Toward Real-World Super Resolution With Adaptive Self-Similarity Mining

Zejia Fan, Wenhan Yang, *Member, IEEE*, Zongming Guo, *Member, IEEE*, and Jiaying Liu, *Senior Member, IEEE*

***Abstract*— Despite efforts to construct super-resolution (SR) training datasets with a wide range of degradation scenarios, existing supervised methods based on these datasets still struggle to consistently offer promising results due to the diversity of real-world degradation scenarios and the inherent complexity of model learning. Our work explores a new route: integrating the sample-adaptive property learned through image intrinsic self-similarity and the universal knowledge acquired from large-scale data. We achieve this by uniting internal learning and external learning by an unrolled optimization process. With the merits of both, the tuned fully-supervised SR models can be augmented to broadly handle the real-world degradation in a plug-and-play style. Furthermore, to promote the efficiency of combining internal/external learning, we apply an attention-based weight-updating method to guide the mining of self-similarity, and various data augmentations are adopted while applying the exponential moving average strategy. We conduct extensive experiments on real-world degraded images and our approach outperforms other methods in both qualitative and quantitative comparisons. Our project is available at: https://github.com/ZahraFan/AdaSSR/.**

***Index Terms*— Super-resolution, real-world SR, semi-supervised learning, self-similarity.**

I. Introduction

SINGLE-IMAGE super-resolution is a fundamental problem in computer vision and image processing, attracting significant research interest in recent years [1], [2], [3], [4], [5], [6], [7], [8]. The task aims to recover a high-resolution (HR) image from its low-resolution (LR) observation. Many existing super-resolution (SR) methods focus on restoring the fine details of high-resolution images by learning the mapping from the low-resolution input to the high-resolution output. However, the majority of these methods [2], [9], [10], [11], [12], [13], [14] are trained using synthetic datasets or simplified scenarios, which may fail to super-resolve the complex and challenging real-world images. In real-world scenarios, images often suffer from diverse degradations, *e.g.*, noise, blur, compression artifacts, and low resolution, which can significantly impair their visual quality and limit their practical utility. Considering this, researchers have conducted studies on super-resolution with unknown degradation [3], [15], [16]. To be more specific, they aim to recover high-resolution images from low-resolution images degraded by unknown and complex degradations and further aims to achieve stable and outstanding SR performance in the real world.

As for real-world SR, the mainstream super-resolution methods can be divided into two categories. The first category is *external learning*, which mainly relies on supervised learning. Some researches [15], [16], [17] use various synthetic degradation models consisting of blur, downsampling, noise, and compression artifacts to synthesize training data. These approaches aim to enhance the model's generalization to various degradation factors, with the goal of improving the model's usability in real-world scenarios. Some works [18], [19], [20] apply generative adversarial networks [21] (GAN) to learn the implicit degradation model from the data. In external learning, a large amount of paired LR-HR data is collected as external datasets. The models are trained on diverse synthetic degradations, with fixed parameters during testing. However, due to the complexity of degradation in real scenarios, this kind of method may fail to model unknown degradations and to restore the LR real-world images adaptively.

The second category is *internal learning*, which only performs training based on the single LR image itself. These methods [22], [23], [24] mainly rely on self-supervised learning. One basis is that the degradation condition is generally shared in the entire LR image. Furthermore, real-world images show prevalent characteristics of self-similarity [25], that image patches with similar content tend to recur among different locations and scales within the same image. Some researchers take advantage of self-similarity that lies in images and mine degradation information directly from the LR input image [26], [27]. For instance, ZSSR [22] trains a small-scale convolutional neural network (CNN) during inference with the specified LR input. KernelGAN [23] utilizes GAN to estimate blur kernels and enhance the quality of reconstruction. Therefore, the internal learning approach can capture sample-dependent degradation properties. Despite its

Received 14 February 2024; revised 28 July 2024; accepted 21 September 2024. Date of publication 9 October 2024; date of current version 22 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62332010, in part by the Program of Beijing Municipal Science and Technology Commission Foundation under Grant Z241100003524010, and in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515010454. The associate editor coordinating the review of this article and approving it for publication was Dr. Nikos Deligiannis. *(Corresponding author: Jiaying Liu.)*

Zejia Fan, Zongming Guo, and Jiaying Liu are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China (e-mail: zejia@pku.edu.cn; guozongming@pku.edu.cn; liujiaying@pku.edu.cn).

Wenhan Yang is with the Pengcheng Laboratory, Shenzhen 518066, China (e-mail: yangwh@pcl.ac.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2024.3473320, provided by the authors.

Digital Object Identifier 10.1109/TIP.2024.3473320

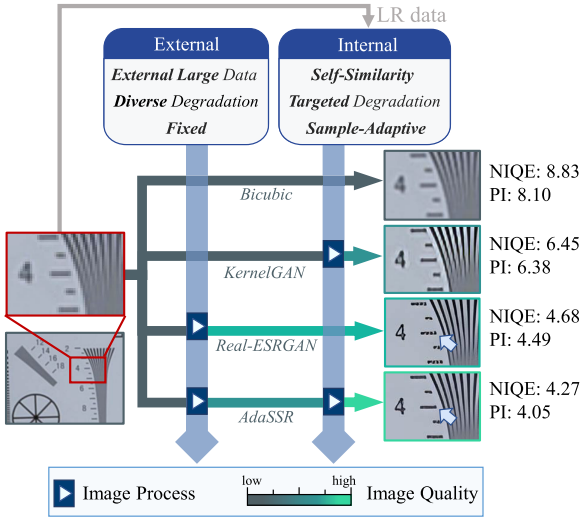


Fig. 1. The proposed adaptive self-similarity mining super-resolution (AdaSSR) is a semi-supervised method. Two kinds of real-world SR methods, Real-ESRGAN [3] (supervised external learning method) and KernelGAN [23] (self-supervised internal learning method) are compared. AdaSSR combines the strengths of both. The NIQE and PI metrics are better when smaller, and the luminance of the green color indicates the image quality.

strengths, possessing less knowledge of general signal priors from images limits its restoration capabilities, resulting in its insufficiency in producing visually satisfactory results.

The ideal real-world super-resolution method should possess both powerful and generalizable capabilities for describing degradation and visual priors, as well as the ability of sample-adaptive degradation perception and processing. Thus, the ideal paradigm of SR should combine the process of fitting to a wide range of diverse degradation and performing targeted modeling for a specific given testing image/scene. Combining internal learning and external learning, Tirer et al. [28] use LR input to fine-tune the CNN denoiser in IDBP [29]. MZSR [30] carries out rapid adaptation of model-agnostic meta-learning based on internal data. However, these endeavors have failed to accurately analyze and effectively integrate the respective advantages of internal and external learning for real-world SR scenarios. Particularly, critical issues such as how these two are combined, the extent of their contribution during integration, and the specific interaction between them have not been well addressed. Consequently, it leaves room for further performance improvement.

In this paper, we propose a simple and practical framework, named Adaptive Self-similarity mining Super-Resolution (AdaSSR), a semi-supervised method combining the advantages of external and internal learning. As shown in Fig. 1, AdaSSR aims to integrate the sample-dependent degradation characteristics learned adaptively from LR self-similarity and the general sense knowledge extracted from a large dataset. We achieve this by constructing a deep network inspired by the unrolling process from the alternating direction method of multiplier (ADMM) solver. To make the knowledge of internal/external learning better integrated, we also introduce a self-similarity-informed attention-based method to guide the internal learning process. We generate the attention map with the self-attention mechanism, which can emphasize regions with high self-similarity in the image. Our framework can be applied to work jointly with various external learning methods and offers better performance in a plug-and-play manner. Through quantitative and qualitative comparisons, we demonstrate that our approach outperforms state-of-the-art methods on real-world LR images.

In summary, our contribution is threefold:

- We propose a novel Adaptive Self-similarity mining Super-Resolution (AdaSSR) framework, which is inspired by the unrolled process of an optimization function that unifies the external and internal learning for SR. It adaptively learns degradation characteristics from image self-similarity and generalizable knowledge from external datasets.
- We introduce the self-similarity-informed attention maps to guide the internal learning process, resulting in a plug-and-play solution to boost the performance of external learning models, and can handle the sample-dependent degradations presented in real-world environments.
- We analyze the strengths and limitations of internal learning and external learning in SR with unknown degradation scenarios and study how to combine the advantages of both. The analysis indicates that internal learning relies more on self-similarity capable of describing sample-dependent degradation, while external learning is more dependent on the feature extraction capabilities learned from a diverse range of data.

The paper is organized as follows. We first review the related work in Section II. Then, in Section III, we conduct an empirical analysis of internal learning and external learning to show their characteristics, and build the proposed semi-supervised AdaSSR inspired from the unrolling process from an optimization function that integrates internal and external learning. Experimental results are presented in Section IV. Conclusions are summarized in Section V.

II. Related Work

A. Image Super-Resolution Network Architecture

Image super-resolution has drawn great research interest in recent years [2], [5], [10], [31], [32], [33]. As a pioneer work of deep-learning based image SR method, SRCNN [31], which builds a three-layer CNN to reconstruct HR images from bicubic-degraded LR images, proves the ability of CNNs to extract local correlation of images. Because such ability is critical to image SR tasks, CNNs have been used widely since that. Ledig et al. [2] propose to build a deep CNN with residual connections to improve the reconstruction performance. Zhang et al. [33] take channel-attention mechanism to maintain better local correlation of features. Liang et al. [5] apply vision transformers [34], [35] in SR networks. Researchers have also discovered the powerful role of the self-attention mechanism in SR tasks. Liang et al. [5] use a shift window local self-attention mechanism for deep feature extraction and propose a strong Transformer model for image restoration. Chen et al. [4] combine both channel attention and window-based self-attention mechanisms for activating more pixels for better restoration. Lei and Shi [36] propose a hybrid-scale self-attention-based network for remote sensing image SR,

considering self-similarity is strong in the input. Liu et al. [37] incorporate contrastive learning-based degradation representation extraction with the self-attention mechanism for image SR with unknown degradation.

Along with the development of generative models, they are introduced into SR tasks to improve perceptual quality. Ledig et al. [2] build a generative adversarial network-based [21] SR model. Menon et al. [38] use a pre-trained GAN as a generative prior and explore the latent space to find appropriate high-resolution images that can be degraded to the input LR images as SR results. Some researchers [39], [40] embed pre-trained GANs into SR networks as knowledge prior to assisting the SR task on LR images in certain domains. Dahl et al. [41] leverage auto-regressive models to generate SR images in a pixel-wise way. Lugmayr et al. [42] use normalizing flow models [43] to get multiple reasonable SR results corresponding to one LR image input. Saharia et al. [14] and Ma et al. [44] regard LR images as conditions of diffusion models [45] and achieve the goal of SR by LR-conditional generating. The mentioned works are representative works of different network structure designs in the SR field, and various network designs can be adapted into our method AdaSSR in a plug-and-play manner.

