model consists of a backbone and a set of upsamplers. In addition, we employ

a classifier that measures the difficulty of restoring target pixels on the HR space (*i.e.*, query pixels). LR input image is feed-forwarded to the backbone and corresponding LR feature is generated. Then, the classifier determines the restoration difficulty for each query pixel and its output RGB value is computed through the corresponding upsampler.

**Backbone.** We propose a pixel-wise computation distributing method for efficient large image SR. It is possible to use any existing deep SR networks as our backbone to fit a desired model size. For example, small-sized FSRCNN [7], medium-sized CARN [2], large-sized SRResNet [14], and also other models can be adopted.

**Classifier.** We introduce a lightweight classifier which is an MLP-based network, to obtain the probability of belonging to each upsampler (or class) in a pixel-wise manner. Given a query pixel coordinate  $x_q$ , our classifier assigns it to one of the corresponding upsamplers depending on the classification probability to predict its RGB value. By properly assigning easy pixels to a lighter upsampler instead of a heavier upsampler, we can save on computational resources with minimal performance drop.

Let an LR input be  $X \in \mathbb{R}^{h \times w \times 3}$ , and its corresponding HR be  $Y \in \mathbb{R}^{H \times W \times 3}$ . And let  $\{y_i\}_{i=1 \dots HW}$  be the coordinate of each pixel within the HR  $Y$  and  $\{Y(y_i)\}_{i=1 \dots HW}$  be the corresponding RGB values. Firstly, an LR feature  $Z \in \mathbb{R}^{h \times w \times D}$  is calculated from the LR input using the backbone. Then, given the number of classes  $M$ , classification probability  $p_i \in \mathbb{R}^M$  is obtained by the classifier  $C$ :

$$p_i = \sigma(C(Z, y_i; \theta_C)), \quad (3)$$

where  $\sigma$  is a softmax function. The MLP-based classifier operates similarly to an upsampler, with the main difference being that its output dimension is  $M$ . Please see Eq. (4) for detailed information.

**Upsampler.** We employ LIIF [5] as our upsampler, which is suitable for pixel-level processing. We first normalize  $y_i$ , which is previously defined, from the HR space to map it to the coordinate  $\hat{y}_i \in \mathbb{R}^2$  in the LR space. Given the LR feature  $Z$ , we denote  $z_i^* \in \mathbb{R}^D$  as the nearest (by Euclidean distance) feature to the  $\hat{y}_i$  and  $v_i^* \in \mathbb{R}^2$  as the corresponding coordinate of that feature. Then the upsampling process is summarized as:

$$I^{SR}(y_i) = U(Z, y_i; \theta_U) = U([z_i^*, \hat{y}_i - v_i^*]; \theta_U), \quad (4)$$

where  $I^{SR}(y_i) \in \mathbb{R}^3$  is an RGB value at the  $y_i$  and  $[\cdot]$  is a concatenation operation. We can obtain the final output  $I^{SR}$  by querying the RGB values for every  $\{y_i\}_{i=1 \dots HW}$  and combining them (Please refer to [5] for more details of LIIF processing). In our proposed method,  $M$  parallel upsamplers  $\{U_0, U_1, \dots, U_{M-1}\}$  can be exploited to handle a variety range of restoration difficulties (*i.e.* from heavy to light capacity).

### 3.3 Training

During the training phase, we feed-forward a query pixel through all  $M$  upsamplers and aggregate the outputs to effectively back-propagate the gradient as follows:

$$\hat{Y}(y_i) = \sum_{j=0}^{M-1} p_{i,j} \times U_j(Z, y_i; \theta_{U_j}), \quad (5)$$

where  $\hat{Y}(y_i) \in \mathbb{R}^3$  is an RGB output at the  $y_i$  and  $p_{i,j}$  is the probability of that query pixel being in an upsampler  $U_j$ .

