Component Divide-and-Conquer for Real-World Image Super-Resolution

Pengxu Wei[1], Ziwei Xie[1], Hannan Lu[2], Zongyuan Zhan[1], Qixiang Ye[3], Wangmeng Zuo[2], and Liang Lin[⋆1,4]

[1] Sun Yat-sen University, Guangzhou, China
[2] Harbin Institute of Technology, Harbin, China
[3] University of Chinese Academy of Sciences, Beijing, China
[4] DarkMatter AI
weipx3@mail.sysu.edu.cn xiezw5@mail2.sysu.edu.cn hannanlu@hit.edu.cn zhanzy178@gmail.com qxye@ucas.ac.cn wmzuo@hit.edu.cn linliang@ieee.org

Abstract. In this paper, we present a large-scale Diverse Real-world image Super-Resolution dataset, *i.e.*, DRealSR, as well as a divide-and-conquer Super-Resolution (SR) network, exploring the utility of guiding SR model with low-level image components. DRealSR establishes a new SR benchmark with diverse real-world degradation processes, mitigating the limitations of conventional simulated image degradation. In general, the targets of SR vary with image regions with different low-level image components, *e.g.*, smoothness preserving for flat regions, sharpening for edges, and detail enhancing for textures. Learning an SR model with conventional pixel-wise loss usually is easily dominated by flat regions and edges, and fails to infer realistic details of complex textures. We propose a Component Divide-and-Conquer (CDC) model and a Gradient-Weighted (GW) loss for SR. Our CDC parses an image with three components, employs three Component-Attentive Blocks (CABs) to learn attentive masks and intermediate SR predictions with an intermediate supervision learning strategy, and trains an SR model following a divide-and-conquer learning principle. Our GW loss also provides a feasible way to balance the difficulties of image components for SR. Extensive experiments validate the superior performance of our CDC and the challenging aspects of our DRealSR dataset related to diverse real-world scenarios. Our dataset and codes are publicly available at `https://github.com/xiezw5/Component-Divide-and-Conquer-for-Real-World-Image-Super-Resolution`

Keywords: Real-world Image Super-Resolution; Image Degradation; Corner Point; Component Divide-and-Conquer; Gradient-Weighted Loss.

1 Introduction

Single image Super-Resolution (SR) is an inherently ill-posed inverse problem that reconstructs High-Resolution (HR) images from Low-Resolution (LR) coun-

⋆ Corresponding Author

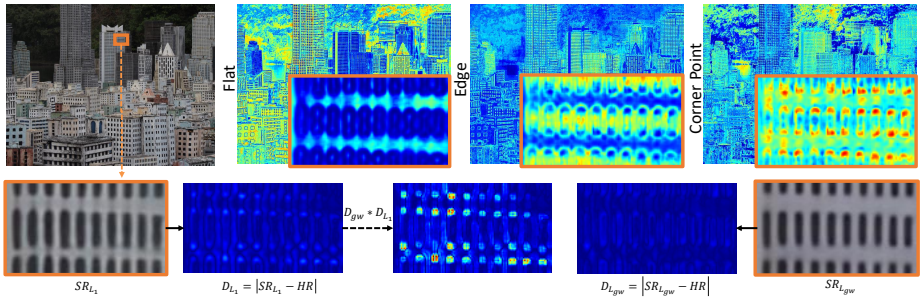

Fig. 1: Image degradation reflected by loss values. SR_{L_1} and $SR_{L_{gw}}$ are SR results trained with L_1 loss and our GW loss, respectively. The difference map D_{L_1} which presents different reconstruction difficulties demonstrates the complex image degradation. It is observed that regions from small to large values in D_{L_1} are relatively consistent with flat, edges and corner regions, which motivates us to explore these components and introduce a weighting strategy for D_{L_1} which drives models to attend to hard regions. The first row shows three attentive masks learnt by our CDC which well predict the confidence of flat, edges and corner regions, respectively.

terparts with image quality degradations. As a fundamental research topic, it has attracted a long-standing and considerable attention in the computer vision community [7][28]. SR methods based on Convolutional Neural Network (CNN) (*e.g.*, SRCNN [5], SRGAN [13], EDSR [14], ESRGAN [26] and RCAN [33]) have achieved a remarkable improvement over conventional SR methods [7].

However, such improvements remain limited for real-world SR applications. The first reason is that SR models have to be trained on datasets with simulated image degradation, as LR images are obtained by simplified downsampling methods (*e.g.*, bicubic) due to the difficulty of HR-LR pair collection. Such simulated degradation usually deviates from real ones, making the learned model not applicable to many real-world applications [32,2]. The second reason is that homogenous pixel-wise losses (*e.g.*, MSE) would lead to model overfitting or attend to regions for easy reconstruction. Intuitively, the targets of SR vary with LR regions with different low-level image elements, *e.g.*, smoothness preserving for flat regions, sharpening for edges, and detail enhancing for textures. Considering that flat regions and edges are the most frequent in an image, the models learned by homogeneous pixel-wise loss prefer addressing flat regions and edges, but usually fail to infer realistic details of complex textures. In Fig. 1, an SR image from EDSR [14] trained with L_1 loss presents different reconstruction difficulties in different regions, specifically in flat, edge and corner point regions. In Fig. 2, we analyze proportions of three components (flat, edges and corners) for L_1 loss in EDSR [14] and evaluate their respective effects for SR reconstruction with averaged pixel-wise loss. Three components are observed to have different recovery difficulties: smooth regions and edges have lower loss while corner points