B. Real-World Image Super-Resolution

The gap between real LR images and synthesized LR images during training could cause issues for practical real-world SR. To the end of mitigating the effect, an intuitive way to train the SR model is to build degraded images close to real-world LR images and force the model to learn such an unknown degradation. Pioneer work, SRMD [17], builds LR images by blurring and adding noise and provides the features of blur kernel and noise as model input. Cai et al. [46] collect a real LR image dataset which can be used to evaluate the performance of real-world SR models. Zhang et al. [47] design a random degradation process to generate LR images imitating real-world LR images. Wang et al. [3] improve the process to make it closer to real-degradation and use an existing GAN-based SR model [11], achieving impressive results. In practice, the synthesized LR-HR pair can be used to train any end-to-end SR models and enable them to process real-world LR images. Huang et al. [15] propose to iteratively apply a kernel-estimator and a reconstructor to refine the results of each other. Ates et al. [48] also apply an iterative method which contains kernel reconstruction, noise estimation and SR reconstruction in each iteration. Wang et al. [20] utilize contrastive learning methods to extract abstract degradation representations of LR images. Li et al. [49] extend the framework of DASR [20] to images with more general degradation.

Some works have achieved real-world SR capabilities by learning from real-world HR and LR data. Song et al. [50] simulate the unknown downsampling process through an adversarial training framework, eliminating the need for restrictive prior knowledge or paired examples. Zhang et al. [51] propose an auxiliary-LR generator and AdaSTN to address the misalignment between short-focus low-resolution images and telephoto ground-truth images, which achieves self-supervised learning for the real-world dual camera scene. Chen et al. [52] combine supervised pre-training with self-supervised learning to improve the adaptability of image SR models on real-world images while training an LR reconstructor with paired real-world data. Compared to these works, our method does not require any real degradation data for the test images.

Due to the self-similarity of images, several one-shot methods that attempt to mine degradation information in the input LR image itself have been proposed. Shocher et al. [22] train a small-scale CNN of the specified LR input during inference. Bell-Kligler et al. [23] leverage a GAN to estimate blur kernels to improve the reconstruction quality. Soh et al. [30] transfer a pre-trained SR model with few parameter updates. Yang et al. [53] propose an unsupervised kernel estimation model with a Markov chain Monte Carlo sampling process on random kernel distributions. DGDML-SR [24] samples HR and LR patches according to the depth map and unites the degradation network and the SR network together. The proposed AdaSSR integrates the strengths of both external and internal learning, leveraging the robust feature extraction capability acquired from external data and the sample-adaptive image reconstruction ability derived from self-similarity.

III. Adaptive Self-Similarity Mining Super-Resolution (AdaSSR)

A. Empirical Analysis: Internal Learning vs. External Learning

Internal learning SR leverages the information inherent in the LR image itself to enhance the performance of SR. On the contrary, *External learning SR* utilizes valuable knowledge from external datasets to reconstruct LR images.

We first analyze the characteristics of internal/external learning for SR empirically, and the results are shown in Fig. 2. Below is the experimental setup. We first conduct internal learning and subsequently proceed with external learning on the inherited network parameter. Circular points represent the results obtained after the convergence of internal learning, while each data point on the line represents the results obtained through external learning after training with a different number of external images. We employ a lightweight CNN-based SR network, and each data point represents convergence under different settings. For the internal learning phase, we mainly followed the settings in ZSSR [22]. We generate a set of LR-HR patch pairs from the test image y under 4 different settings: **Oracle**, **Desired**, **Estimated** and **Mismatched**. After the internal learning, we proceed to the second phase, external learning. The horizontal axis, $\#External$, represents the quantity of external data used. The test image and external images undergo $4\times$ down-sampling using random blur kernels, with each image a unique blur kernel, while the kernel generation is inspired by classic works in real-world SR [3]. As for the LR-HR pairs used in each internal learning setting, **Oracle** utilizes ground truth HR x and LR y generated with kernel k^g for training, representing the theoretical upper bound achieved by the network. **Desired** involves down-sampling y using the kernel k^g, resulting in y_d^g, paired with y. **Estimated** estimates the down-sampling blur kernel k^e from y and down-samples y to obtain y_d^e, paired with LR. **Mismatched** utilizes bicubic

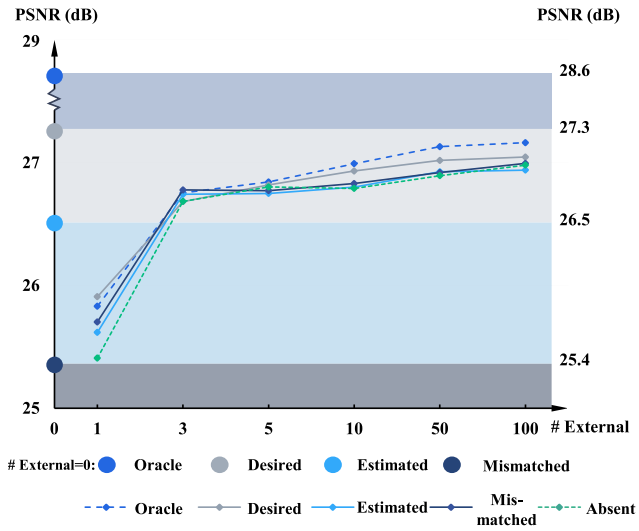


Fig. 2. An empirical analysis of the characteristics of using internal/external learning for SR with unknown degradations. A brief description of the experimental setup is provided below. **Oracle**: ground truth x paired with LR y that generated with kernel k^g. **Desired**: y_d^g obtained with kernel k^g, paired with y. **Estimated**: y_d^e obtained with the estimated blur kernel k^e, paired with y. **Mismatched**: y_d^b obtained by bicubic downsampling and paired with LR. **Absent**: no internal learning, randomly initialize the network. $\#External$ refers to the number of external data used, with the leftmost end representing complete internal learning, and the increase in values showcasing the impact of the second stage, external learning, on different scales of external data.

downsampling generated y_d^b paired with y. **Absent** indicates the absence of the internal learning stage, with the network parameters randomly initialized. The external learning process trains on high-quality external data and corresponding LR generated with random blur kernels.

From experiment results shown in Fig. 2, we can qualitatively observe certain characteristics of internal learning and external learning:

- The restoration ability of external learning primarily stems from the generalization obtained by fitting a large amount of data. With an increase in external training data, although the images are affected by various mismatched types of degradation, the feature extraction capabilities of the network get stronger, leading to enhanced generalization and the capacity to handle diverse degradation scenarios.
- Internal learning trains in a sample-dependent manner to adapt to the targeted degradation environment. It lacks training data and relies on the self-similarity of the test image, making the estimation of the current degradation scene a key factor in its performance improvement. When the estimation of degradation is not accurate enough, its performance may lag behind that of external learning supported by large datasets. However, if the degradation estimation approaches the **Desired** level, it can achieve superior results with minimal computational cost compared to external learning.
- The combination of the advantages of internal learning and external learning is susceptible to catastrophic forgetting. In internal learning, a clear decline in performance is observed in the order of **Oracle**, **Desired**, **Estimated**, and **Mismatched** settings. However, after external learning, such as $\#External = 100$, the performances of these settings become similar, resembling the **Absent** setting. Preserving the knowledge acquired during the internal learning phase is challenging for neural networks.

Inspired by the above observations, we propose AdaSSR, a semi-supervised method that combines the merits of fully-supervised and self-supervised approaches. Our approach takes advantage of the generalization performance of external learning models and the sample-adaptive capacity of internal learning. By comprehensively leveraging the strengths of both sides, we can give a better SR result for real-world LR images.

B. Formulation and Optimization Function

This section formulates the optimization function, from which we derive an alternative solution and inspires the construction of our AdaSSR. Let y_r and x_r respectively denote the low-quality image and the ideal high-quality image in real-world testing scenarios. Let (y_i, x_i) denote the paired training data and x_i is from a large-scale collected external dataset Φ, where $i = 1, 2, \ldots, M$ and M is the total pair number. y_i is downsampled from x_i with a synthetic degradation. The network is expected to learn to combine the advantages of internal learning and external learning in a semi-supervised learning manner as follows:

$$\min_{\Theta} \mathcal{L}_p \left(N\left(y_r|\Theta\right), x_r\right) + \lambda \sum_{x_i \in \Phi} \mathcal{L}_p \left(N\left(y_i|\Theta\right), x_i\right), \quad (1)$$

where Θ represents the parameters of an SR neural network, $\mathcal{L}_p\left(N(y|\Theta), x\right)$ measures the difference between the SR result from y under parameters Θ and high-resolution image x, $N(\cdot)$ represents the neural network operation, and λ is the weighting parameter. It can be considered as the generalized reconstruction loss during network training in a broad sense.

Ideally, the results can be obtained via solving the optimization problem (1), and it is expected to obtain the SR model by unrolling the optimization function. However, in real-world image SR inference, each image undergoes variations attributed to degradation factors. There should exist a strong correlation between Θ and characteristics of y_r, while Φ can remain stable for all inference time comparatively. To perform SR on LR images in real-world scenarios, we can employ the same large-scale synthetic dataset Φ. Consequently, taking into account the generality and ease of method design, it is preferable to optimize these two losses separately, allowing us to focus on the characteristics of y_r and, to extent, avoid redundant training on Φ. Based on the above considerations, we derive a plug-and-play method that is compatible with most existing fully-supervised SR models. An unfolding inference for solving problem (1) is developed as the clues for the network construction that are inspired by ADMM [54]. The ADMM solves the optimization problem as follows:

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmin}}\, f(\Theta) + \lambda g(\Theta). \quad (2)$$

The optimization problem (1) can be converted into formula as Eq. (2) with $f(\Theta) = \mathcal{L}_p \left(N\left(y_r|\Theta\right), x_r\right)$ and $g(\Theta) = \sum_{x_i \in \Phi} \mathcal{L}_p \left(N\left(y_i|\Theta\right), x_i\right)$. The unconstrained optimization problem Eq. (2) can be converted into a constrained problem by

introducing an auxiliary variable Θ_e:

$$(\widehat{\Theta}, \widehat{\Theta}_e) = \underset{\Theta,\Theta_e}{\operatorname{argmin}} f(\Theta) + \lambda g(\Theta_e), \text{ s.t. } \Theta = \Theta_e. \quad (3)$$