Then we leverage two kinds of loss functions: reconstruction loss  $L_{recon}$ , and average loss  $L_{avg}$  which is similar one used in ClassSR [13]. The reconstruction loss is defined as the L1 loss between the RGB values of the predicted output and the target. Here, we consider the target as the difference between the ground-truth HR patch and the bilinear upsampled LR input patch. The reason is that we want the classifier to perform the classification task well, even with a very small capacity, by emphasizing high-frequency features. Therefore, the loss can be written as:

$$L_{recon} = \sum_{i=1}^{HW} |(Y(y_i) - upX(y_i)) - \hat{Y}(y_i)|, \quad (6)$$

where  $upX(y_i)$  is the RGB value of the bilinear upsampled LR input patch at the location  $y_i$ . For the average loss, we encourage a uniform assignment of pixels across each class by defining the loss as:

$$L_{avg} = \sum_{j=1}^M \left| \sum_{n=1}^N \sum_{i=1}^{HW} p_{n,i,j} - \frac{NHW}{M} \right|, \quad (7)$$

where  $p_{n,i,j}$  is probability of the  $i$ -th pixel of the  $n$ -th HR image (*i.e.* batch dimension, with batch size  $N$ ) being in the  $j$ -th class. Here, we consider the probability for being in each class as the effective number of pixel assignments to that class. We set the target as  $\frac{NHW}{M}$  because we want to allocate the same number of pixels to each class (or upsampler), out of a total of  $NHW$  pixels. Finally, total loss  $L$  is defined as:

$$L = w_{recon} \times L_{recon} + w_{avg} \times L_{avg}. \quad (8)$$

Since jointly training all modules (*i.e.*, backbone  $B$ , classifier  $C$ , upsamplers  $U_{j \in [0, M]}$ ) from scratch can lead to unstable training, we adopt multi-stage training strategy. Assuming that the capacity of the upsampler decreases from  $U_0$  to  $U_{M-1}$ , the upper bound of the model's performance is determined by the backbone  $B$  and the heaviest upsampler  $U_0$ . Thus, we initially train  $\{B, U_0\}$  only using the reconstruction loss. And then, starting from  $j = 1$  to  $j = M - 1$ , the following process is repeated: Firstly, freeze  $\{B, U_0, \dots, U_{j-1}\}$  that are trained already. Secondly, attach  $U_j$  to the backbone (and also newly attach  $C$  for  $j = 1$ ). Lastly, jointly train  $\{U_j, C\}$  using the total loss.

### 3.4 Inference

In the inference phase of PCSR, the overall process is similar to training, but a query pixel is assigned to a unique upsampler branch based on the predicted classification probabilities. While one can allocate the pixel to the branch with the highest probability, we provide users controllability for traversing the computation-performance trade-off without re-training. To this end, FLOP count is considered in the decision-making process. We define and pre-calculate the impact of each upsampler  $U_{j \in [0, M)}$  in terms of FLOPs as:

$$\text{cost}(U_j) = \sigma(\text{flops}(B; (h_0, w_0)) + \text{flops}(U_j; (h_0, w_0))), \quad (9)$$

where  $\sigma$  is the softmax function and  $\text{flops}(\cdot)$  refers to FLOPs of the module, given the fixed resolution  $(h_0, w_0)$ <sup>3</sup>. The branch allocation for pixel at  $y_i$  is then determined as follows:

$$\text{argmax}_j \frac{p_{i,j}}{[\text{cost}(U_j)]^k}, \quad (10)$$

where  $k$  is a hyperparameter and  $p_{i,j}$  is the probability of that query pixel being in  $U_j$ , as mentioned previously. By the definition, setting lower  $k$  value results in more pixels being assigned to the heavier upsamplers, minimizing performance degradation while increasing computational load. Conversely, a higher  $k$  value assigns more pixels to the lighter upsamplers, accepting a reduction in performance in exchange for lower computational demand.

**Adaptive Decision Making (ADM).** While our method allows users to manage the computation-performance trade-off, we also provide an additional functionality that automatically allocates pixels based on probability values with considering statistics across the entire image. It proceeds as follows: Given  $\forall p_{i,j}$  for a single input image and considering  $U_{j \in [0, \lfloor (M+1)/2 \rfloor)}$  as heavy upsamplers,  $\text{sum}_{0 \leq j < \lfloor (M+1)/2 \rfloor} p_{i,j}$  is computed to represent the restoration difficulty of that pixel, resulting in total number of  $i$  values. Then we group the values into  $M$  clusters using a clustering algorithm. Finally, by assigning each group to the upsamplers ranging from the heaviest  $U_0$  to the lightest  $U_{M-1}$  based on the its centroid value, all pixels are allocated to the appropriate upsampler. We especially employ the K-means clustering to minimize computational load. As we uniformly initialize the centroid values, the process is deterministic. We demonstrate the efficacy of ADM in the appendix.