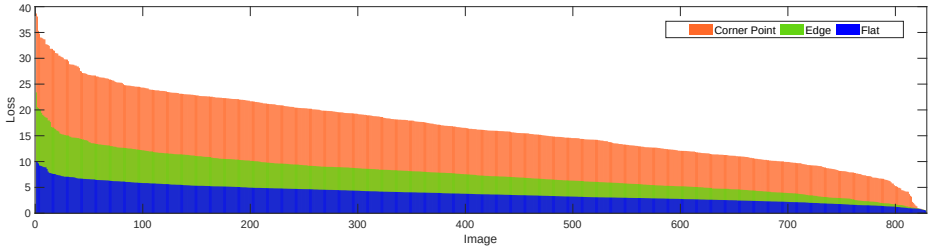

Fig. 2: Component analysis for real SR. To investigate the challenging aspects, we analyze proportions of three components (flat, edges and corners) for L_1 loss in EDSR [14] and evaluate their respective effects for SR reconstruction with averaged pixel-wise loss. Three components are observed to have different recovery difficulties: smooth regions and edges have a lower loss while corner points have a higher loss.

have higher loss. Thus, these observations inspire us to investigate the utility of these three components in the SR task.

In this paper, we establish a large-scale Diverse Real-world SR benchmark, DRealSR, and propose a Component Divide-and-Conquer model (CDC) to address real-world SR challenges, *i.e.*, (i) the gap between simulated and real degradation processes and (ii) the diverse image degradation processes caused by different devices. CDC is inspired by the mechanism of Harris corner point detection algorithm [19]. An image can be disentangled into three low-level components (*i.e.*, flat, edge and corner) with respect to the importance of the information they convey. Flat regions have almost constant pixel values, edges can be regarded as the boundary of different flat regions, and multiple edges interweave into corners. In CDC, three low-level elements which facilitate an implicit composition optimization are treated as guidance to regularize the SR task.

Specifically, we first develop a base model, named HGSR, based on a stacked hourglass network. HGSR learns multi-scale features with repeated bottom-up and top-down inference across all scales. With HGSR, CDC builds three Component-Attentive Blocks (CABs) which are associated with flat, edges and corners, respectively. Each CAB focuses on learning one of the three low-level components with the Intermediate Supervision (IS) strategy. CDC takes the flat regions, edges and corners extracted from HR images only in the training stage and then incorporates them separately into three different branches with CABs. These three CABs form a progressive paradigm and are aggregated to yield the final SR reconstruction. Considering that different image regions convey different gradients in all directions, we propose a Gradient-Weighted (GW) loss function for SR reconstruction. More complex a region is, larger impacts on the loss function it has. Our GW loss, in a way like Focal loss [15] for training object detectors, adapts the model training based on different image reconstruction difficulties.

In brief, our contributions are summarized as follows:

- A large-scale real-world SR benchmark (DRealSR), which is collected from five DSLR cameras. DRealSR mitigates the limits of conventional simulated image degradation and establishes a new SR benchmark related to real-world challenges.
- A Component Divide-and-Conquer model (CDC), which, inspired by corner point detection, aims at addressing real-world SR challenges in a divide-and-conquer manner. CDC employs three Component-Attentive Blocks to learn attentive masks for different components and predicts intermediate SRs with an intermediate supervision learning strategy.
- A Gradient-Weighted loss, which fully utilizes image structural information for fine-detailed image SR. GW loss explores the imbalance learning issue for different image regions in the pixel-wise SR task and provides a promising solution, which can be extended to other low-level vision tasks.

2 Related Work

Datasets. In the area of SR, widely-used SR datasets include Set5 [1], Set14 [29], BSD300 [17], Urban100 [9] and DIV2K [24]. Due to the difficulty of collecting HR-LR pairs, non-blind SISR approaches usually adopt a simulated image degradation for training and testing, *e.g.*, bicubic downsampling. Consequently, images in those SR datasets are usually regarded as HR images whose LR counterparts are obtained by HR downsampling. However, the real-world degradation process can be much more complex and even nonlinear. This simulated degradation limits related SR researches to a rather ideal SR simulation with approximately linear kernels and causes a great gap for practical SR applications [2][3][30][32].

To fill this gap, City100 dataset [3] with 100 aligned image pairs is built for SR modeling in the realistic imaging system. However, City100 is captured for the printed postcards under an indoor environment. To capture real-world natural scenes, SR-RAW dataset is introduced for super-resolution from raw data via optical zoom [32]. RealSR dataset [2] provides a well-prepared benchmark for real-world single image super-resolution, which is captured with two DSLR cameras. In this work, we build a larger and more challenging real SR dataset with five DSLR cameras, with the target to further explore SR degradation in real-world scenarios.

Methods. Recent years have witnessed an evolution of image super-resolution research with widely-explored deep learning, which has significantly improved SR performance against traditional methods [5]. Sequentially, deep SR networks derived from various CNN models, *e.g.*, VDSR [10], EDSR [14], SRRestNet [13], LapSR [12] and RCAN [33], are presented to further improve the SR performance. To regularize the model for ill-posed SR problem, several works are suggested to incorporate image priors, *e.g.*, edge detection [4][6], texture synthesis [20] or semantic segmentation [25]. Despite their progress, most existing approaches are still tested on synthesized SR datasets with bicubic downsampling or downsampling after Gaussian blurring, while few researches are devoted to real-world SR problems.

Recently, a contextual bilateral loss (CoBi) is introduced to mitigate the misalignment issue in a real-world SR-RAW dataset [32]. Besides, LP-KPN [2] proposes a Laplacian pyramid based method to deal with the non-uniform blur kernels for SR. However, it remains limited in considering the complexity and diversity of real degradation processes among different devices, hindering the applications of real-world SR. In this work, we neither train an SR model by treating uniformly all the pixels/regions/components in an image nor bias towards only edges or textures. We parse an image into three low-level components (flat, edge and corner), explore their different importance, develop a CDC model in a divide-and-conquer learning framework and propose a GW loss to adaptively balance the pixel-wise reconstruction difficulties.