Considering the corresponding augmented Lagrangian function:

$$\mathcal{L}(\Theta, \Theta_e, \boldsymbol{u}) = f(\Theta) + \lambda g(\Theta_e) + \boldsymbol{u}^T(\Theta - \Theta_e) + \mu\|\Theta - \Theta_e\|^2, \quad (4)$$

where μ is the penalty parameter. The minimizer of Eq. (3) can be found by solving a sequence of subproblems:

$$\Theta_e^{(k+1)} = \underset{\Theta_e}{\operatorname{argmin}} \lambda g(\Theta_e) + \mu \left\|\Theta_e - \widetilde{\Theta}_e^{(k)}\right\|^2, \quad (5)$$

$$\Theta^{(k+1)} = \underset{\Theta}{\operatorname{argmin}} f(\Theta) + \mu \left\|\Theta - \widetilde{\Theta}^{(k)}\right\|^2, \quad (6)$$

$$\overline{\boldsymbol{u}}^{(k+1)} = \overline{\boldsymbol{u}}^{(k)} + \left(\Theta_e^{(k+1)} - \Theta^{(k+1)}\right), \quad (7)$$

in which $\overline{\boldsymbol{u}}^{(k)} \overset{\text{def}}{=} (2/\mu)\boldsymbol{u}^{(k)}$ is the scaled Lagrange multiplier, $\widetilde{\Theta}_e^{(k)} \overset{\text{def}}{=} \Theta^{(k)} - \overline{\boldsymbol{u}}^{(k)}$ and $\widetilde{\Theta}^{(k)} \overset{\text{def}}{=} \Theta_e^{(k+1)} + \overline{\boldsymbol{u}}^{(k)}$. When both functions $f(x)$ and $g(x)$ exhibit properties of being closed, proper, and convex, and assuming the existence of a saddle point for the Lagrangian function Eq. (4), it can be demonstrated that the iterations given by Eq. (5)–(7) converge towards a solution for Eq. (3). This assumption holds true for a linear SR network with common MSE training loss. For more complex networks, such as Real-ESRGAN, we experimentally demonstrate the convergence in Section IV-C. Specifically, when the number of iterations is set to 1, we can pre-train on a large-scale external dataset. For any given real image, sharing the pre-trained model and only conducting internal learning, we can obtain a trade-off in the solution for Eq. (2) in a resource-efficient way.

1) External Learning: Process on external dataset Φ. Substituting $g(\Theta) = \sum_{x_i \in \Phi} \mathcal{L}_p(N(y_i|\Theta), x_i)$ into the iteration formula (5), we obtain:

$$\Theta_e^{(k+1)} = \underset{\Theta_e}{\operatorname{argmin}} \lambda \sum_{x_i \in \Phi} \mathcal{L}_p(N(y_i|\Theta_e), x_i) + \mu \left\|\Theta_e - \widetilde{\Theta}_e^{(k)}\right\|^2. \quad (8)$$

This process focuses on fitting to the external dataset. Especially for the first iteration, the initialization of Θ can be used to support the SR in the general scene, and the regularization term $\|\Theta_e - \widetilde{\Theta}_e^{(k)}\|^2$ can be ignored. In this way, the iteration is entirely consistent with supervised learning methods:

$$\Theta_e^{(1)} = \arg\min_{\Theta_e} \sum_{x_i \in \Phi} \mathcal{L}_p(N(y_i|\Theta_e), x_i). \quad (9)$$

Our framework initializes the model with a supervised method pre-trained parameters $\Theta_e^{(1)}$ following the above form.

2) Internal Learning: Substituting $f(\Theta) = \mathcal{L}_p(N(y_r|\Theta), x_r)$ into the iteration formula (6), we obtain the following optimization problem:

$$\Theta^{(k+1)} = \underset{\Theta}{\operatorname{argmin}} \mathcal{L}_p(N(y_r|\Theta), x_r) + \mu \left\|\Theta - \widetilde{\Theta}^{(k)}\right\|^2. \quad (10)$$

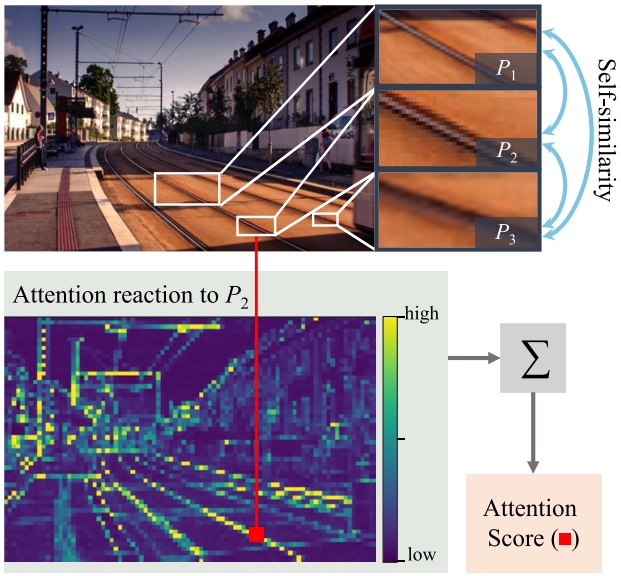


Fig. 3. Visualization of self-similarity and attention score calculation. Three patches at different scale levels are outlined with white boxes, which share similar content. The blue arrows indicate the inherent similarity among them. The lower part depicts the attention reactions, which are from the attention model, of all regions to Patch P_2, marked with the red square. Similar patches contribute significantly to the attention reactions. The sum of attention reactions from the image yields the attention score for the patch represented by the red square.

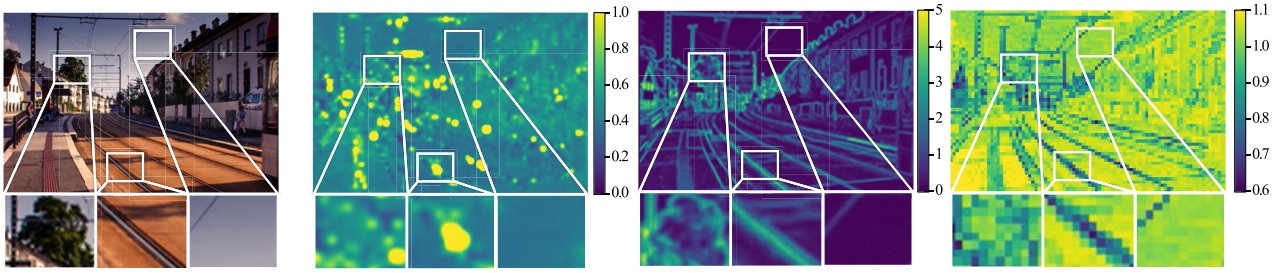
For the term $\mathcal{L}_p(N(y_r|\Theta), x_r)$, we cannot obtain the x_r for training supervision. Following most internal learning SR methods [22], [23] built on *self-similarity*, we also obtain training pairs by further down-sampling the input y_r with the estimated kernel to derive $(y_{r,d}, y_r)$ pair, where d is the downsampling scale. The kernel is estimated in a self-learning manner. As shown in Fig. 3, the self-similarity means that similar patches recur across scales within natural images. For instance, patches P_1, P_2 and P_3 are similar to each other. The restoration progress of P_1 in scale of $(y_{r,d}, y_r)$ can guide that of P_2 in scale of (y_r, x_r). Namely, the (y_d, y) pair has the potential to provide rich useful guidance to fit x taking as the input y. Thus, we use $\mathcal{L}_p\left(N\left(y_{r,d}|\Theta\right), y_r\right)$ to approximate $\mathcal{L}_p(N(y_r|\Theta), x_r)$.

Especially for the first iteration, the initialization of $\overline{\boldsymbol{u}}^{(0)}$ can support image SR in the general sense, and the regularization term $\|\Theta - \widetilde{\Theta}^{(k)}\|^2$ can be ignored.

C. Internal Learning Optimization Strategy

To better solve the optimization problem Eq. (10) for internal learning, we adopt the following three optimizations in the implementation. We 1) use attention map to guide the exploration of self-similarity, 2) adopt the exponential moving average (EMA) updating mechanism to constrain the process of model adaptation, reducing the learning burden of the loss term $\|\Theta - \widetilde{\Theta}^{(k)}\|^2$, 3) employ data augmentation techniques to provide more diverse and comprehensive training supervision.

1) Self-Similarity Based Attention Map Guidance: We make the network acquire the capability of reconstructing x_r through the training supervision of the pair $(y_{r,d}, y_r)$, relying primarily on the cross-scale self-similarity of natural images. To efficiently and precisely explore the self-similarity of images, we



(a) Input image (b) Attention map (c) Gradient map (d) Similarity map

Fig. 4. Display of attention map guidance. (a) is the corresponding input image. (b) is the attention map for (a). The gradient map and similarity map are also shown for comparison in (c) and (d), respectively. The color of the border of the image patches below is indicated by the color bar on the right, representing the magnitude of map weights. It can be observed that the parts with larger weights usually have clearer textures and greater self-similarity in the attention map, while the areas with smaller weights have less information.

propose an attention map to guide the training process as the weight of loss.

Our proposed attention map is based on the transformer-based image enhancement network, specifically using SwinIR [5]. SwinIR consists of three main parts: shallow feature extraction, deep feature extraction, and HQ image reconstruction. The deep feature extraction is formed by stacking Residual Swin Transformer Blocks. We leverage the last multi-head self-attention in the last block, where deeper layers focus more on high-level information, aiding in better exploration of self-similarity.

Let us first review the multi-head self-attention in the Swin transformer [35] layer. Taking an input of size $H \times W \times C$, the network initially transforms it into an $HW/M^2 \times M^2 \times C$ feature by dividing the input into non-overlapping $M \times M$ local windows, where HW/M^2 is the total number of windows. Subsequently, it calculates standard self-attention independently for each window, referred to as local attention. For a local window feature $X \in \mathbb{R}^{M^2 \times C}$, the query, key, and value matrices Q, K, and V are computed:

$$Q = XP_Q, \quad K = XP_K, \quad V = XP_V. \tag{11}$$

As for our attention map calculation, after obtaining an input of size $H \times W \times C$, we perform pooling operations by a factor of n and transform the input to obtain G, a feature of size $HW/n^2 \times C$. We then compute global attention as follows:

$$Q = GP_Q, \quad K = GP_K, \tag{12}$$

$$A = \text{SoftMax}\left(QK^T/\sqrt{C}\right), \tag{13}$$

$$\text{AttentionMap}(G) = \sum_i A_{i,j}, \tag{14}$$

Considering the multi-head self-attention mechanism, the attention map is averaged among all heads. The attention map is used during training to weight the loss on a pixel-wise basis, thereby guiding the training of internal learning:

$$\Theta_{t+1}^{(k)} \leftarrow \Theta_t^{(k)} - \eta \nabla \left(w \circ f(\Theta_t^{(k)}) + \mu \|\Theta_t^{(k)} - \tilde{\Theta}^{(k)}\|^2 \right) \tag{15}$$

where t is the current updating step, η is the learning rate, w is the attention map and $\circ$ refers to pixel-wise multiplication. The formula is also shown in line 13 of Algorithm 1.

In Fig. 4, we show the attention map and compare it with other choices for guidance. The gradient map in Fig. 4(c) is commonly used in the field of image enhancement [55], [56], capable of extracting fine edges and textures. However, the features extracted by this map are low-level, lacking the understanding of the semantic information and self-similarity, which results in assigning high weights to noise or less frequently occurring parts.