**Pixel-wise Refinement.** Since the RGB value for each pixel is predicted by the independent upsampler, artifacts can arise when adjacent pixels are assigned to upsamplers with different capacities. To address this issue, we propose a simple solution: we again treat the lower half of the upsamplers by capacity as light upsamplers and the upper half as heavy upsamplers, performing refinement when

<sup>3</sup> It doesn't matter whatever the values of  $h_0$  and  $w_0$  are, as FLOPs of the module is proportional to the input resolution. We use sufficiently small values for pre-calculating the  $\text{cost}(\cdot)$  to reduce computational load.



adjacent pixels are allocated to different types of upsamplers. To be specific, for pixels assigned to  $U_j$  where  $\lfloor (M+1)/2 \rfloor \leq j < M$  (*i.e.*, light upsamplers), if at least one neighboring pixel has been assigned to  $U_j$  with  $0 \leq j < \lfloor (M+1)/2 \rfloor$  (*i.e.*, heavy upsamplers), we replace its RGB value with the average value of the neighboring pixels (including itself) in the SR output. Our pixel-wise refinement algorithm works without needing any extra forward processing, effectively reducing artifacts with only a small amount of extra computation and having minimal effect on the overall performance.

## 4 Experiments

### 4.1 Settings

**Training.** To ensure a fair comparison, we aligned the overall training settings to match those of ClassSR and ARM. We densely cropped DIV2K [1] (from index 0001-0800) into 1.59 million 32x32 LR sub-images for training dataset and random rotation and flipping are applied for data augmentation. We adopt existing FSRCNN [7], CARN [2], and SRResNet [14] as backbones with their original parameters of 25K, 295K, and 1.5M respectively. Throughout all training phases for both the original models and PCSR, the batch size is 16 and the initial learning rate is set at 0.001 for FSRCNN and 0.0002 for CARN and SRResNet with cosine annealing scheduling. Adam optimizer is used. Both the original models and the initial PCSR (which includes only the backbone and the heaviest upsampler) are trained with 2,000K iterations, while subsequent stages of PCSR’s training use 500K iterations. In the initial PCSR, we fine-tuned the hidden dimension of the backbone and adjusted the MLP size of the heaviest upsampler to maintain performance parity with the original models in terms of PSNR and FLOPs. In our implementation, we simply set  $M = 2$  as it shows the decent performance with its simplicity, which will be verified in the Sec. 4.3.

**Evaluation.** We mainly evaluate our method on the Test2K/Test4K/Test8K [13] which are downsampled from DIV8K [9], and the Urban100 [11] which consists of much larger images than the commonly used benchmarks such as Set5 [3] and Set14 [22]. For the evaluation metrics, we use PSNR (Peak Signal-to-Noise Ratio) to assess the quality of the SR images, and FLOPs (Floating Point Operations) to measure the computational efficiency. PSNR is calculated on the RGB space and FLOPs are measured on the full image. Unless specified, the original model and our PCSR is evaluated at full resolution, while ClassSR and ARM are evaluated on an overlapped patch basis. Other evaluation protocols follow those of ClassSR and ARM. When comparing PCSR with comparison groups, pixel-wise refinement is always employed and hyperparameter  $k$  is adjusted to match their performance or ADM is used.

**Table 1:** The comparison of the previous patch-level methods and our pixel-level method PCSR on the large image SR benchmarks: Test2K, Test4K, Test8K, and Urban 100 with  $\times 4$  SR. The lowest FLOPs values are highlighted in bold.