3 DRealSR: A Large-scale Real-world SR Dataset

To further explore complex real-world SR degradation, we build a large-scale diverse SR benchmark, named DRealSR, by zooming DSLR cameras to collect real LR and HR images. DRealSR covers 4 scaling factors (*i.e.*, $\times 1 \sim \times 4$).

Dataset collection. These images are captured from five DSLR cameras (*i.e.,* Canon, Sony, Nikon, Olympus and Panasonic) in natural scenes and cover indoor and outdoor scenes avoiding moving objects, *e.g.*, advertising posters, plants, offices, buildings, *etc.* For each scaling factor, we adopt SIFT method [16] for image registration to crop an LR image to match the content of its HR counterpart. To refine the registration results, an image is cropped into patches and an iterative registration algorithm and brightness match are employed. To better facilitate the model training, considering that their image sizes are 4,000×6,000 or 3,888×5,184, these training images are cropped into 380×380, 272×272, 192×192 patches for $\times 2 \sim \times 4$, respectively. Since the misalignment between HR and LR possibly induces to severely blurry SR results, after each step of registration, we conduct a careful manual selection for patches. More details on the dataset construction are provided in the supplementary file.

Challenges in real-world SR. Due to the difficulty to capture high-resolution and low-resolution image pairs in real world, extensive SR methods are demonstrated on datasets with simulated image degradation (*e.g.*, bicubic downsampling). Compared with simulated image degradation, real-world SR exhibits the following new challenges.

- **More complex degradation against bicubic downsampling**. Bicubic downsampling simply applies the bicubic downsampler to an HR image to obtain the LR image. In real scenarios, however, downsampling usually is performed after anisotropic blurring, and signal-dependent noise may also be added. Thus, the acquisition of LR images suffers from both blurring, downsampling and noise. Also it is affected by in-camera signal processing (ISP) pipeline. This non-uniform property of realistic image degradation can be verified based on the reconstruction difficulty analysis of different image regions/components, Fig.2. Usually, SR models trained on bicubic degradation exhibit poor performance when handling real-world degradation.

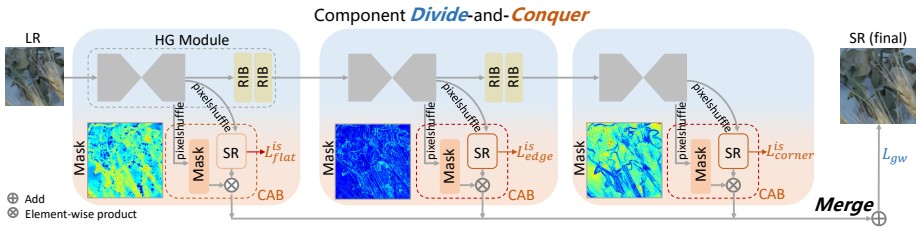

Fig. 3: CDC framework. The stacked architecture enables CDC to incorporate flat regions, edges and corners in three CABs separately. Each CAB branch produces an attentive mask and an intermediate SR. This mask regularizes the produced intermediate SR to collaborate with other branches and seamlessly blend them to yield natural SR images.

– **Diverse degradation processes among devices.** In practical scenarios, differences among lens and sensors of cameras determine the different imaging patterns, which is the primary reason for explaining the diverse degradation processes in real-world SR. Consequently, SR models learned on a real LR dataset may generalize poorly to other datasets and real-world LR images, raising another challenging issue for applications SR.

Nonetheless, different kinds of regions exhibit robustness characteristics to degradation. For example, a flat region is less affected by the diversity of the degradation process while changes in degradation settings produce quite different results for regions with edges and corners, Fig. 2. Accordingly, this motivates us to parse an image into flat, edges and corners for easing model training.

4 Real-word Image Super-Resolution

To handle diverse image degradation, one intuitive solution is to learn anisotropic blur kernels. Due to complex contents in a natural image, however, it is hard to propose a universal solution to estimate anisotropic kernels. Inspired by Harris corner points [8], image contents are divided into three primary visual components: flat, edges and corner points according to their gradient changes. These components are considered in our work. Because they represent the complexity of image contents, which indicates their reconstruction difficulty, as demonstrated in Fig. 1 and Fig. 2. For example, corners possess crucial orientation cues that control the shape or appearance of edges or textures [8] and are potentially beneficial for image reconstruction. Thus, these three components, *i.e.*, flat, edges and corners, are explored to facilitate SR model training being free from the limits to diverse degradation processes.

Specifically, in this work, considering the reconstruction difficulty of different components, we build an HGSR network with a stacked architecture and propose a Component Divide-and-Conquer model with a Gradient-Weighted loss to address the real SR problem. *Divide* arranges the introduction order as flat,

edges and corners to facilitate the feature learning in the network from easy to hard; *conquer* separately produces intermediate SR results for each component which are *merged* into the final SR prediction.

- *Divide*: Consider the complexity of image contents in flat, edges and corner point regions, we guide three HG modules to emphatically learn component-attentive masks from LR images respectively, with the component parsing guidance from HR images. It is noted that, we do not directly detect three components from LR images with off-the-shelf methods but predict their maps coherent with the HR image. The main reason is that the low quality of LR images hinders the more accurate corner point detection and yields undesirable detection results. Another reason is that this strategy avoids corner point detection for each image in the test stage.
- *Conquer*: Three Component-Attentive Blocks produce different component-attentive masks and intermediate SR predictions. The generated attentive maps present remarkable characteristics of three components. Meanwhile, intermediate SR results are consistent with the characteristics of three regions.
- *Merge*: To yield the final SR result, we collaboratively aggregate three intermediate SR outputs weighted by the corresponding component-attentive maps. In particular, a GW loss is proposed to drive the model to adapt learning objectives to their reconstruction difficulties.