The similarity map shown in Fig. 4(d) is a naive approach focused on self-similarity. It obtains features for each patch through the network feature extraction module, calculates cosine similarity pairwise to get the similarity matrix, and sums along one dimension of the matrix. The resulting value for every single patch can approximately measure to which degree the entire image is similar to that patch. However, this similarity score can be high for parts with less information, such as patches of sky or flat ground that are similar to other solid-colored regions. These patches with less information should not be focused on for model enhancement ability.

The attention map differs from the other two. The computation mechanism of self-attention ensures that each patch gives a higher attention score to beneficial patches for its reconstruction, *i.e.*, focusing more on patches with useful information. The presentation in Fig. 4(b) demonstrates that the extraction of the attention map aligns with our design methodology. Patches with higher self-similarity, such as the railroad tracks, receive higher weights, less frequent tree crowns obtain moderate weights, and parts with less information like blank sky have low weights.

2) EMA Updating: Overfitting on generated paired data $(y_{r,d}, y_r)$ can potentially affect the feature extraction and image restoration capabilities of external learning, so a balance between internal learning and external learning is critical. In Eq. (10), the loss $\|\Theta - \tilde{\Theta}^{(k)}\|^2$ is the penalty term to satisfy $\Theta = \Theta_e$ in Eq. (3). So we initialize the parameter Θ with $\Theta_e^{(k+1)}$. We also adopt the EMA strategy that is widely used for semi-supervised learning, domain adaption, and unsupervised learning as a convenient and efficient regularizer:

$$\Theta_{A_t} = \alpha \Theta_{A_{t-1}} + (1-\alpha)\Theta_t, \tag{16}$$

where Θ_t represents the parameters of a continuously updated model at step t guided by $\mathcal{L}_p\left(N\left(y_{r,d}|\Theta\right), y_r\right)$. Θ_{A_t} is the EMA model parameter, which satisfies $\Theta_{A_0} = \Theta_e^{(k+1)}$. α is the decay rate in the range of [0, 1], which is responsible for determining to which degree the network parameter Θ stays

Algorithm 1 Training Process of AdaSSR

Input: LR image y_r, external HR data $\mathcal{X}$, external learning pretrained parameters $\Theta_e^{(1)}$, model operation N, attention-based newtork $Attn$, learning rate η, total iterations T_{iter}, max steps T_{step}, EMA decay parameter α, hyper parameters λ and μ.

Output: SR result x_r^*.

$w \leftarrow Attn(y_r)$ // Calculate attention map

for $k = 1$ **to** T_{iter} **do**

  Update $\tilde{\Theta}_e^{(k)}, \tilde{\Theta}^{(k)}, \bar{\boldsymbol{u}}^{(k)}$ with ADMM formula

  /* External learning process */

  if $k > 1$ **then**

    $\Theta_{e,1}^{(k)} \leftarrow \Theta^{(k-1)}$

    for $t = 1$ **to** T_{step} **do**

      $\Theta_{e,t+1}^{(k)} \leftarrow \Theta_{e,t}^{(k)} - \eta\nabla\left(\lambda g(\Theta_{e,t}^{(k)}) + \mu\|\Theta_{e,t}^{(k)} - \tilde{\Theta}_e^{(k)}\|^2\right)$ // Calculate on $\mathcal{X}$

      $\Theta_e^{(k)} \leftarrow \alpha\Theta^{(k-1)} + (1-\alpha)\Theta_{e,t+1}^{(k)}$

    end

  end

  /* Internal learning process */

  $\Theta_1^{(k)} \leftarrow \Theta_e^{(k)}$

  for $t = 1$ **to** T_{step} **do**

    $\Theta_{t+1}^{(k)} \leftarrow \Theta_t^{(k)} - \eta\nabla\left(w \circ f(\Theta_t^{(k)}) + \mu\|\Theta_t^{(k)} - \tilde{\Theta}^{(k)}\|^2\right)$ // Calculate on y_r

    $\Theta^{(k)} \leftarrow \alpha\Theta_e^{(k)} + (1-\alpha)\Theta_{t+1}^{(k)}$

    $x_t^* \sim N(y_t \mid \Theta^{(k)})$

    if *IQA*(x_t^*) *trends downward* **then**

      Break

    end

  end

end

close to $\Theta_e^{(k+1)}$. This training strategy ensures a more stable network training process and alleviates the learning burden on the penalty term $\|\Theta - \widetilde{\Theta}^{(k)}\|^2$. The training is stopped based on the change in non-reference subjective evaluation metrics of the super-resolved $\widehat{x}$ image, and Θ_{A_T} is taken as $\Theta^{(k+1)}$, where T is the last step.

3) Data Augmentation: Random sampling directly from $(y_{r,d}, y_r)$ pair results in few samples and insufficient diversity. Due to the small volume of internal data, data augmentation is of great importance. We utilize affine transformations, noise injection, RandAugment [58], and unsharp masking (USM) operations by chance. Affine transformations, particularly scaling, help AdaSSR to deeply mine the self-similarity features on images with different scales, while noise injection helps the network become more robust to noise and learn high-frequency textures more accurately. The random USM sharpening on LR stresses the edge information, assisting in the learning of the main object in the image. A more detailed description is presented in the Section IV.

Algorithm 1 shows the pseudo code for AdaSSR framework.

IV. Experiments

A. Experimental Setup

1) Internal Training: For the internal learning iteration, we only use the real-world LR images to be tested as the data source. During each internal learning iteration, the network parameters are first initialized by the results of the external iteration. Then, the pair data is generated from the LR input y_r through downsampling and data augmentation. Each step is updated according to Eq. (15) and parameters are updated in an EMA manner. We set T_{step} to 1,500 while employing NIQE of SR results for early stopping. We perform testing every 100 steps. If the decrease in the metric is less than 0.05 for three consecutive times, we terminate the training.

For the attention weighting map, we perform a weighted average between the attention map (weighted by 0.9) and a map of all ones (weighted by 0.1), then clipped to the region within [0.1, 3.0]. The attention map only needs to be computed once for each LR input.

For random augmentation, we use rotation, flip, random affine transformations, noise injection, RandAugment, and unsharp masking operations by chance. The scale transformation in affine transformation can have a significant impact on the experimental results. Excessively large scale values lead to over-blur while too small values hinder the utilization of self-similarity and result in a severe difference from the test condition. Therefore, we uniformly sample scale values from [0.5, 1.0]. We also add Gaussian and Poisson noise with a probability of {0.24, 0.16}.

2) External Training: For the first external learning iteration, we use the pre-trained model parameters from each external learning method as the optimization result of Eq. (9). In subsequent external learning iterations, we use paired data generated with the HR images from DIV2K [59] dataset and synthetic downsampling process as in Real-ESRGAN [3]. T_{step} is also 1,500 with EMA mechanism applied.

3) Implementation Details: We conduct experiments on a single RTX 2080 Ti GPU using PyTorch. The batch size is set to 4. As for the Eq. (3), we set $\lambda = 1\times10^{-3}$ and $\mu = 1\times10^{-4}$. Unless otherwise noted, the experimental results presented are obtained by applying AdaSSR on Real-ESRGAN [3] and the ADMM process takes iteration T_{iter} as 1. We apply Adam optimizer with a learning rate of 1×10^{-4}. The decay parameter α of EMA is set to 0.999.

4) Testing Datasets: RealSR dataset [46] is a challenging dataset for the task of real-world SR. The dataset contains real-world paired LR-HR images of the same scene captured using two full-frame DSLR cameras (Canon 5D3 and Nikon D810) with four focal lengths. We take the 100 paired test data for experiments. Due to the use of automatic settings for focus, white balance, and exposure, as well as other technical limitations in acquiring the dataset, the HR and LR images are not perfectly aligned and the HR images can only be a reference. We represent it as Ref GT in Fig. 5.

DRealSR [60] is also designed for complex real-world SR degradations and collects real LR and HR image pairs from five different DSLR cameras (Canon, Sony, Nikon, Olympus, and Panasonic) in indoor and outdoor scenes.

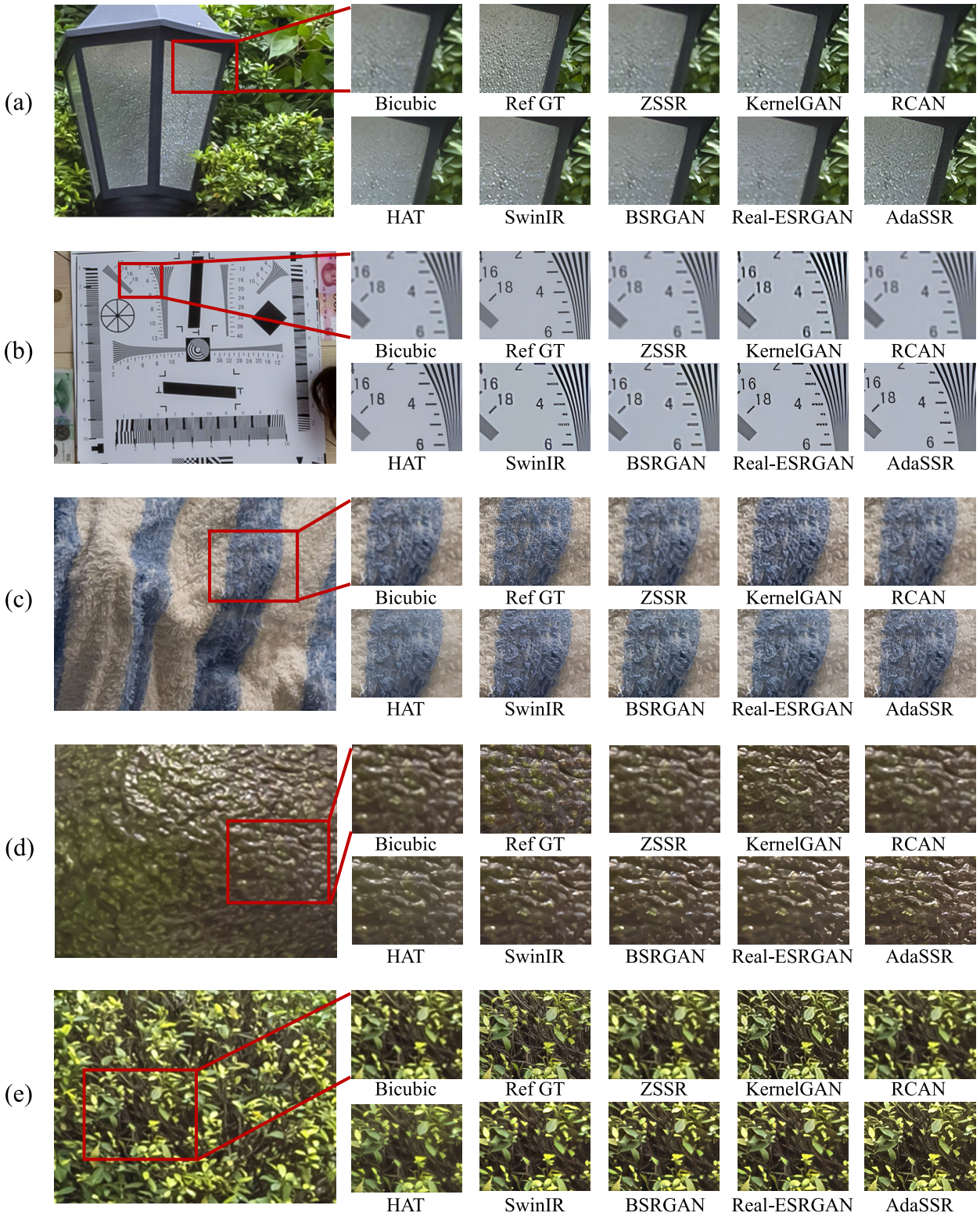


Fig. 5. Visual comparison on real-world LR input from RealSR. Ref GT refers to the reference ground truth obtained by shooting with a different focal length. AdaSSR result is based on Real-ESRGAN. **[Zoom in for best view]**.