Models	Params.	Test2K(dB)	GFLOPs	Test4K(dB)	GFLOPs
FSRCNN	25K	25.69	45.3 (100%)	26.99	185.3 (100%)
FSRCNN-ClassSR	113K	25.61	38.4 (85%)	26.91	146.4 (79%)
FSRCNN-ARM	25K	25.61	35.6 (79%)	26.91	152.9 (83%)
FSRCNN-PCSR	25K	25.61	<b>8.5 (19%)</b>	26.91	<b>32.6 (18%)</b>
CARN	295K	26.03	112.0 (100%)	27.45	457.8 (100%)
CARN-ClassSR	645K	26.01	101.7 (91%)	27.42	384.1 (84%)
CARN-ARM	295K	26.01	99.8 (89%)	27.42	379.2 (83%)
CARN-PCSR	169K	26.01	<b>64.0 (57%)</b>	27.42	<b>260.0 (58%)</b>
SRResNet	1.5M	26.24	502.9 (100%)	27.71	2056.2 (100%)
SRResNet-ClassSR	3.1M	26.20	446.7 (89%)	27.66	1686.2 (82%)
SRResNet-ARM	1.5M	26.20	429.1 (85%)	27.66	1742.2 (85%)
SRResNet-PCSR	1.1M	26.20	<b>245.6 (49%)</b>	27.66	<b>981.0 (48%)</b>

Models	Params.	Test8K(dB)	GFLOPs	Urban100(dB)	GFLOPs
FSRCNN	25K	32.82	1067.8 (100%)	23.05	19.9 (100%)
FSRCNN-ClassSR	113K	32.73	709.2 (66%)	22.89	20.8 (105%)
FSRCNN-ARM	25K	32.73	746.7 (70%)	22.89	19.9 (100%)
FSRCNN-PCSR	25K	32.73	<b>196.6 (18%)</b>	22.89	<b>3.4 (17%)</b>
CARN	295K	33.29	2638.6 (100%)	24.03	49.3 (100%)
CARN-ClassSR	645K	33.25	1829.9 (69%)	24.00	51.7 (105%)
CARN-ARM	295K	33.26	1783.2 (68%)	23.99	50.8 (103%)
CARN-PCSR	169K	33.25	<b>1355.1 (51%)</b>	24.00	<b>29.6 (60%)</b>
SRResNet	1.5M	33.55	11850.7 (100%)	24.65	221.3 (100%)
SRResNet-ClassSR	3.1M	33.50	7996.0 (67%)	24.54	226.5 (102%)
SRResNet-ARM	1.5M	33.50	7865.3 (66%)	24.54	245.2 (111%)
SRResNet-PCSR	1.1M	33.52	<b>5093.7 (43%)</b>	24.54	<b>124.9 (56%)</b>

## 4.2 Main Results

As demonstrated in Tab. 1, our proposed method, PCSR, exhibits better computational efficiency compared to previous patch-based efficient SR models [4, 13] on four benchmarks, Test2K/Test4K/Test8K, and Urban100. We assess the computational costs (FLOPs) of the existing SR models [4, 10, 13] while ensuring their PSNR performance remain comparable.

We also provide qualitative results with the PSNR and FLOPs of each generated image for better comparisons in Fig. 4. Patch-level approaches such as ClassSR and ARM fail in fine-grained restoration difficulty classification. In contrast, our method can process input image more precisely due to pixel-level classification, resulting in efficient and effective SR outputs. For more detailed analysis, in Fig. 4a, ClassSR and ARM classify the shown patch area as easy one due to the dominance of the flat region, so they fail to restore thin lines well. On the other hand, our method properly classifies those lines in pixel-level difficulty classification, so it recovers them well. In Fig. 4b, due to over-computation by the patch-based methods, our approach demonstrates much better computa-

**Table 2:** The comparison of the MGA and our PCSR on Test2K, Test4K, and Urban100 with  $\times 4$  SR. The lowest FLOPs values are highlighted in bold.

Models	Params.	Test2K(dB)	GFLOPs	Test4K(dB)	GFLOPs	Urban100(dB)	GFLOPs
FSRCNN	25K	25.68	45.3 (100%)	26.98	185.3 (100%)	23.02	19.9 (100%)
FSRCNN-MGA	43K	25.66	29.2 (64%)	26.94	101.7 (55%)	23.01	14.6 (73%)
FSRCNN-PCSR	25K	25.66	<b>12.8 (28%)</b>	26.94	<b>37.8 (20%)</b>	23.01	<b>4.3 (22%)</b>
SRResNet	1.5M	26.30	502.9 (100%)	27.79	2056.2 (100%)	24.87	221.3 (100%)
SRResNet-MGA	2.0M	26.20	249.2 (50%)	27.66	871.9 (42%)	24.55	124.0 (56%)
SRResNet-PCSR	0.9M	26.20	<b>191.0 (38%)</b>	27.66	<b>755.3 (37%)</b>	24.55	<b>97.3 (44%)</b>

**Table 3:** Comparison of our PCSR and ClassSR according to the patch size, on Test2K ( $\times 4$ ). To ensure a fair comparison, the original model (CARN) and our model (CARN-PCSR) are also evaluated on decomposed input patches. The LR input size is cropped to multiples of 128 without overlap to maintain consistency across patch sizes.

