Learning Omni-frequency Region-adaptive Representations for Real Image Super-Resolution

Xin Li*, Xin Jin* Tao Yu, Simeng Sun, Yingxue Pang, Zhizheng Zhang, Zhibo Chen†
CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System,
University of Science and Technology of China
Hefei 230027, China
{lixin666, jinxustc, yutao666, smsun20, pangyx, zhizheng}@mail.ustc.edu.cn, chenzhibo@ustc.edu.cn

Abstract

Traditional single image super-resolution (SISR) methods that focus on solving *single* and *uniform* degradation (*i.e.*, bicubic down-sampling), typically suffer from poor performance when applied into real-world low-resolution (LR) images due to the complicated realistic degradations. The key to solving this more challenging real image super-resolution (RealSR) problem lies in learning feature representations that are both *informative* and *content-aware*. In this paper, we propose an **O**mni-frequency **R**egion-adaptive **Net**work (OR-Net) to address both challenges, here we call features of all low, middle and high frequencies omni-frequency features. Specifically, we start from the frequency perspective and design a Frequency Decomposition (FD) module to separate different frequency components to comprehensively compensate the information lost for real LR image. Then, considering the different regions of real LR image have different frequency information lost, we further design a Region-adaptive Frequency Aggregation (RFA) module by leveraging dynamic convolution and spatial attention to adaptively restore frequency components for different regions. The extensive experiments endorse the effective, and scenario-agnostic nature of our OR-Net for RealSR.

Introduction

With the development of deep learning, single image super-resolution (SISR) has achieved great success either on PSNR values (Dong et al. 2015; Haris, Shakhnarovich, and Ukita 2018; Kim, Kwon Lee, and Mu Lee 2016; Lim et al. 2017; Zhang et al. 2018a; Dai et al. 2019; Mei et al. 2020; Pan et al. 2020) or on visual quality (Ledig et al. 2017; Sajjadi, Scholkopf, and Hirsch 2017). In general, these traditional SISR methods typically focus on restoring the low-resolution (LR) image with *single* and *uniform* synthetic degradation, such as bicubic down-sampling and Gaussian down-sampling. However, the degradations in real-world LR images are usually far more complicated, which makes most SISR models become less effective when directly applied to practical scenarios.

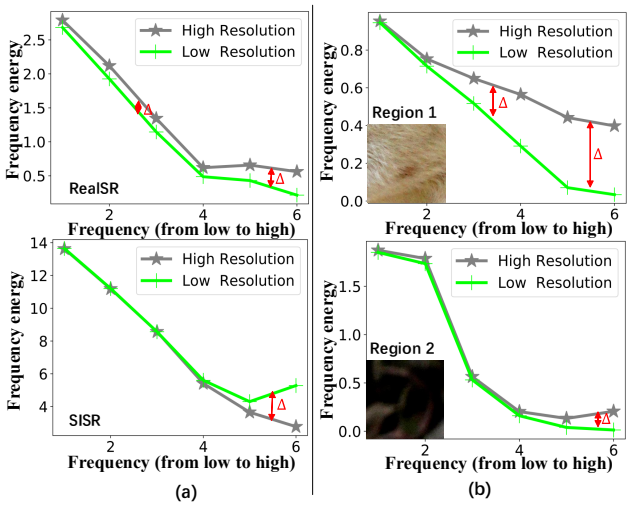

Figure 1: Difference analysis in frequency domain between RealSR and the conventional SISR: (a) the overall frequency distribution comparison between RealSR and SISR. (b) the frequency components distribution comparison between different regions of the same real LR image. Δ means the absolute difference of frequency degradation.

In recent years, some studies that focus on solving the real image super-resolution (RealSR) problem have attracted more and more attention (Cai et al. 2019; Wei et al. 2020). Unlike the *single* and *uniform* synthetic degradation in SISR, the LR and HR images of RealSR are captured with digital single lens reflex (DSLR) cameras, which typically contains various/complex non-uniform real-world degradations, including blur, noise, and down-sampling. That is why those classic conventional SISR methods (RCAN (Zhang et al. 2018a), EDSR (Lim et al. 2017), SAN (Dai et al. 2019) *etc.*) cannot handle the RealSR problem well.