5) Testing Methods: We compare our method with popular single-image super-resolution (SISR) methods. The self-supervised internal learning methods include ZSSR [22] and KernelGAN [23]. The external learning methods include follows: SRCNN [31], RCAN [33], LDL [57], DASR [61], HAT [4], SwinIR [5], BSRGAN [47], Real-ESRGAN [3] and DiffIR [6]. SRCNN is a foundational SR convolutional network consisting of three convolutional layers. RCAN is an advanced CNN-based SISR method, while HAT and SwinIR represent state-of-the-art Transformer-based methods

and DiffIR represents the diffusion-based model. In addition, LDL and DASR are supervised methods designed for real-world SR. MZSR [30] is a semi-supervised method for real-world SR. Taking into account that some methods have released multiple models for selection, we made the following choices: We test on the real SR GAN-based HAT model, Swin-BSRGAN-Large model and DiffIR-RealSR model. Specifically, Swin-BSRGAN-Middle model is applied for DrealSR inference since the image size is large.

As demonstrated in Section III-A, the scale of external data used has a nonnegligible influence on external learning performance. We show the details for the training process of comparison external learning models in the supplementary material Section II-A.

6) Evaluation Metrics: We test 5 metrics: NIQE [62], BRISQUE [63], NRQM [64], PI [65], LPIPS [66], which are commonly used for measuring image perceptual quality to evaluate the reconstruction performance.

B. Experimental Results

1) Visualization Comparison: In Fig. 5, we show the visualization comparison on real-world LR images that contain various content. Additional results are presented in the supplementary materials.

External learning methods might fail in regions with out-of-scope degradation, and it is difficult to control and diagnose such errors in the reconstructed results. The visualization comparison of external learning methods on real-world scenes in Fig. 5 shows the following characteristics:

- External learning method heavily relies on the data distribution of the dataset used for training. In (a), supervised methods fail to recover clear results due to the unique texture of the lamp glass cover, whose patterns take a relatively low proportion in the external dataset.
- The feature reconstruction methods based on external learning may not be effectively adapted to well handle random real-world scenarios, such as in (c), where the HAT method excels in reconstructing soft and delicate high-definition textures but fails to match the features of the towel input LR.
- Attention mechanisms like HAT can use non-local patch-based LR to reconstruct SR, but the reconstructed result in (c) has obvious color deviation and reduced contrast, possibly due to the improper utilization of LR information by the attention mechanism.

Internal learning methods often learn to over-blurred results. Although they learn some specific degradation prior from the LR, they have a poor ability to extract and restore high-quality textures and structures, and to utilize semantic information within the LR to recover corresponding high-frequency details. In most scenarios, the ZSSR restoration results are similar to Bicubic upsampling, showing a trend toward over-blurriness. Although KernelGAN estimates the blur kernel of the LR image, it can introduce additional noise and artifacts when the blur kernel is not accurately estimated, as demonstrated in Fig. 5(b) and (c) with overshoot and ringing artifacts.

Our proposed AdaSSR method exhibits excellent performance in visualization comparison. In Fig. 5(a), it successfully reconstructs the glass texture and makes the leaf contours and textures more distinct. In Fig. 5(b), SwinIR and Real-ESRGAN both incorrectly reconstruct the scale lines, while KernelGAN results in severe ringing artifacts, HAT exhibits bias in the recovery of number "6", and the overall color of HAT's result becomes lighter. The reconstruction results of other methods are not clear enough, while AdaSSR performs better. In Fig. 5(c), AdaSSR fully explores self-similarity and effectively reconstructs towel textures. In Fig. 6, we present additional examples of AdaSSR applied to real-world scenarios, showcasing the performance of our method that our approach demonstrates advantages in revealing details and avoiding artifacts. More illustrations will be provided in the supplementary material.

2) Quantitative Comparison: In Table I, we show the quantitative comparison on RealSR dataset among SISR methods, including SOTA external and internal learning methods. The AdaSSR framework can be combined with any external learning network structures and here we show the results on SRCNN, SwinIR, BSRGAN, Real-ESRGAN, and DiffIR. It can be observed that the internal learning methods show significant improvement over the Bicubic method, but still fall short compared to the external learning methods. SRCNN and RCAN are trained on simple Bicubic synthesized datasets and therefore perform worse than other external learning methods that use synthetic datasets based on more complicated degradation models and have stronger generalization capacity. In the case of $\times$ 2 downsampling, the combination of SRCNN and AdaSSR shows a significant improvement. This is attributed to the substantial gap between bicubic downsampling and real-world degraded images. Internal learning effectively bridges this gap. On the other hand, models like SwinIR, BSRGAN, Real-ESRGAN, DiffIR, *etc.*, utilize complex synthetic processes to simulate the degradation of $\times$ 2 downsampling images well. AdaSSR provides a relatively smaller improvement in this context. However, in the case of $\times$ 4 downsampling, where image degradation is severe, a noticeable gap exists between synthetic data and degraded data. AdaSSR brings a significant improvement for external learning methods under these circumstances. For a more comprehensive comparison, we present the experimental results on DrealSR in Table II. AdaSSR demonstrates outstanding performance in more diverse shooting scenarios.

3) Analysis on Self-Similarity: Self-similarity is an intrinsic property of natural images [25], [67]. Internal learning highly relies on the self-similarity of images, which not only helps the modeling of the HR-LR relationship but also assists the estimation of the current degradation environment.

Below we discuss the impact of self-similarity on AdaSSR. In Fig. 7, we display a set of images with varying degrees of self-similarity: (a) exhibiting low self-similarity, (b) featuring a certain level of self-similarity in the person's hair, and (c) showcasing strong self-similarity in the bushes.

We evaluate the effect of self-similarity on our method. The results are shown in Table III. AdaSSR is based on Real-ESRGAN as a baseline. The *Difference* demonstrates the improvements of our method compared to Real-ESRGAN. It can be observed that our approach brings significant

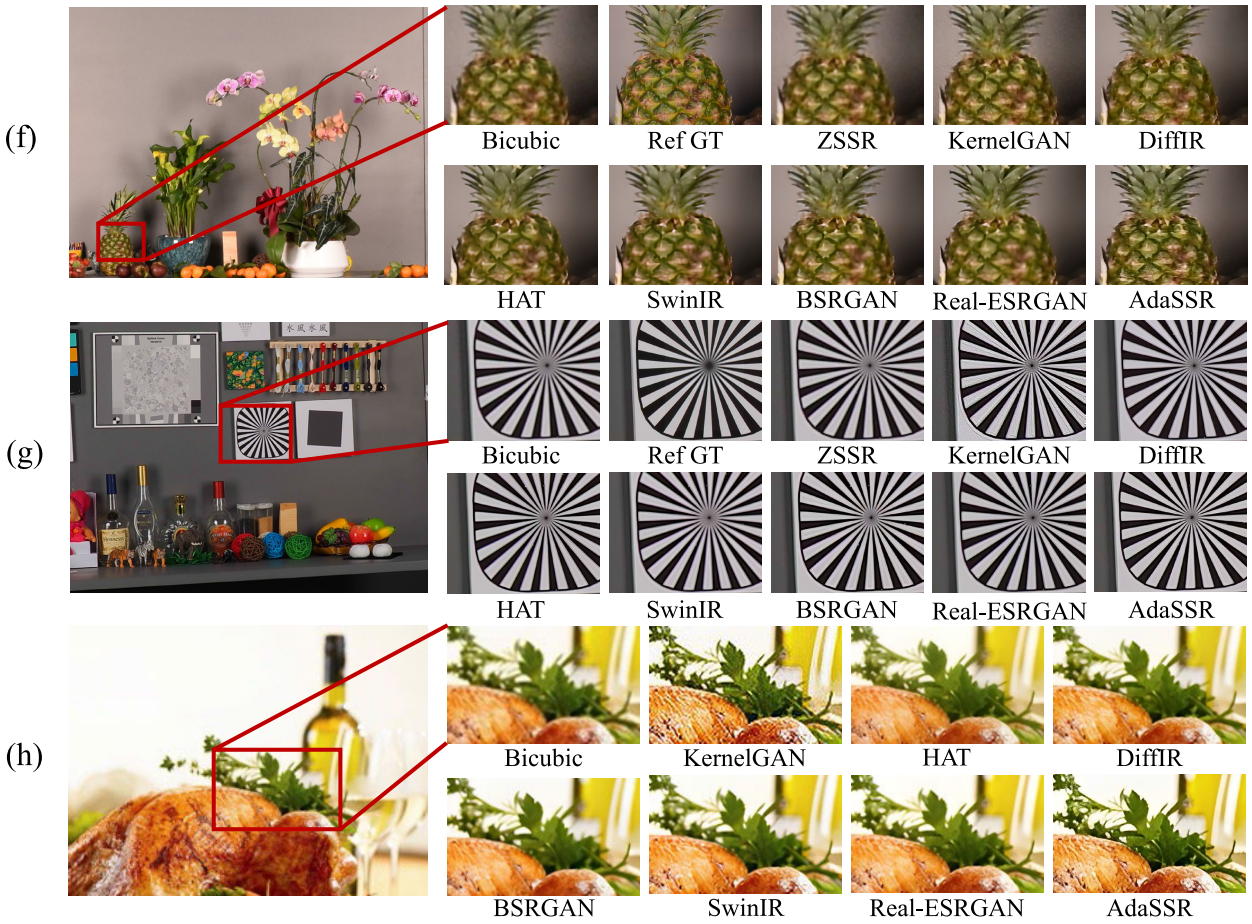


Fig. 6. Visualization comparison on real-world LR images. Cases (f) and (g) are from DRealSR while (h) is from the internet. AdaSSR result is based on Real-ESRGAN. **[Zoom in for best view]**.