Patch Size	16		32		64		128	
	PSNR(dB)	GFLOPs	PSNR(dB)	GFLOPs	PSNR(dB)	GFLOPs	PSNR(dB)	GFLOPs
CARN	26.04	98.6 (100%)	26.13	98.6 (100%)	26.18	98.6 (100%)	26.20	98.6 (100%)
CARN-ClassSR	26.03	66.7 (68%)	26.12	69.8 (71%)	26.16	72.5 (74%)	26.17	75.8 (77%)
CARN-PCSR	26.03	<b>61.1 (62%)</b>	26.12	<b>60.3 (61%)</b>	26.16	<b>56.9 (58%)</b>	26.17	<b>54.5 (55%)</b>


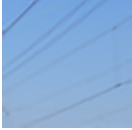
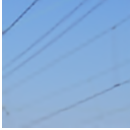
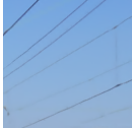
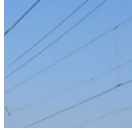



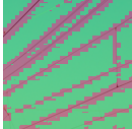
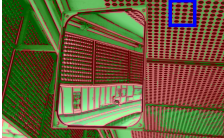
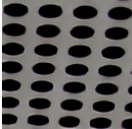
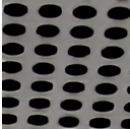
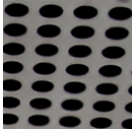
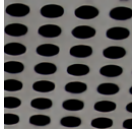
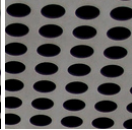


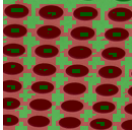
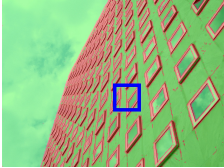
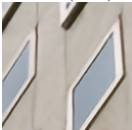
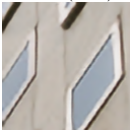
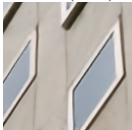
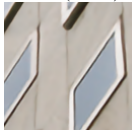
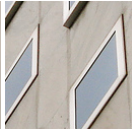
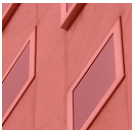


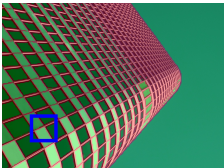
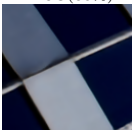
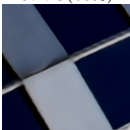
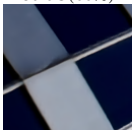
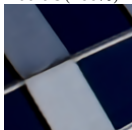

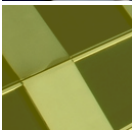
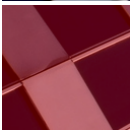
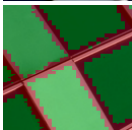
tional savings. This is attributed to our method’s efficient distribution of computational resources, allowing us to achieve comparable or better performance while minimizing computational overhead. In Fig. 4c, ClassSR waste computational resources, while ARM reduced computations excessively, resulting in inferior output quality. In contrast, our pixel-level approach enables more effective utilization of resources, leading to improved performance.

In Tab. 2, we further evaluate our method with the per-image processing efficient SR method, MGA [10]. To make a fair comparison, we use the same training dataset and input patch size as used in MGA and retrain our model. Even when compared to the per-image processing method, our model shows better efficiency with much fewer parameters, demonstrating its broad applicability and overall effectiveness.

### 4.3 Ablation Studies

**Input Patch Size.** As shown in Tab. 3, our experiments demonstrate that efficiency of the patch-distributing method [13] decreases as the size of the patch increases. This decline occurs because larger patches are more likely to contain a mix of easy and hard regions at the pixel level, making precise prediction of patch difficulty more challenging. In contrast to the patch-level approach, our method employs a pixel-level approach, allowing any patch sizes without computational efficiency decline. Our method is more efficient than the patch-level approach at all patch sizes, with the gap becoming more pronounced as the patch size increases.