4.1 Formulation

In the real SR, given N LR-HR pairs, we estimate an SR image $\hat{\mathbf{x}}_i$ by minimizing the loss function $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{rec}(\hat{\mathbf{x}}_i, \mathbf{x}_i)$, where $\mathcal{L}_{rec}(\cdot)$ is a reconstruction loss function. The network learns a mapping function $\mathcal{F}$ from the LR image $\mathbf{y}_i$ to the HR image $\mathbf{x}_i$; namely, $\hat{\mathbf{x}}_i = \mathcal{F}(\mathbf{y}_i; \Theta)$, where Θ is the network model parameter. In general, the realistic image degradation is complex and diverse as claimed above. To make it relatively tractable, our CDC employs three CABs and learn models with the intermediate supervision in a divide-and-conquer manner, rather than directly learning LR-HR mapping function or estimating blur kernels. Thus, our loss function is defined as follows,

$$\mathcal{L} = \frac{1}{N}\sum\nolimits_{i=1}^{N}[\mathcal{L}_{rec}(\hat{\mathbf{x}}_i, \mathbf{x}_i) + \sum\nolimits_{e=1}^{3}\mathcal{L}_{is}(\tilde{\mathbf{x}}_i^e, \mathbf{x}_i)], \tag{1}$$

where $\mathcal{L}_{is}$ is the intermediate loss function, the index e represents an CAB module that is specific to either *flat*, *edge* or *corner*, and $\tilde{\mathbf{x}}_i^e$ is the intermediate SR prediction ($\tilde{\mathbf{x}}_e$ for simplicity in the following sections).

4.2 Hourglass Super-Resolution Network

We propose a basemodel, Hourglass Super-Resolution network (HGSR), which has a stacked hourglass architecture [18] followed by a pixelshuffle layer [21]. The hourglass (HG) architecture is motivated to capture information at every

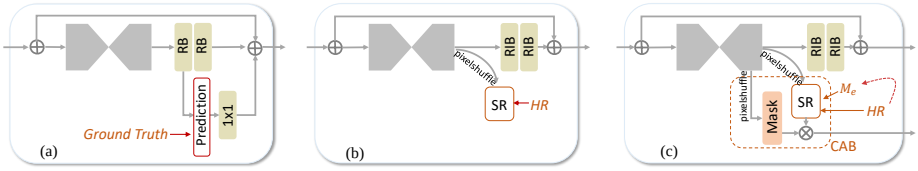

Fig. 4: CAB and intermediate supervision in the HG module. We compare (a) the IS in the basic HG [18] with those in HGSR (b) and our CDC (c). In [18], an intermediate prediction recursively joins into the next HG module after 1×1 convolution for human pose estimation. For SR, our IS strategy in (b) and (c) avoids the recursive operation which tends to invite large disturbance for feature learning in the backbone.

scale and has a superior performance for keypoint detection [18]. Its hourglass module can be regarded as an encoder-decoder with skip connections to preserve spatial information at each resolution and bring them together to predict pixel-wise outputs. In the HG module, an input passes through a convolutional layer firstly and then is downsampled to a lower resolution by a maximum pooling layer. During Top-Down inference, it repeats this procedure until reaching the lowest resolution. Next, a Bottom-Up inference performs constantly upsampling by nearest neighbor interpolation and combines features across scales by skip-connection layers until the original resolution is restored.

The conventional connection between two HG modules are two Residual Blocks (RBs) [18], as shown in Fig. 4(a). HGSR replaces RBs with Residual Inception Blocks (RIBs) [23]. RIBs have a parallel cascade structure and concatenate feature maps produced by filters of different sizes. Besides, HGSR utilizes the Intermediate Supervision (IS) strategy for model learning. The main difference of the IS module in [18] is that HGSR does not recursively feed the IS prediction to the next HG module, as shown in Fig. 4(b). The intermediate loss function $\mathcal{L}_{is}$ in HGSR is the $\mathcal{L}_1$ loss.

4.3 Component Divide-and-Conquer Model

Our Component Divide-and-Conquer model takes HGSR as the backbone and follows the divide-and-conquer principle to learn the model. Specifically, CDC focuses on three image components, *i.e.*, flat, edges and corners, rather than edges or/and complex textures. This makes it relatively tractable to solve the ill-posed real SR problem. These components are explicitly extracted from HR images with Harris corner detection algorithm, separately in CABs and are implicitly blended seamlessly to yield natural SR results by minimizing a GW loss. Although the guidance of three components is from HR images, CDC infers the component probability maps in the test stage without any detection.