Fig. 7. Image samples with varying degrees of self-similarity.

improvements to images with strong self-similarity, and also exhibits certain performance gains on images with little self-similarity. Additionally, during the internal learning process, the use of non-reference metrics, *e.g.* NIQE, to seek optimal results ensures that the performance is bounded by the external learning method.

4) Analysis on Benefits from External Learning: In order to demonstrate how internal learning benefits from external learning, we list three kinds of experiments that show the necessity and benefits of combining the two instead of only applying internal learning. As shown in Fig. 8, we show the visualization comparison results of the internal learning method, AdaSSR w/o external learning, and AdaSSR. The comparison reveals that using only the internal learning method results in severe artifacts, poor image clarity, and significant ringing and overshoot. Internal learning methods ZSSR, KernelGAN, and MZSR all apply to the light model. As for the AdaSSR w/o external learning case, the amount of internal learning data is insufficient to support the training of a normal size network (e.g., RRDBNet [11] in this case), resulting in abnormal output performance. In contrast, the results of AdaSSR are more consistent with human subjective perception.

5) Experiments on Different External Learning Methods: In Fig. 9, we compare the visualization results before and after applying AdaSSR to various external learning methods. External learning methods exhibited domain gap factors in real-world SR scenarios vary. For instance, the Real-ESRGAN network tends to simplify output textures, resulting in overly flat SR results, as shown in Fig. 9(a); SWINIR produces relatively blurry results in real-world SR; and DiffIR tends to generate color casts, as shown in Fig. 9(c). AdaSSR, by utilizing test image-specific self-learning, makes targeted corrections to these external learning methods, resulting in SR outcomes that align more closely with human subjective perception.

In the AdaSSR framework, the improvement in different external learning methods is mainly influenced by two factors: the capacity of the network itself and the generalization capacity of the model. For the first point, if the network has low capacity, it poorly fits the data distribution in the external dataset. In real-world image cases, it benefits more from

TABLE I

QUANTITATIVE COMPARISON AMONG SISR METHODS. WE SHOW THE PERFORMANCE OF ADASSR AND COMPARISON METHODS ON REALSR DATASET OF SCALE ×2, ×4. THE SYMBOL ↑ INDICATES THAT HIGHER METRIC VALUES ARE BETTER, WHILE ↓ INDICATES THE OPPOSITE. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

			NIQE ↓	BRISQUE ↓	PI ↓	LPIPS ↓	NRQM ↑
Scale ×2	*Interpolation method*	Bicubic	7.06	55.37	6.80	0.223	3.55
	Internal learning method	ZSSR [22]	6.82	52.78	6.53	0.189	3.89
		KernelGAN [23]	6.60	43.91	5.47	0.173	5.66
	External learning method	SRCNN [31]	7.03	53.35	6.63	0.182	3.92
		RCAN [33]	6.82	53.45	6.59	0.200	3.74
		LDL [57]	4.79	32.44	4.51	0.157	5.89
		HAT [4]	6.82	53.49	6.58	0.201	3.74
		SwinIR [5]	4.43	29.31	4.18	0.149	**6.20**
		BSRGAN [47]	5.55	31.31	4.89	0.146	5.84
		Real-ESRGAN [3]	4.63	31.37	4.44	0.149	5.87
		DiffIR [6]	5.04	**28.21**	4.50	0.144	6.15
	Semi-supervised method	MZSR [30]	6.91	50.71	5.36	0.229	6.12
		AdaSSR+SRCNN	6.26	47.33	5.82	0.145	4.81
		AdaSSR+SwinIR	**4.39**	29.35	**4.16**	0.147	**6.20**
		AdaSSR+BSRGAN	5.49	31.80	4.87	0.143	5.85
		AdaSSR+Real-ESRGAN	4.59	31.07	4.40	0.147	5.90
		AdaSSR+DiffIR	4.93	28.71	4.47	**0.140**	6.16
			NIQE ↓	BRISQUE ↓	PI ↓	LPIPS ↓	NRQM ↑
Scale ×4	*Interpolation method*	Bicubic	8.83	67.15	8.10	0.477	2.72
	Internal learning method	ZSSR [22]	7.91	62.65	7.54	0.406	3.03
		KernelGAN [23]	6.45	52.53	6.38	0.317	3.80
	External learning method	SRCNN [31]	8.15	61.56	7.71	0.385	2.91
		RCAN [33]	8.55	66.68	7.84	0.442	3.01
		LDL [57]	4.89	30.45	5.68	0.278	4.71
		DASR [20]	5.97	44.07	4.06	0.311	6.06
		HAT [4]	5.26	32.23	5.08	0.247	5.38
		SwinIR [5]	4.68	33.79	4.71	0.252	5.50
		BSRGAN [47]	4.65	25.35	4.46	0.269	5.98
		Real-ESRGAN [3]	4.68	29.13	4.49	0.273	5.87
		DiffIR [6]	5.39	31.11	4.96	0.253	5.66
	Semi-supervised method	MZSR [30]	6.60	53.42	6.28	0.304	4.13
		AdaSSR+SRCNN	7.24	57.56	6.99	0.343	3.35
		AdaSSR+SwinIR	4.56	32.84	4.56	**0.244**	5.67
		AdaSSR+BSRGAN	4.54	25.15	4.34	0.263	6.09
		AdaSSR+Real-ESRGAN	**4.24**	**24.47**	**3.99**	0.266	**6.34**
		AdaSSR+DiffIR	4.93	29.57	4.51	0.245	5.91

TABLE II

QUANTITATIVE COMPARISON AMONG SISR METHODS ON DREALSR DATASET OF SCALE ×4. BELOW IS THE SUPER-RESOLUTION PERFORMANCE OF ADASSR AND COMPARISON METHODS. THE SYMBOL ↑ INDICATES THAT HIGHER METRIC VALUES ARE BETTER, WHILE ↓ INDICATES THE OPPOSITE. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

	NIQE ↓	BRISQUE ↓	PI ↓	LPIPS ↓	NRQM ↑
Bicubic	9.66	67.61	8.60	0.438	2.56
SwinIR	4.57	29.43	4.74	0.284	5.39
BSRGAN	4.68	27.86	4.78	0.293	5.46
Real-ESRGAN	4.72	29.87	4.78	**0.282**	5.43
AdaSSR+SwinIR	4.53	29.33	4.73	**0.282**	5.42
AdaSSR+BSRGAN	4.64	**27.72**	4.75	0.291	5.46
AdaSSR+Real-ESRGAN	**4.44**	29.24	**4.53**	**0.282**	**5.60**

the supplementary information provided by internal learning, resulting in significant improvement. We present the improvement in metric LPIPS for different external learning methods in Table VI. SRCNN network has a small capacity and learns poor reconstruction ability from the external dataset, resulting in a large improvement under the AdaSSR framework. In contrast, networks like Real-ESRGAN, SWINIR, DiffIR, and BSRGAN have a large capacity and inherently strong SR capabilities, so their improvement under the AdaSSR framework is at the same level.

For the second point, the greater the impact of the domain gap on the method's performance, the higher the potential for improvement. A network with poor generalization capacity is more affected by the domain gap when facing unknown

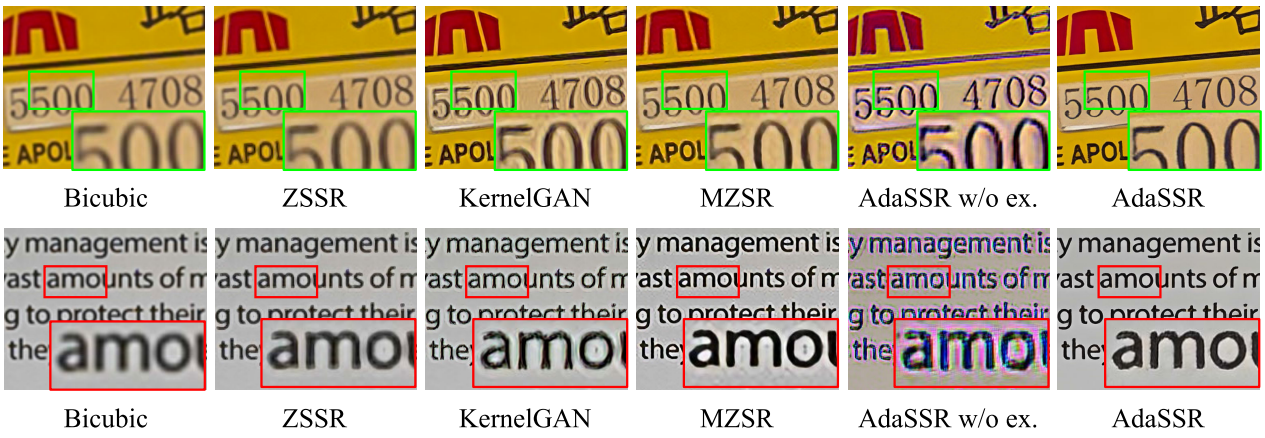


Fig. 8. Visualization comparison between internal learning methods, AdaSSR w/o external learning and AdaSSR.