**Impact of the number of classes.** In Table 4, we explore the impact of the number of classes on the efficiency of PCSR by comparing cases with  $M=2$

	Classification (Ours)	ClassSR	ARM	Ours	Backbone	GT
		26.64dB 78.2G(65%)	26.66dB 77.0G(64%)	26.85dB 72.4G(60%)	26.87dB 120.3G(100%)	
(a)						
						
		21.21dB 45.0G(97%)	21.10dB 45.8G(99%)	21.47dB 34.5G(75%)	21.31dB 46.3G(100%)	
(b)						
						
		25.07dB 37.3G(67%)	25.07dB 35.4G(64%)	25.40dB 33.3G(60%)	25.37dB 55.5G(100%)	
(c)						
						
		25.23dB 44.3G(80%)	25.18dB 37.7G(68%)	25.66dB 36.6G(66%)	25.43dB 55.5G(100%)	
(d)						
						

**Fig. 4:** Qualitative results of previous methods [4, 13] and our method with  $\times 4$  SR.

**Table 4:** Comparison depending on the number of classes  $M$  with  $\times 4$  SR.

Models	Params.	Test2K(dB)	GFLOPs	Test4K(dB)	GFLOPs	Urban100(dB)	GFLOPs
CARN	295K	26.03	112.0 (100%)	27.45	457.8 (100%)	24.03	49.3 (100%)
CARN-PCSR-2class	169K	26.01	64.0 (57%)	27.42	260.0 (58%)	24.00	29.6 (60%)
CARN-PCSR-3class	181K	26.01	62.4 (56%)	27.42	245.1 (54%)	24.00	28.6 (58%)

**Table 5:** Comparison of multi-scale PCSR and ARM on Test2K. Our model (CARN-PCSR) is retrained in a multi-scale training setting with a scale range of [2,4].

Models	Total Params.	x2			x4			x8		
		Params.	PSNR	FLOPs	Params.	PSNR	FLOPs	Params.	PSNR	FLOPs
CARN-original	<b>885K</b>	258K	30.79dB	335G	295K	26.03dB	112G	332K	23.51dB	57G
CARN-ARM	<b>885K</b>	258K	30.57dB	181G	295K	25.85dB	60G	332K	23.17dB	31G
CARN-PCSR	<b>169K</b>	169K	30.57dB	233G	169K	25.85dB	56G	169K	23.48dB	31G

and  $M=3$ . While both scenarios exhibit high efficiency compared to the original model, the case with fewer classes has minimal impact on efficiency while using fewer parameters. Therefore, for simplicity, we choose  $M=2$ .

**Multi-scale SR.** By leveraging LIIF [5] as our upsampler, our model inherently benefits from LIIF’s key feature of multi-scale SR. It allows us to maintain efficiency that only a single model is required to accommodate diverse scale factors, unlike other methods which necessitate individual models for each scale factor. We demonstrate this advantage of LIIF-based upsampling in Tab. 5. Furthermore, our model can extend to arbitrary-scale SR, including non-integer scales, a capability not achievable with conventional patch-based approaches.

**Pixel-wise Refinement.** In a patch-level approach, using individual models based on patch-wise difficulties can result in artifacts when adjacent areas are assigned to different models. This issue can be mitigated by employing patch overlapping, where overlapped areas are averaged with multiple patch-level SR outputs. However, this solution harms computational efficiency by increasing the number of patches per image. Similarly, using upsamplers based on pixel-wise difficulties can cause artifacts if neighboring pixels are assigned to different upsamplers. Our pixel-wise refinement algorithm does not require any additional forward processing, allowing artifacts to be effectively mitigated with minor additional computations and minimal impact on performance. Fig. 5 illustrates the efficacy of our simple yet effective pixel-wise refinement algorithm.

## 5 Limitation and Future Works

Our PCSR dynamically allocates resources based on the restoration difficulty of each pixel, thus pursuing further efficiency improvements through finer granularity. Nevertheless, a limitation exists: since our classifier operates based on LR features from backbone, the lower bound of PCSR’s FLOPs is determined by the



**Fig. 5:** Visualization of the artifact reduction by the pixel-wise refinement.

size of the backbone. This can lead to unnecessary computation for images with predominantly flat regions. To mitigate this, we plan to have the classifier work on the backbone’s earlier layers or use a lookup table for straightforward pixel processing through bilinear interpolation from the LR input, significantly reducing computational costs compared to neural network processing. Additionally, for future works, applying the PCSR to generative models to enhance efficiency, as well as integrating it with techniques such as model compression, pruning, and quantization, presents promising opportunities.