Component-Attentive Block. CDC has three CABs which respectively correspond to either flat, edges or corners. Since it inherits the advantages of HGSR

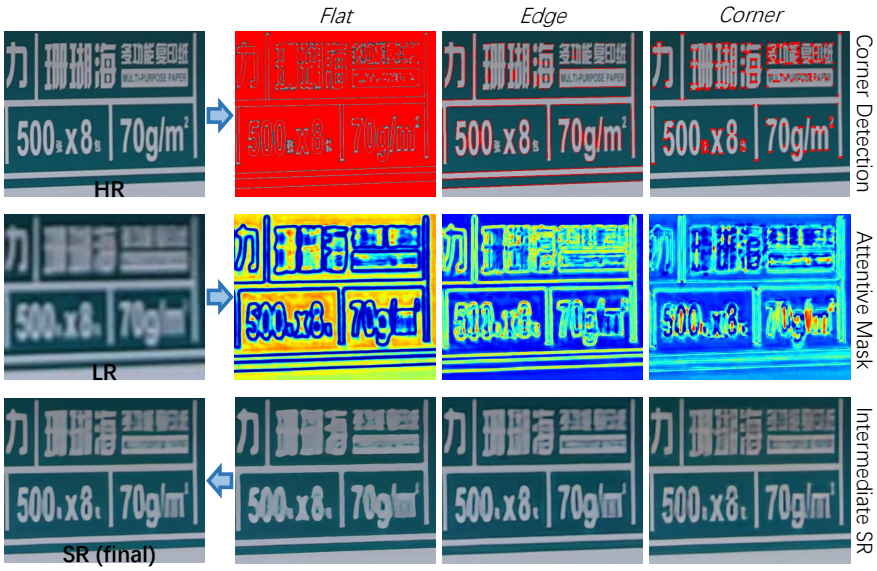

Fig. 5: Harris corner point detection, learned component-attentive masks and intermediate SR images from three CABs. Component-attentive masks from each CAB present a high similarity to flat, edges and corners, respectively.

with a cascaded hourglass network, CDC is suitable to incorporate the intermediate supervision. As shown in Fig.4(c), CAB consists of two pixel-shuffle layers. One is used to generate a coarse intermediate SR result. The other one is used to generate a mask which indicates the component probability map. It weights this coarse SR for the final SR reconstruction together with outputs of other CABs. In the training stage, CDC leverages the HR image as intermediate supervision to generate an IS loss weighted by the guidance of the component mask from HR. Accordingly, the intermediate loss function in an CAB is defined as

$$\mathcal{L}_{is} = l(\boldsymbol{M}_e * \mathbf{x}, \boldsymbol{M}_e * \tilde{\mathbf{x}}_e), \tag{2}$$

where e is similar to that defined in Equ. 1 and $*$ denotes the entry-wise product; $\boldsymbol{M}_e$ is the component guidance mask extracted from HR images. In general, $l(\cdot)$ can be any loss functions; we adopt widely-used L_1 loss function in the CDC.

As shown in Fig. 5, the learned component-attentive masks in three CABs exhibit their own characteristics in indicating flat regions, edges and textures, respectively. Accordingly, their intermediate SR results are also consistent with these characteristics. To further aggregate these three types of information, we will describe how to collaboratively aggregate them to yield the final SR result with a gradient-weighted loss.

Gradient-Weighted Loss. For conventional pixel-wise loss, regions in an image are treated identically. However, flat regions and edges dominate the loss function due to their large quantity in images. Thus, the learned SR models

incline to address flat regions and edges, but fail to infer realistic details of complex textures. Inspired by Focal loss [15], we propose to suppress a large number of simple regions while emphasizing the hard ones. Notably, this strategy is also crucial for low-level vision tasks. In our work, the solution of flat, edge and corner point detection provides a plausible disentanglement of images according to their importance, which can thus be used to determine the easy and hard regions and obtain the final SR prediction $\hat{\mathbf{x}}$ as the sum of outputs from three CABs, namely, $\hat{\mathbf{x}} = \sum_e \boldsymbol{A}_e * \tilde{\mathbf{x}}_e$, where $\boldsymbol{A}_e$ is a component-attentive mask.

We propose a Gradient-Weighted loss to dynamically adjust their roles for minimizing the SR reconstruction loss. Following this philosophy, the flat and single edge regions are naturally classified as simple regions. Corners are categorized as difficult regions since they possess the fine-details in images. Considering the diversity in the first-order gradient of different regions, the new reconstruction loss function for SR, named GW loss, is defined as

$$\mathcal{L}_{gw} = l(D_{gw} * \mathbf{x}, D_{gw} * \hat{\mathbf{x}}), \tag{3}$$

where $D_{gw} = (1+\alpha D_x)(1+\alpha D_y)$; $D_x = |G_x^{sr} - G_x^{hr}|$ and $D_y = |G_y^{sr} - G_y^{hr}|$ represent gradient difference maps between SR and HR in the horizontal and vertical directions; α is a scalar factor to determine the quantity for this weighting in the loss function. Generally, $l(\cdot)$ can be also any loss function and we adopt L_1 loss in this paper. If $\alpha = 0$, GW loss becomes the original loss $l(\mathbf{x}, \hat{\mathbf{x}})$. α is 4 in our experiments. This GW loss is regarded as the reconstruction loss $\mathcal{L}_{rec}$.

5 Experiments

5.1 Experimental settings

Dataset. We conduct experiments on an existing real-world SR dataset, RealSR, and our DRealSR. **RealSR** [2] has 595 HR-LR image pairs captured from two DSLR cameras. 15 image pairs at each scaling factor of each camera are selected randomly for building the testing set and the rest pairs are training set. Their image sizes are in the range of [700, 3100] and [600, 3500] and each training image is cropped in 192×192 patches. For ×2∼×4, our **DRealSR** has 884, 783 and 840 image pairs respectively, where 83, 84 and 93 image pairs are randomly selected for testing respectively and the rest are for training at each scaling factor.

Network Architecture. CDC cascades six HG modules. In each HG module, a residual block followed by a max-pooling layer for the top-down process and nearest neighbor method for the bottom-up process. For shortcut connection across two adjacent resolutions, we also use a residual block consisting of three convolution layers: 1×1, 3×3 and 1×1 filters. Between two hourglass modules, there are two connection layers using Residual Inception Block [23] for multi-scale processing. To introducing the intermediate supervision, those six HG modules are divided into three groups and the last HG in a group generates a coarse SR image by an upsampling layer of pixelshuffle.