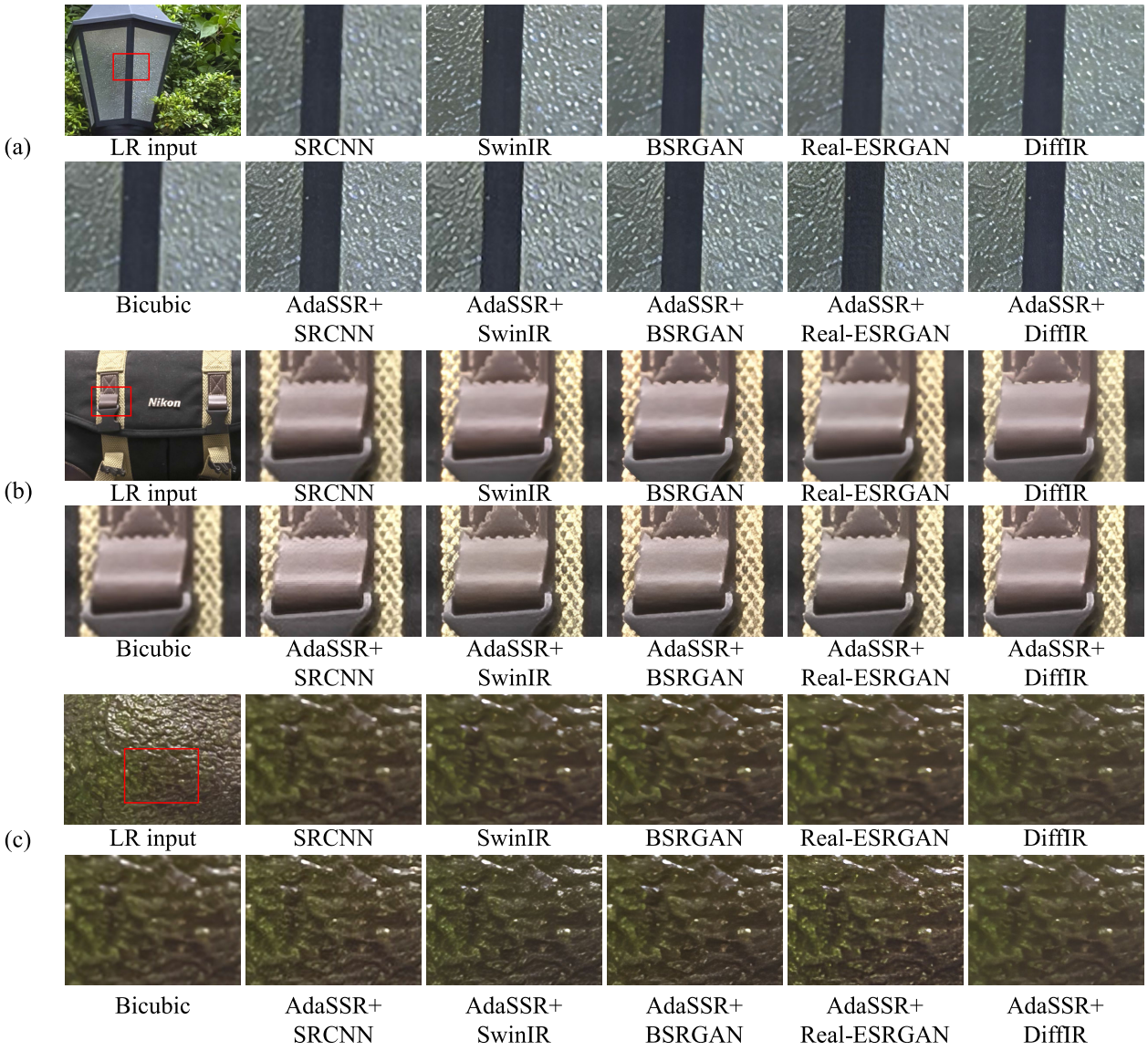


Fig. 9. The visualization comparison of different external learning methods before/after applying AdaSSR.

real-world degradation cases, leading to blurred, artifact-ridden, and low-quality SR results. The AdaSSR framework can provide specific domain knowledge tailored to the current scenario, resulting in notable improvement. As shown in Table VII and Fig. 9(a), we show the quantitative and visualization comparison of different external learning methods

TABLE III
THE EFFECT OF SELF-SIMILARITY ON PERFORMANCE

Image		(a)	(b)	(c)
Real-ESRGAN	NIQE ↓	4.52	4.24	6.59
	LPIPS ↓	0.182	0.368	0.378
AdaSSR	NIQE ↓	4.42	3.60	4.43
	LPIPS ↓	0.169	0.317	0.250
Difference (Δ)	NIQE ↓	-0.10	-0.64	-2.14
	LPIPS ↓	-0.013	-0.051	-0.128

TABLE IV
EXPERIMENTS ON THE EFFECT OF THE NUMBER OF ITERATIONS ON THE PERFORMANCE OF ALTERNATING UPDATING PROCESS

	NIQE ↓	BRISQUE ↓	LPIPS ↓
Θ_e^1	5.36	37.06	0.418
Θ^1	4.79	33.78	0.376
Θ_e^2	5.06	34.94	0.384
Θ^2	4.74	32.94	0.375
Θ_e^4	4.88	33.34	0.383
Θ^4	4.70	32.34	0.374

TABLE V
ABLATION STUDY ON PARAMETER α FOR EMA UPDATING

Value α	NIQE ↓	BRISQUE ↓	PI ↓	LPIPS ↓	NRQM ↑
0.9999	4.63	29.04	4.45	0.270	5.89
0.999	4.24	**24.47**	3.99	**0.266**	**6.34**
0.99	**4.12**	25.36	**3.96**	0.280	6.29
0.9	4.13	25.96	3.97	0.279	6.29

TABLE VI
THE IMPROVEMENT IN DIFFERENT EXTERNAL LEARNING METHODS ON REALSR DATASET OF SCALE ×4

Performance (LPIPS)	AdaSSR+ SRCNN	AdaSSR+ Real-ESRGAN	AdaSSR+ SWINIR	AdaSSR+ DiffIR	AdaSSR+ BSRGAN
Before AdaSSR	0.385	0.252	0.269	0.273	0.253
After AdaSSR	0.343	0.244	0.263	0.266	0.245
Difference Δ	-0.042	-0.008	-0.006	-0.007	-0.008

TABLE VII
THE IMPROVEMENT IN DIFFERENT EXTERNAL LEARNING METHODS ON NIKON_006_LR4 IN REALSR

Performance (LPIPS)	AdaSSR+ SRCNN	AdaSSR+ Real-ESRGAN	AdaSSR+ SWINIR	AdaSSR+ DiffIR	AdaSSR+ BSRGAN
Before AdaSSR	0.442	0.340	0.310	0.293	0.302
After AdaSSR	0.340	0.275	0.257	0.248	0.250
Difference Δ	-0.102	-0.065	-0.053	-0.045	-0.051

before/after applying AdaSSR on image Nikon_006_LR4 from the RealSR dataset. It can be observed that the results of SRCNN are too blurry, and the reconstruction results of Real-ESRGAN appear over-smooth, not aligning with the expected real-world HR results. AdaSSR brings a remarkable performance improvement on SRCNN and the second-largest improvement on Real-ESRGAN, with great improvements also seen in the other three methods.

6) Multi-Image Condition: Images captured by the same photographic equipment exhibit similarity in degradation conditions. Therefore, we also evaluated the performance of AdaSSR in multi-image scenarios, as shown in Table IX. The RealSR dataset comprises two sets of data captured by Canon and Nikon devices, each containing 50 groups of images. For the multi-image scenario, we train for 10,000 iterations for each capture type and then test the performance among images taken by the same shooting equipment.

TABLE VIII
THE COMPUTATION TIME OF ADASSR

Setting	AdaSSR+ Real-ESRGAN	AdaSSR+ SRCNN	AdaSSR+ Real-ESRGAN multi.
Inference Time	9min	2min	65s

It can be observed that, compared to the single-image learning setting, the improvement in the multi-image testing scenario is less significant. However, AdaSSR still brings a substantial improvement over the corresponding external learning methods with fewer total iterations.

7) Computation Time: We show the computation time of AdaSSR and identify several methods to accelerate AdaSSR in practical applications. As shown in Table VIII, we show the inference time for single real-world image based on the backbone Real-ESRGAN and SRCNN. We also show the inference time under the same setting of Table IX, which is referred to *AdaSSR+Real-ESRGAN multi.*. As indicated in Table I, we can use a lightweight backbone like SRCNN to surpass the results of larger external learning SR models such as RCAN and HAT. As demonstrated in the results of the multi-image case, for a set of images captured under the same unknown shooting conditions (same camera equipment), we can achieve significant performance improvements in reconstructing real images within a short training time. By using an external learning model and applying the AdaSSR framework for minimal, intermittent learning on images captured by the device, we can significantly enhance performance with a relatively low computational cost, thereby improving the usability of our method in real-world applications.

C. Ablation Study

We conduct the ablation study on the proposed design of AdaSSR, including the number of alternating iterations, the EMA updating parameter, and other components.

1) Number of Alternating Iterations: As demonstrated in Eq. (5)-(7), our algorithm updates parameters through an alternating updating process. Table IV demonstrates the impact of the number of iterations on performance. Experimental results are the average from 10 runs on images in the RealSR dataset.

It can be observed that, with an increase in the number of iterations, the performance of results obtained in Eq. (6) gradually improves. External learning and internal learning together contribute to the overall effectiveness. Considering higher efficiency, unless otherwise specified, the presented results in the paper are based on one round of iteration updates.

2) EMA Updating Parameter: We carry out an ablation study on α for EMA updating in the self-supervised learning part, as in Eq. (16). The experiment tests on RealSR with scale ×4. As shown in Table V, we find that the performance is optimal when alpha lies in the range around 0.99 to 0.999. When alpha is too large, such as 0.9999, or too small, such as 0.9, the performance deteriorates. This further confirms the significance of our research's focus on the combination

TABLE IX

QUANTITATIVE COMPARISON FOR MULTI-IMAGE CASE. CANON AND NIKON REPRESENT TWO DATA ACQUISITION SCENES IN THE REALSR DATASET. * MEANS THE INTERNAL LEARNING PROCESS IN ADASSR CARRIED OUT AMONG IMAGES TAKEN BY THE SAME SHOOTING EQUIPMENT

Sensors	Canon					Nikon				
Method	NIQE ↓	BRISQUE ↓	PI ↓	LPIPS ↓	NRQM ↑	NIQE ↓	BRISQUE ↓	PI ↓	LPIPS ↓	NRQM ↑
Bicubic	9.03	68.26	8.25	0.467	2.60	8.63	66.03	7.94	0.486	2.84
SwinIR	4.45	31.79	4.53	0.243	5.64	4.90	35.78	4.90	0.260	5.35
BSRGAN	4.53	25.34	4.36	0.257	6.03	4.77	25.37	4.57	0.280	5.94
Real-ESRGAN	4.45	26.35	4.25	0.261	6.07	4.90	31.92	4.73	0.285	5.67
AdaSSR+SwinIR	4.45	31.55	4.51	**0.240**	5.69	4.67	34.13	4.62	**0.249**	5.64
AdaSSR+BSRGAN	4.46	24.51	4.31	0.255	6.06	4.61	25.80	4.38	0.271	6.11
AdaSSR+Real-ESRGAN	**4.15**	**23.78**	**3.95**	0.258	**6.33**	4.34	25.16	4.04	0.275	6.34
AdaSSR+SwinIR*	4.43	31.79	4.52	0.242	5.65	4.60	34.44	4.62	0.255	5.60
AdaSSR+BSRGAN*	4.51	25.21	4.35	0.257	6.05	4.64	25.94	4.47	0.276	5.97
AdaSSR+Real-ESRGAN*	4.31	25.48	4.13	0.257	6.16	**4.27**	**21.37**	**3.92**	0.283	**6.37**

TABLE X

ABLATION STUDY ON COMPONENTS IN THE ADASSR FRAMEWORK

Backbone	Internal Learning	Kernel Estimation	Data Augmentation	Attention Map Guidance	NIQE ↓	BRISQUE ↓	PI ↓	LPIPS ↓	NRQM ↑
✓					4.68	29.13	4.49	0.273	5.87
✓	✓				4.40	26.15	4.17	0.267	6.10
✓	✓	✓			4.35	25.78	4.15	0.268	6.11
✓	✓	✓	✓		4.29	25.63	4.04	0.267	6.28
✓	✓	✓	✓	✓	4.24	24.47	3.99	0.266	6.34

between *internal learning* and *external learning*. The best reconstruction performance is achieved when both aspects are combined, while suboptimal results occur when either aspect dominates excessively.