Among recent representative RealSR methods, Cai *et al.* (Cai et al. 2019) first construct a RealSR dataset by capturing the LR-HR image pairs on the same scene with different focal lengths of a digital camera. And then, they propose a Laplacian pyramid based kernel prediction network (LP-KPN) to handle such non-uniform degradation. Concur-

*The first two authors contribute equally to this work.
†Corresponding author.

rently, Wei *et al*. (Wei et al. 2020) also propose a large-scale diverse real-world image super-resolution dataset named DRealSR. Considering the targets of RealSR vary with image regions, they further design a component divide-and-conquer (CDC) model to adaptively restore the LR image. But, LP-KPN and CDC both only focus on super-resolving LR images by collaborating different pixel-wise local restorations (*e.g*., flat regions, edges and corners), they ignore to make full use of hierarchical features across different frequency domains to comprehensively enhance texture details for real LR images.

In this paper, we first analyse the latent and essential challenges of RealSR from a completely different perspective of frequency distributions. In particular, we perform an experimental comparison between RealSR and conventional SISR in the frequency domain to explore their difference. Specifically, we visualize two LR-HR pairs in frequency domain through the wavelet transform tool, where one pair belongs to SISR and the other belongs to RealSR. As shown in Figure 1(a), we observe that the degradation (from HR image to LR image) of the general SISR mainly exists in the high-frequency component. In contrast, the degradation of RealSR exists in all frequency components. Besides, in Figure 1(b), we also see that the degradation of different regions of an image are usually distributed in the different frequency components.

Based on the above analysis, we argue that an effective RealSR model should learn feature representations that are both *informative* and *content-aware*. *Informative* property promises to restore sufficient realistic texture details across multiple frequency domains for the degraded LR, and *content-aware* property satisfies the varied targets of RealSR for different image regions (*i.e*., smoothing for flat regions and sharpening for edges). In this paper, we propose a **O**mni-frequency **R**egion-adaptive **Net**work (OR-Net) to efficiently solve the RealSR problem. Specifically, we first design a Frequency Decomposition (FD) module to decompose LR image into low-frequency, middle frequency , and high frequency components. Then, we employ multiple interactive branches to enhance the corresponding frequency factors. Second, to achieve the content-aware super-resolution for real LR images, we further design a Region-adaptive Frequency Aggregation (RFA) module by combining the dynamic convolution and spatial attention to selectively restore different frequency components for the different positions of HR images. The contributions of this paper can be summarized as follows:

- We analyse the essential difference between generalSR and RealSR from the frequency perspective, to answer the question that why the classic SISR methods cannot handle RealSR problem well.
- Based on our analysis, we propose an **O**mni-frequency **R**egion-adaptive **Net**work (OR-Net) for RealSR, which contains two technical novelties–1) Frequency Decomposition (FD) module that aims to achieve the LR image content separation in frequency domain and enhance texture details across all frequency components, 2) Region-adaptive Frequency Aggregation (RFA) module that aims to appropriately restore different frequency components for real HR images in different positions.
- Extensive experiments on multiple RealSR benchmarks have validated the effectiveness and superiority of our OR-Net. Sufficient intuitive visualization results/analysis are also provided to support/verify the expected functions of the proposed FD and RFA modules.

Related Work

Conventional Single Image Super-Resolution

In the last decade, the traditional SISR has achieved great progress, especially for deep learning based approaches (Dong et al. 2015; Dong, Loy, and Tang 2016; Lim et al. 2017; Haris, Shakhnarovich, and Ukita 2018; Zhang et al. 2018a; Dai et al. 2019). These methods usually perform well on the the synthetic degradation (*e.g*., bicubic down-sampling) but generalize poorly to realistic complicated distortions in real-world scenarios. This is problematic especially in practical applications, where the target scenes typically have hybrid/complex non-uniform degradations (*e.g*., blur, noise, and down-sampling), and also, there is always no readily available paired data for training.