Implementation Details. Harris Corner detection method [8] with OpenCV is used. In our experiments, we use Adam optimizer [11] and set 0.9 and 0.999 for its exponential decay rates. The initial learning rate is set to 2e-4 and then reduced to half every 100 epochs. For each training batch, we randomly extract 16 LR patches with the size of 48 × 48. All of our experiments are conducted in PyTorch. Three common image quality metrics are used to evaluate SR models, *i.e.*, peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [27] and Learned Perceptual Image Patch Similarity (LPIPS) [31]. PSNR is evaluated on the Y channel; SSIM and LPIPS are on RGB images.

5.2 Model Ablation Study

Evaluation on HG blocks. Our base model, HGSR, adopts a stacked hourglass network [18] as the backbone. We provide experimental evaluations on the number of HG blocks in Table 1. It is observed that the SR performance has a sustaining boost when the number of HG blocks in HGSR increases from 2 to 4 while it has a stable performance when the number of HG blocks is larger than 6. Thus, the number of HG blocks is set 6 in our experiments.

Table 1: Evaluation of the number of HG blocks

Method	HG Blocks	PSNR	SSIM	LPIPS
HGSR	2	31.55	0.847	0.336
	4	31.80	0.854	0.312
	6	**31.95**	**0.854**	0.304
	8	31.94	0.854	**0.303**

Table 2: Ablation study of the proposed CDC

Method	PSNR	SSIM	LPIPS
HGSR(baseline, w/o IS)	31.95	0.854	0.304
HGSR(baseline)	32.13	0.855	0.310
HGSR+RIB	32.15	0.857	0.310
HGSR+RIB+CAB	32.27	0.858	0.302
HGSR+RIB+CAB+GW	**32.42**	**0.861**	**0.300**

Table 3: Evaluation of CDC in flat, edge, and corner regions

Method	Regions			PSNR	SSIM	LPIPS
	flat	*edge*	*corner*			
CDC	✓			32.03	0.856	0.310
		✓		32.25	0.858	0.307
			✓	32.37	0.861	0.302
	✓	✓		32.23	0.860	0.301
	✓		✓	32.39	0.861	0.300
		✓	✓	32.40	0.861	**0.298**
	✓	✓	✓	**32.42**	**0.861**	0.300

Table 4: Comparison results of our proposed GW loss with L_1 loss

Method	Loss	PSNR	SSIM	LPIPS
SRResNet [13]	L_1	31.63	0.847	0.341
	L_{gw}	**31.93**	**0.853**	**0.321**
EDSR [14]	L_1	32.03	**0.855**	0.307
	L_{gw}	**32.27**	0.857	**0.304**
HGSR(Our baseline)	L_1	32.15	0.857	**0.310**
	L_{gw}	**32.25**	**0.857**	0.313
CDC(Ours)	L_1	32.27	0.858	0.302
	L_{gw}	**32.42**	**0.861**	**0.300**

Evaluation on IS and RIB. We leverage the intermediate supervision strategy to hierarchically supervise the model learning. As shown in Table 2, HGSR with IS achieves 0.18dB PSNR gains. In the following parts, if no special claim, HGSR denotes the base model with IS. Besides, two convolution layers are added between the two HG modules in HGSR to build their connections. To aggravate multi-scale information, these two layers are substituted by two RIBs. This modification slightly improves the base model with 0.02dB PSNR gains. Besides, our

Table 5: Performance comparison on RealSR [2] and DRealSR datasets

Method	Scale	Training Set: DRealSR						Training Set: RealSR					
		Test on RealSR [2]			Test on DRealSR			Test on RealSR [2]			Test on DRealSR		
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Bicubic	×2	31.67	0.887	0.223	32.67	0.877	0.201	31.67	0.887	0.223	32.67	0.877	0.201
SRResNet [13]		32.65	0.907	0.169	33.56	0.900	0.163	33.17	0.918	0.158	32.85	0.890	0.172
EDSR [14]		32.71	0.906	0.172	34.24	0.908	0.155	33.88	0.920	0.145	32.86	**0.891**	0.170
ESRGAN [26]		32.25	0.900	0.185	33.89	0.906	0.155	33.80	0.922	0.146	32.70	0.889	0.172
RCAN [33]		**32.88**	0.908	0.173	34.34	0.908	0.158	33.83	0.923	0.147	**32.93**	0.889	0.169
LP-KPN [2]		32.14	-	-	33.88	-	-	-	-	-	-	-	-
DDet [22]		32.58	-	-	33.92	-	-	33.22	-	-	32.77	-	-
CDC (Ours)		32.81	**0.910**	**0.167**	**34.45**	**0.910**	**0.146**	**33.96**	**0.925**	**0.142**	32.80	0.888	**0.167**
Bicubic	×3	28.63	0.809	0.388	31.50	0.835	0.362	28.61	0.810	0.389	31.50	0.835	0.362
SRResNet [13]		28.85	0.832	0.290	31.16	0.859	0.272	30.65	0.862	0.228	31.25	0.841	0.267
EDSR [14]		29.50	0.841	0.266	32.93	0.876	0.241	30.86	0.867	0.219	31.20	0.843	**0.264**
ESRGAN [26]		29.57	0.841	0.266	32.39	0.873	0.243	30.72	0.866	0.219	31.25	0.842	0.268
RCAN [33]		**29.68**	0.841	0.267	33.03	0.876	**0.241**	30.90	0.864	0.225	31.76	0.847	0.268
LP-KPN [2]		29.20	-	-	32.64	-	-	30.60	-	-	**31.79**	-	-
DDet [22]		29.48	-	-	32.13	-	-	30.62	-	-	31.77	-	-
CDC (Ours)		29.57	**0.841**	**0.261**	**33.06**	**0.876**	0.244	**30.99**	**0.869**	**0.215**	31.65	**0.847**	0.276
Bicubic	×4	27.24	0.764	0.476	30.56	0.820	0.438	27.24	0.764	0.476	30.56	0.820	0.438
SRResNet [13]		27.63	0.785	0.368	31.63	0.847	0.341	28.99	0.825	0.281	29.98	0.822	0.347
EDSR [14]		27.77	0.792	0.339	32.03	0.855	0.307	29.09	0.827	0.278	30.21	0.817	**0.344**
ESRGAN [26]		27.82	0.794	0.340	31.92	0.857	0.308	29.15	0.826	0.279	30.18	0.821	0.353
RCAN [33]		27.93	0.795	0.341	31.85	0.857	0.305	29.21	0.824	0.287	30.37	0.825	0.349
LP-KPN [2]		27.79	-	-	31.58	-	-	28.65	-	-	**30.75**	-	-
DDet [22]		27.83	-	-	31.57	-	-	28.94	-	-	30.12	-	-
CDC (Ours)		**28.11**	**0.800**	**0.330**	**32.42**	**0.861**	**0.300**	**29.24**	**0.827**	**0.278**	30.41	**0.827**	0.357