3) Other Components: We conducted an ablation study on RealSR for key components of the AdaSSR framework, and the results are presented in Table X. The *Backbone* corresponds to the results of the pre-trained model from Real-ESRGAN, representing the model parameters Θ_e^1. *Internal learning* refers to carrying out random sampling of $(y_{r,d}, y_r)$ for model updates, where $y_{r,d}$ is obtained by bicubic downsampling of y_r. *Kernel estimation* incorporates the estimation of degradation conditions during internal learning, with $y_{r,d}$ obtained from the estimated blur kernel. *Data augmentation* and *attention map guidance* follow the descriptions in Section III. These components further enlarge the sample-adaptive advantage of internal learning, making them critical to the effectiveness of AdaSSR. From the performance exhibited across various evaluation metrics, each component of AdaSSR contributes significantly to the final results.

V. CONCLUSION

In this work, we propose a novel semi-supervised framework, AdaSSR, for real-world super-resolution tasks. We analyze that internal learning relies more on a sample-dependent degradation prediction, while the effectiveness of external learning is more dependent on the feature extraction capabilities learned from a diverse range of data. AdaSSR integrates the strengths of both internal learning and external learning in an ADMM iteration manner in combination with attention map guidance. AdaSSR has demonstrated outstanding performance in real-world super-resolution tasks, prompting inspiration for the integration of internal learning and external learning.

REFERENCES

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[2] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4681–4690.

[3] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Oct. 2021, pp. 1905–1914.

[4] X. Chen, X. Wang, J. Zhou, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 22367–22377.

[5] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Oct. 2021, pp. 1833–1844.

[6] B. Xia et al., "DiffIR: Efficient diffusion model for image restoration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 13049–13059.

[7] J.-N. Su, M. Gan, G.-Y. Chen, W. Guo, and C. L. P. Chen, "High-similarity-pass attention for single image super-resolution," *IEEE Trans. Image Process.*, vol. 33, pp. 610–624, 2024.

[8] Q. Cai et al., "HIPA: Hierarchical patch transformer for single image super resolution," *IEEE Trans. Image Process.*, vol. 32, pp. 3226–3237, 2023.

[9] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[10] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[11] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. IEEE Eur. Conf. Comput. Vis. Workshops*, Sep. 2018, pp. 63–79.

[12] B. Xia, Y. Hang, Y. Tian, W. Yang, Q. Liao, and J. Zhou, "Efficient non-local contrastive attention for image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 3, pp. 2759–2767.

[13] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 23593–23606.

[14] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2023.

[15] Y. Huang, S. Li, L. Wang, and T. Tan, "Unfolding the alternating optimization for blind super resolution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 5632–5643.

[16] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1604–1613.

[17] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3262–3271.

[18] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 701–710.

[19] M. Fritsche, S. Gu, and R. Timofte, "Frequency separation for real-world super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop*, Oct. 2019, pp. 3599–3608.

[20] L. Wang et al., "Unsupervised degradation representation learning for blind super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 10576–10585.

[21] I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[22] A. Shocher, N. Cohen, and M. Irani, "'Zero-shot' super-resolution using deep internal learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3118–3126.

[23] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-GAN," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.

[24] X. Cheng, Z. Fu, and J. Yang, "Zero-shot image super-resolution with depth guided internal degradation learning," in *Proc. IEEE Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 265–280.

[25] M. Zontak and M. Irani, "Internal statistics of a single natural image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 977–984.

[26] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Trans. Graph.*, vol. 30, no. 2, pp. 1–11, Apr. 2011.

[27] C.-Y. Yang, J.-B. Huang, and M.-H. Yang, "Exploiting self-similarities for single frame super-resolution," in *Proc. IEEE Asia Conf. Comput. Vis.*, Nov. 2011, pp. 497–510.

[28] T. Tirer and R. Giryes, "Super-resolution via image-adapted denoising CNNs: Incorporating external and internal learning," *IEEE Signal Process. Lett.*, vol. 26, no. 7, pp. 1080–1084, Jul. 2019.

[29] T. Tirer and R. Giryes, "Image restoration by iterative denoising and backward projections," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1220–1234, Mar. 2019.

[30] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3516–3525.

[31] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 184–199.

[32] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4799–4807.

[33] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.

[34] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[35] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 10012–10022.

[36] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2021.

[37] Q. Liu, P. Gao, K. Han, N. Liu, and W. Xiang, "Degradation-aware self-attention based transformer for blind image super-resolution," *IEEE Trans. Multimedia*, vol. 26, pp. 7516–7528, 2024.

[38] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: Self-supervised photo upsampling via latent space exploration of generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2437–2445.

[39] K. C. K. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "GLEAN: Generative latent bank for large-factor image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 14245–14254.

[40] T. Yang, P. Ren, X. Xie, and L. Zhang, "GAN prior embedded network for blind face restoration in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 672–681.

[41] R. Dahl, M. Norouzi, and J. Shlens, "Pixel recursive super resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5449–5458.

[42] A. Lugmayr, M. Danelljan, L. Van Gool, and R. Timofte, "SRFlow: Learning the super-resolution space with normalizing flow," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 715–732.

[43] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–10.

[44] Y. Ma, H. Yang, W. Yang, J. Fu, and J. Liu, "Solving diffusion ODEs with optimal boundary conditions for better image super-resolution," 2023, *arXiv:2305.15357*.

[45] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[46] J. Cai, H. Zeng, H. Yong, Z. Cao, and L. Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 3086–3095.

[47] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 4791–4800.

[48] H. F. Ates, S. Yildirim, and B. K. Gunturk, "Deep learning-based blind image super-resolution with iterative kernel reconstruction and noise estimation," *Comput. Vis. Image Understand.*, vol. 233, Aug. 2023, Art. no. 103718.

[49] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, "All-in-one image restoration for unknown corruption," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 17431–17441.

[50] S. Son, J. Kim, W.-S. Lai, M.-H. Yang, and K. M. Lee, "Toward real-world super-resolution via adaptive downsampling models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8657–8670, Nov. 2022.

[51] Z. Zhang, R. Wang, H. Zhang, and W. Zuo, "Self-supervised learning for real-world super-resolution from dual and multiple zoomed observations," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 20, 2024.

[52] H. Chen et al., "Low-res leads the way: Improving generalization for super-resolution by self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 25857–25867.

[53] Z. Yang et al., "A dynamic kernel prior model for unsupervised blind image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2024, pp. 26046–26056.

[54] S. Boyd et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[55] C. Ma, Y. Rao, Y. Cheng, C. Chen, J. Lu, and J. Zhou, "Structure-preserving super resolution with gradient guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 7769–7778.

[56] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[57] J. Liang, H. Zeng, and L. Zhang, "Details or artifacts: A locally discriminative learning approach to realistic image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 5657–5666.

[58] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2020, pp. 702–703.

[59] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 1122–1131.

[60] P. Wei et al., "Component divide-and-conquer for real-world image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 101–117.

[61] J. Liang, H. Zeng, and L. Zhang, "Efficient and degradation-adaptive network for real-world image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 574–591.

[62] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Apr. 2012.

[63] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[64] C. Ma, C.-Y. Yang, X. Yang, and M.-H. Yang, "Learning a no-reference quality metric for single-image super-resolution," *Comput. Vis. Image Understand.*, vol. 158, pp. 1–16, May 2017.
[65] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6228–6237.
[66] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[67] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Sep. 2009, pp. 349–356.

Zejia Fan received the B.S. degree in computer science from Peking University, Beijing, China, in 2021, where she is currently pursuing the Ph.D. degree with the Wangxuan Institute of Computer Technology. Her current research interests include image enhancement and deep learning.

Wenhan Yang (Member, IEEE) received the B.S. and Ph.D. (Hons.) degrees in computer science from Peking University, Beijing, China, in 2012 and 2018, respectively. He is currently an Associate Researcher with the Pengcheng Laboratory, Shenzhen, Guangdong, China. His current research interests include image/video processing/restoration, bad weather restoration, and human–machine collaborative coding. He has authored over 50 technical articles in refereed journals and proceedings and holds nine granted patents. He received the 2023 IEEE Multimedia Rising Star Runner-Up Award, the IEEE ICME-2020 Best Paper Award, the IFTC 2017 Best Paper Award, the IEEE CVPR-2018 UG2 Challenge First Runner-Up Award, and the MSA-TC Best Paper Award of ISCAS 2022. He was the Candidate of CSIG Best Doctoral Dissertation Award in 2019. He served as the Area Chair for IEEE ICME-2021/2022/2023/2024, the Session Chair for IEEE ICME-2021, and the Organizer for IEEE CVPR-2019/2020/2021 UG2+ Challenge and Workshop.

Zongming Guo (Member, IEEE) received the B.S. degree in mathematics and the M.S. and Ph.D. degrees in computer science from Peking University, Beijing, China, in 1987, 1990, and 1994, respectively. He is currently a Professor with the Wangxuan Institute of Computer Technology, Peking University. His current research interests include video coding, processing, and communication. He is an Executive Member of China-Society of Motion Picture and Television Engineers. He was a recipient of the First Prize of the State Administration of Radio Film and Television Award in 2004, the First Prize of the Ministry of Education Science and Technology Progress Award in 2006, the Second Prize of the National Science and Technology Award in 2007, the Wang Xuan News Technology Award and the Chia Tai Teaching Award in 2008, the Government Allowance granted by the State Council in 2009, and the Distinguished Doctoral Dissertation Advisor Award of Peking University in 2012 and 2013.

Jiaying Liu (Senior Member, IEEE) received the Ph.D. degree (Hons.) in computer science from Peking University, Beijing, China, in 2010.

She is currently an Associate Professor and a Boya Young Fellow with the Wangxuan Institute of Computer Technology, Peking University, China. She has authored more than 100 technical articles in refereed journals and proceedings and holds 70 granted patents. She was a Visiting Scholar with the University of Southern California, Los Angeles, CA, USA, from 2007 to 2008. Her current research interests include multimedia signal processing, compression, and computer vision. She was a Visiting Researcher with Microsoft Research Asia in 2015, supported by the Star Track Young Faculties Award. She is a Senior Member of IEEE/CSIG and a Distinguished Member of CCF. She has served as a member for Multimedia Systems and Applications Technical Committee (MSA TC) and Visual Signal Processing and Communications Technical Committee (VSPC TC) in IEEE Circuits and Systems Society. She received the IEEE ICME 2020 Best Paper Award and IEEE MMSP 2015 Top10% Paper Award. She has also served as the Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS SYSTEMS FOR VIDEO TECHNOLOGY, and *Journal of Visual Communication and Image Representation*; the Technical Program Chair for ACM MM Asia-2023/IEEE ICME-2021/ACM ICMR-2021/IEEE VCIP-2019; the Area Chair for CVPR-2021/ECCV-2020/ICCV-2019; an ACM ICMR Steering Committee Member; and the CAS Representative at the ICME Steering Committee. She was the APSIPA Distinguished Lecturer (2016–2017).