Real Image Super-Resolution

RealSR has drawn more and more attention in recent years. Different from the general SISR that typically focuses on the simple and uniform synthetic degradation, RealSR mainly aims to solve these realistic complicated degradations in real-world scenarios. To capture the distortion of real scene, Chen *et al*. (Chen et al. 2019a) design two novel data acquisition strategies. Cai *et al*. (Cai et al. 2019) build a real-world super-resolution (RealSR) dataset by adjusting the focal length of a digital camera, and introduce the Laplacian pyramid based kernel prediction network (LP-KPN) to solve such non-uniform distortions. Recently, Wei *et al*. (Wei et al. 2020) present a large-scale diverse real-world image super-resolution dataset (DRealSR) and a component divide-and-conquer (CDC) model with gradient weighted (GW) loss, achieving great performance in RealSR.

However, above methods ignore to consider and study the difference between general SISR and RealSR. They didn't design solutions based on the essential difference analysis, which is sub-optimal and un-targeted. In this paper, we start from analysing the difference between two kinds of super-resolution tasks in frequency domain in detail. Besides, we also analyse the degradation of different regions in the same LR image. Based on these analysis, we study how to design a generalizable and efficient RealSR framework that can exploit the merits of previous works while more targeting on the specific characteristics of RealSR itself.

Frequency Decomposition and Content Adaptation

Lots of excellent low-level restoration studies explore to enhance/reconstruct content details from the frequency decomposition perspective, including image denoising/deraining (Fu et al. 2017), rescaling (Xiao et al. 2020) and super-resolution (Fritsche, Gu, and Timofte 2019; Pang et al. 2020). Particularly, Chen *et al*. introduce an octave

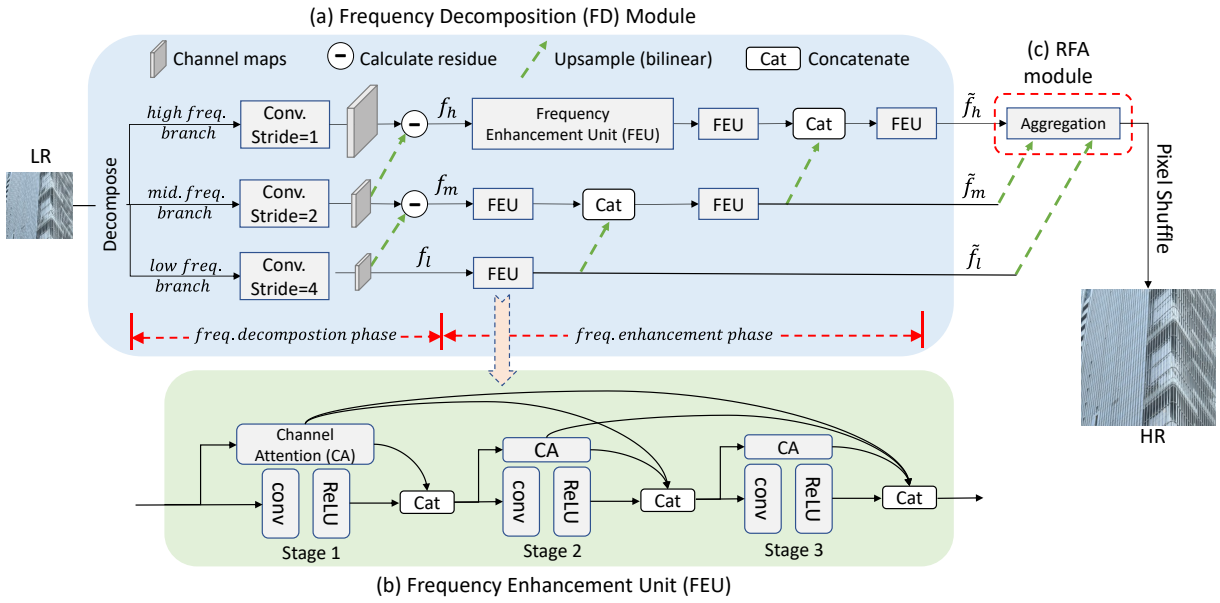

Figure 2: Architecture of our proposed OR-Net, which consists of two critical modules: (a) Frequency Decomposition (FD) module, which employs an multiple-branch architecture to decompose the input real LR image content in frequency domain and enhance texture details across all frequency components, (c) Region-adaptive Frequency Aggregation (RFA) module, which aggregates the enhanced omni-frequency components for different regions of a LR image. Moreover, in FD module, we additionally design a (b) Frequency Enhancement Unit (FEU) to strengthen the feature representation capability.