CAB and the GW loss have 0.12 and 0.15 dB PSNR improvements, respectively. Particularly, our final version, *i.e.*, CDC, exhibits an impressive improvement (*i.e.*, 0.29dB) compared with the base model HGSR.

Evaluation on Component-Attentive Block. As demonstrated in Table 2, our CAB brings an improvement of 0.12 dB in PSNR. In order to analyze CAB, we conduct experiments on different guidance from flat, edge and corner regions, as shown in Table 3. Without corner branches, our model has a significant drop of 0.19 dB by PSNR. Among three components (*i.e.*, flat, edges and corners), corners that represent important information play a crucial role in the SR task, although they have a small quantity in an image. This observation is encouraging to pay more attention to exploring corner points in the SR task, as well as directly on edges or/and textures.

Evaluation on Gradient-Weighted Loss. In Table 4, in comparison with L_1 loss, the GW loss respectively introduces 0.30 dB, 0.25 dB, 0.10 dB and 0.15 dB improvement in PSNR for EDSR, SRResNet, HGSR, and our CDC. This indicates that our proposed GW loss can be applied to other SR models to further improve their performance. Notably, the GW loss rooted in L_1 loss achieves a greater improvement than L_1. Therefore, our GW loss provides a new way to understand the SR model learning and can be explored in other loss functions and other low level vision tasks.

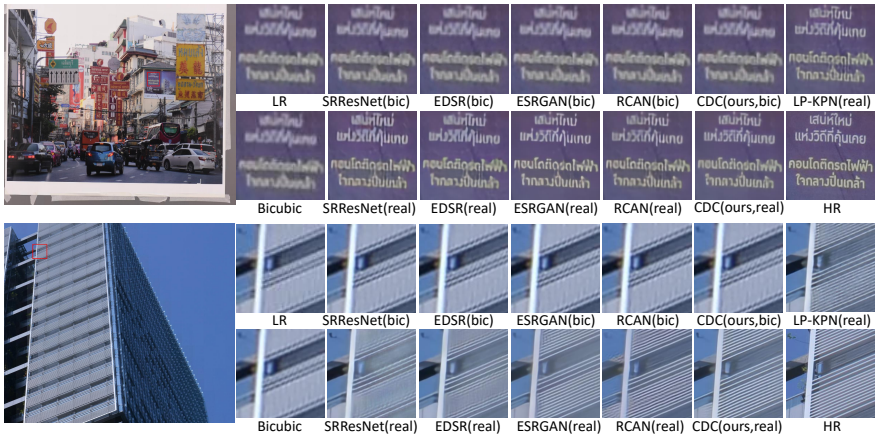

Fig. 6: SR results for $\times 4$ on DRealSR in comparison with state-of-the-art approaches. 'real' indicates models trained on DRealSR while 'bic' indicates those trained on a dataset version by bicubic downsampling DRealSR HR images.

5.3 Comparison with state-of-the-arts on real SR datasets

We compare our method with several state-of-the-art SR methods, including SRResNet [13], EDSR [14], ESRGAN [26], RCAN [33], LP-KPN [2] and DDet [22]. Among these SR methods, LP-KPN [2] and DDet [22] are the only two designed to solve the real-world SR problem. Quantitative comparison results are given in Table 5. LP-KPN [2] and DDet [22] are trained on the Y channel and other methods are trained on RGB images. Considering this difference, LPIPS and SSIM of LP-KPN and DDet are not provided since these two metrics are evaluated on RGB images. Our CDC outperforms the state-of-the-art algorithms on two real-world SR datasets. On DRealSR, CDC achieves the best results in all scales and notably improves the performance by about 0.4 dB for $\times 4$. Similar to the performance on DRealSR, PSNR and SSIM of CDC on RealSR are also superior to the others, validating the effectiveness of our method.

Fig. 6 visualizes SR results of the competing methods and ours. It is observed that existing SR methods (*e.g.*, EDSR, RCAN, LP-KPN) are prone to generate realistic detailed textures with visual aliasing and artifacts. For instance, in the second example in Fig. 6, SRRestNet, EDSR and LP-KPN produce blurry details of the building and the result of RCAN has obvious aliasing effects. In comparison, our proposed CDC reconstructs sharp and natural details.