## 6 Conclusion

This paper introduces the Pixel-level Classifier for Single Image Super-Resolution (PCSR), a novel approach to efficient SR for large images. Unlike existing patch-distributing methods, PCSR allocates computational resources at the pixel level, addressing varying restoration difficulties and reducing redundant computations with finer granularity. It also offers tunability during inference, balancing performance and computational cost without re-training. Additionally, an automatic pixel assignment using K-means clustering and a post-processing technique to remove artifacts are also provided. Experiments show that PCSR outperforms existing methods in the PSNR-FLOP trade-off across various SISR models and benchmarks. We believe our proposed method facilitates the practicality and accessibility of large image SR for real-world applications.

## Acknowledgement

This research was supported and funded by Artificial Intelligence Graduate School Program under Grant (2020-0-01361), the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (NRF-2022R1A2C2004509), and Samsung Electronics Co., Ltd. (Mobile eXperience Business).

## References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017)
2. Ahn, N., Kang, B., Sohn, K.A.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the European conference on computer vision (ECCV). pp. 252–268 (2018)
3. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding (2012)
4. Chen, B., Lin, M., Sheng, K., Zhang, M., Chen, P., Li, K., Cao, L., Ji, R.: Arm: Any-time super-resolution method. In: European Conference on Computer Vision. pp. 254–270. Springer (2022)
5. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8628–8638 (2021)
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence* **38**(2), 295–307 (2015)
7. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14. pp. 391–407. Springer (2016)
8. Gao, G., Wang, Z., Li, J., Li, W., Yu, Y., Zeng, T.: Lightweight bimodal network for single-image super-resolution via symmetric cnn and recursive transformer. *arXiv preprint arXiv:2204.13286* (2022)
9. Gu, S., Lugmayr, A., Danelljan, M., Fritsche, M., Lamour, J., Timofte, R.: Div8k: Diverse 8k resolution image dataset. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 3512–3516. IEEE (2019)
10. Hu, X., Xu, J., Gu, S., Cheng, M.M., Liu, L.: Restore globally, refine locally: A mask-guided scheme to accelerate super-resolution networks. In: European Conference on Computer Vision. pp. 74–91. Springer (2022)
11. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5197–5206 (2015)
12. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)
13. Kong, X., Zhao, H., Qiao, Y., Dong, C.: Classsr: A general framework to accelerate super-resolution networks by data characteristic. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12016–12025 (2021)
14. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
15. Li, Z., Liu, Y., Chen, X., Cai, H., Gu, J., Qiao, Y., Dong, C.: Blueprint separable residual network for efficient image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 833–843 (2022)

16. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. pp. 136–144 (2017)
17. Liu, M., Zhang, Z., Hou, L., Zuo, W., Zhang, L.: Deep adaptive inference networks for single image super-resolution. In: *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV* 16. pp. 131–148. Springer (2020)
18. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1874–1883 (2016)
19. Tai, Y., Yang, J., Liu, X., Xu, C.: Memnet: A persistent memory network for image restoration. In: *Proceedings of the IEEE international conference on computer vision*. pp. 4539–4547 (2017)
20. Wang, S., Liu, J., Chen, K., Li, X., Lu, M., Guo, Y.: Adaptive patch exiting for scalable single image super-resolution. In: *European Conference on Computer Vision*. pp. 292–307. Springer (2022)
21. Xie, W., Song, D., Xu, C., Xu, C., Zhang, H., Wang, Y.: Learning frequency-aware dynamic network for efficient super-resolution. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 4308–4317 (2021)
22. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. *IEEE transactions on image processing* **19**(11), 2861–2873 (2010)
23. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 286–301 (2018)
24. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2472–2481 (2018)
25. Zhao, H., Kong, X., He, J., Qiao, Y., Dong, C.: Efficient image super-resolution using pixel attention. In: *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III* 16. pp. 56–72. Springer (2020)