convolution (Chen et al. 2019b) to decompose features in the frequency domain to reduce spatial redundancy in CNNs. Akbari *et al.* (Akbari et al. 2020) further extend the octave convolution and frequency decomposition idea to the image compression field. These methods typically decompose the degraded image into low-/high-frequency factors, and individually deal with different components to achieve divide-and-conquer. However, they usually ignore the interaction and aggregation between multiple frequency factors. In addition, with the development of some content-adaptation techniques (*e.g.*, attention mechanism (Wang et al. 2018; Woo et al. 2018; Hou et al. 2020; Mei et al. 2020; Zhang et al. 2020), region normalization (Yu et al. 2020) and dynamic convolution (Jia et al. 2016; Mildenhall et al. 2018; Lin et al. 2019; Chen et al. 2020)), some low-level restoration algorithms (*e.g.*, image interpolation (Niklaus, Mai, and Liu 2017), hybrid-distorted image restoration (Li et al. 2020) and denoising (Mildenhall et al. 2018)) also achieve the purpose of 'divide-and-conquer', where the processing operations can be adjusted adaptively according to the changing image content in the reference. In this paper, considering that the different regions of real LR image inevitably have different frequency information lost, we explore to simultaneously utilize dynamic convolution and spatial attention to achieve adaptive frequency information compensation.

Omni-frequency Region-adaptive Network

We aim at designing a generalized and efficient framework for RealSR. During the training, we have access to the annotated real-world SR datasets with LR-HR pairs. The trained model is expected to work well with high generalization capability for the real LR images. Figure 2 shows the overall flowchart of our framework. Particularly, we propose an **O**mni-frequency **R**egion-adaptive **Net**work (OR-Net) to boost the SR performance on the real low-resolution images. OR-Net is designed from a frequency perspective, and contains two novel technical designs: Frequency Decomposition (FD) module and Region-adaptive Frequency Aggregation (RFA) module. As shown in Figure 2(a), FD module first employs an architecture with three branches, which decomposes the input real LR image into low-/middle-/high-frequency components (*i.e.*, omni-frequency) in frequency domain and enhance them in an interactive manner to achieve the comprehensive information compensation. Moreover, the RFA module (see Figure 2(c), 4) adaptively fuses the enhanced omni-frequency for different regions to achieve the content-aware super-resolution.

Frequency Decomposition (FD) Module

To obtain the informative omni-frequency representation for RealSR, we propose the **F**requency **D**ecomposition (FD) module. As shown in Figure 2(a), FD module is consists of two phases: frequency decomposition phase and frequency enhancement phase. The first frequency decomposition phase aims to separate the low-/middle-/high- frequency components from the LR input, and the frequency enhancement phase is dedicated to enhance the different frequency representations. To encourage the interaction between different frequency components and enhance frequency components in a coarse-to-fine (easy-to-hard) manner, we also

progressively utilize the enhanced lower frequency representation to help the enhancement of higher frequency components by concatenating two frequency representations.

For the FD module, we denote the input (which is an RGB LR image) by $I \in \mathbb{R}^{h\times w\times 3}$ and the output includes three kinds of enhanced frequency features: low-frequency enhanced features $\widetilde{f}_l$, mid-frequency enhanced features $\widetilde{f}_m$, and high-frequency enhanced features $\widetilde{f}_h$.