To further validate challenges of RealSR and our DRealSR, we also conduct the cross-testing on the two datasets, *i.e.*, training models on one of them and then testing on the other one. In Table 5, CDC trained on DRealSR maintains a superior performance in all scales when tested on RealSR. However, for models trained on RealSR, the testing performance drops greatly on DRealSR, especially for $\times 4$, which indicates that DRealSR is more challenging than RealSR.

Table 6: SR performance comparison upon the image degradation

Method	Training set		Testing set (DRealSR)					
			Bicubic			Real		
	Dataset	Degradation	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
SRResNet [13]	DRealSR	Bicubic	41.28	0.954	0.103	30.61	0.822	0.422
EDSR [14]	DRealSR	Bicubic	41.49	0.956	0.099	30.60	0.822	**0.421**
RRDB [26]	DRealSR	Bicubic	41.66	0.957	0.097	30.60	0.822	0.425
CDC (Ours)	DRealSR	Bicubic	**41.78**	**0.957**	**0.096**	**30.63**	**0.822**	0.425
SRResNet [13]	DRealSR	Real	31.43	0.864	0.249	31.63	0.847	0.341
EDSR [14]	DRealSR	Real	32.53	0.880	0.231	32.03	0.855	0.307
RRDB [26]	DRealSR	Real	32.37	0.877	0.234	31.92	0.857	0.308
CDC (Ours)	DRealSR	Real	**32.54**	**0.883**	**0.215**	**32.42**	**0.861**	**0.300**

5.4 Analysis on Real and Simulated SR Results

In this section, we analyze the bicubic and real image degradation on DRealSR. In Table 6, the performance of real image degradation on our dataset is very close to that of bicubic image degradation even if the training set is different. Actually, the performance on PSNR, SSIM and LPIPS is close to that of bicubic upsampling method, as shown in Table 5. Thus, no matter which model is used, it is not useful to restore the real image with the model trained on bicubic images. This demonstrates the limited generalization of simulated bicubic degradation. On the other hand, our proposed CDC still achieves an improvement on bicubic images and outperforms most of state-of-the-arts methods. This is also the evidence to prove the superiority and generalization of our method. Fig. 6 visualizes SR results from models trained on simulated SR datasets. One can see that models trained on bicubic degradation produce blurry and poor SR results. This clearly demonstrates that the image degradation of the simulated SR dataset greatly hinders the performance of SR methods in real-world scenarios.

6 Conclusion

In this paper, we establish a large-scale real-world image super-resolution dataset, named DRealSR, to facilitate the further researches on realistic image degradation. To mitigate the complex and diverse image degradation, considering reconstruction difficulty of different components, we build a HGSR network with a stacked architecture and propose a Component Divide-and-Conquer model (CDC) to address the real SR problems. CDC employs three Component-Attentive Blocks (CABs) to learn attentive masks and intermediate SR predictions with an intermediate supervision learning strategy. Meanwhile, a Gradient-Weighted loss is proposed to drive the model to adapt learning objectives to their reconstruction difficulties. Extensive experiments validate the challenging aspects of our DRealSR dataset related to real-world scenarios, while our divide-and-conquer solution and GW loss provide a novel impetus for the challenging real-world SR task or other low-level vision tasks.

References

1. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: British Machine Vision Conference. pp. 1–10 (2012)
2. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: International Conference on Computer Vision (2019)
3. Chen, C., Xiong, Z., Tian, X., Zha, Z., Wu, F.: Camera lens super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1652–1660 (2019)
4. Dai, S., Han, M., Wu, Y., Gong, Y.: Bilateral back-projection for single image super resolution. In: IEEE International Conference on Multimedia and Expo. pp. 1039–1042 (2007)
5. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European conference on computer vision. pp. 184–199 (2014)
6. Fan, Y., Gan, Z., Qiu, Y., Zhu, X.: Single image super resolution method based on edge preservation. In: International Conference on Image and Graphics. pp. 394–399 (2011)
7. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: 2009 IEEE 12th international conference on computer vision. pp. 349–356 (2009)
8. Harris, C.G., Stephens, M., et al.: A combined corner and edge detector. In: Alvey vision conference (1988)
9. 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)
10. Kim, J., Kwon Lee, J., Mu Lee, K.: 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)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2017)
12. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 624–632 (2017)
13. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 105–114 (2017)
14. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1132–1140 (2017)
15. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
17. Martin, D.R., Fowlkes, C.C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings of the Eighth International Conference On Computer Vision. pp. 416–425 (2001)

18. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499 (2016)
19. Rosten, E., Porter, R., Drummond, T.: Faster and better: A machine learning approach to corner detection. IEEE transactions on pattern analysis and machine intelligence **32**(1), 105–119 (2008)
20. Sajjadi, M.S., Scholkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4491–4500 (2017)
21. 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)
22. Shi, Y., Zhong, H., Yang, Z., Yang, X., Lin, L.: Ddet: Dual-path dynamic enhancement network for real-world image super-resolution. IEEE Signal Processing Letters **27**, 481–485 (2020)
23. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI Conference on Artificial Intelligence (2017)
24. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 114–125 (2017)
25. Wang, X., Yu, K., Dong, C., Change Loy, C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 606–615 (2018)
26. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision. pp. 0–0 (2018)
27. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
28. 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)
29. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: International conference on curves and surfaces. pp. 711–730 (2010)
30. Zhang, K., Zuo, W., Zhang, L.: Learning a single convolutional super-resolution network for multiple degradations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3262–3271 (2018)
31. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
32. Zhang, X., Chen, Q., Ng, R., Koltun, V.: Zoom to learn, learn to zoom. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3762–3770 (2019)
33. 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. pp. 286–301 (2018)