Frequency Decomposition Phase. In FD module, we first decompose the input LR image I into different frequency components. Such frequency components separation could be achieved in wavelet transform (**?**) or discrete consine transform (Ahmed, Natarajan, and Rao 1974) in the traditional signal-processing methods. However, with the mathematical operations being deterministic and task-irrelevant, such transforms inevitably discard some critical/detailed information for low-level restoration tasks. To imitate the wavelet transform while avoiding key information lost, we propose to factorizes the mixed feature representations through the *learnable* latent-wise spatial down-sampling, the similar operation can be found in the recently-proposed octave convolution (OctConv) (Akbari et al. 2020) but OctConv aims to reduce channel-wise redundancy like group or depth-wise convolutions. Specifically, we first utilize the convolution layer with the larger stride (*e.g.*, stride$=2$) to downsample the feature represent to extract the coarse features, *i.e.*, the low-frequency components. Then, we remove such relatively low frequency components from the original feature (before downsample) to obtain the rest relatively high frequency feature representations.

In formula, and as shown in Figure 2(a), we downsample the feature space by using convolution layer with stride$=4$ to get the corresponding low-frequency components f_l. Then we get the middle frequency components f_m by removing the f_l from the corresponding original features, which is also downsampled with stride$=2$. Similarly, to get the high frequency components f_h, we remove the downsampled features with stride$=2$ from the features without down-sampling, which has same spatial size with original LR image. The whole process can be denoted as follows:

$$\begin{aligned} f_l &= Conv{\downarrow_2}(Conv{\downarrow_2}(I)),\\ f_m &= Conv{\downarrow_2}(I) - Conv{\downarrow_2}(Conv{\downarrow_2}(I)){\uparrow_2},\\ f_h &= Conv(I) - Conv{\downarrow_2}(I){\uparrow_2}, \end{aligned} \tag{1}$$

where $Conv{\downarrow_2}$ denotes the convolution layer with stride$=2$ and $Conv$ denotes the convolution layer without downsampling. $\uparrow$ means the bilinear upsampling operation. The corresponding interpretive/analysis experimental results that support the reasonableness of such frequency decomposition design can be found in Figure 6.

Frequency Enhancement Phase. After extracting the low-/middle-/high-frequency components from LR image, we enhance these representations by a well-designed Frequency Enhancement Unit (FEU), to make up for low-/middle-/high-frequency information lost. Specifically, the FEU is designed based on the popular GRDB module (Kim, Ryun Chung, and Jung 2019), which can be regarded as an dense connection block. But, as shown in the Figure 2(b) and Figure 3, the features on the different stages of FEU are usually different and contain different information (some features focus on object structure but others focus on texture details). Hence, the common fusion/concatenate (directly sum up all features) that used in traditional dense connection is not consistent with our purpose (*i.e.*, divide-and-conquer). Intuitively, the simple sum-up of all features cannot promise that the low-frequency branch only focuses on the low-frequency features, same issues also exist in other two branches (middle- and high-freq. branch). To address this problem, except of the regular operations (*e.g.*, non-linear transformation achieved by Conv+ReLU, dense connection), we further integrate the channel-wise attention to adaptively adjust the residual information aggregation in FEU as shown in Figure 2(b), which helps each branch selectively fuse the corresponding frequency components at different stages, and thus the representation capability in frequency domain of each branch is significantly improved.

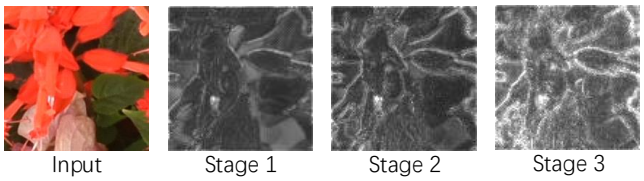

Figure 3: Feature visualization on the different stages of frequency enhancement unit (FEU).

Moreover, considering that the high-frequency feature components are relatively difficult to restore/enhance (Fritsche, Gu, and Timofte 2019; Wei et al. 2020), we propose to enhance such challenging frequency component in a coarse-to-fine/easy-to-hard manner. In detail, we encourage the interaction between different frequency components, and progressively utilize the enhanced lower frequency representations to help the enhancement of higher frequency components by concatenating them together. We define this process as:

$$\begin{aligned} \widetilde{f}_l &= Enhance(f_l),\\ \widetilde{f}_m &= Enhance(f_m, \widetilde{f}_l),\\ \widetilde{f}_h &= Enhance(f_h, \widetilde{f}_m, \widetilde{f}_l), \end{aligned} \tag{2}$$

where $Enhance(\cdot)$ denotes a set of several frequency enhancement unit (FEU), the specific number of FEU of each branch can be found in Figure 2(a).

Region-adaptive Frequency Aggregation (RFA) Module

Different from the conventional SISR, the degradation (from HR$\rightarrow$LR image) of RealSR generally exists in all frequency components. Besides, the frequency information loss of different regions in a real LR image is different. Therefore, it is necessary to adaptively aggregate omni-frequency components for different regions to restore a more realistic HR image with the rich texture details. In this section, we introduce the Region-adaptive Frequency Aggregation (RFA) module in detail (shown in Figure 4) , which achieves the

Method	Category	Scale	DRealSR			Scale	DRealSR			Scale	DRealSR		
			PSNR	SSIM	LPIPS		PSNR	SSIM	LPIPS		PSNR	SSIM	LPIPS
Bicubic	SISR	×2	32.67	0.877	0.201	×3	31.50	0.835	0.362	×4	30.56	0.820	0.438
VDSR			34.02	0.901	0.154		32.60	0.859	0.263		31.43	0.839	0.337
EDSR			34.24	0.908	0.155		32.93	0.876	0.241		32.03	0.855	0.307
RDN			34.46	0.910	0.151		33.08	0.875	0.245		32.08	0.857	0.308
DDBPN			34.26	0.906	0.157		-	-	-		31.80	0.849	0.321
RCAN			34.34	0.908	0.158		33.03	0.876	0.241		32.41	0.861	0.303
LP-KPN	RealSR	×2	33.88	-	-	×3	32.64	-	-	×4	31.58	-	-
CDC			34.45	**0.910**	0.146		33.06	0.876	0.244		32.42	0.861	0.300
OR-Net(Ours)			**34.56**	**0.910**	**0.145**		**33.28**	**0.877**	**0.230**		**32.59**	**0.863**	**0.292**

Table 1: Quantitative results on the DRealSR dataset. We compare our OR-Net to the general SISR methods, including Bicubic, VDSR, EDSR, RDN, DDBPN, RCAN, and RealSR methods, including LP-KPN and CDC. We use PSNR, SSIM and LPIPS as evaluation metrics.

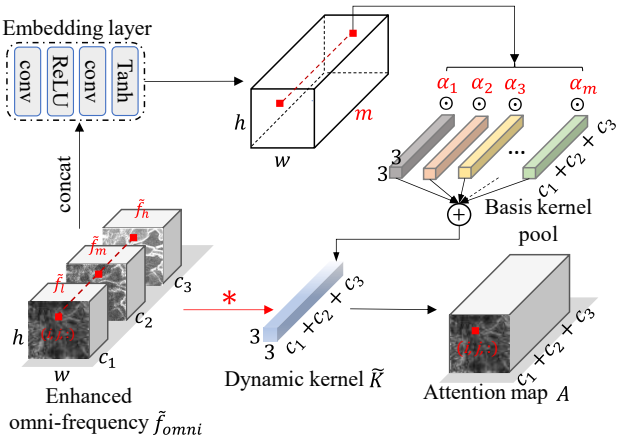

Figure 4: Region-adaptive attention map systhesis in RFA module. $*$ is matrix multiplication and $\odot$ is element multiplication.

content-aware super-resolution by combining the dynamic convolution and spatial attention mechanism.

To achieve the region-adaptive aggregation, a straightforward solution is to utilize spatial attention in (Woo et al. 2018) to fuse the low-/mid-/high-frequency components. However, the general spatial attention only leverages the spatial contextual information but lacks of considering the relationship among low-/mid-/high frequency features, since the attention map for each region is generated through the same convolution filter. Therefore, in our RFA module, to achieve better omni-frequency feature adaptive aggregation, we combine the dynamic convolution (Zhao et al. 2018) and spatial attention (Woo et al. 2018) to achieve adaptive frequency aggregation. Specifically, we utilize dynamic convolution kernels to synthesize the region-adaptive attention map from omni-frquency features. With such attention map, we achieve flexible and accurate frequency component fusion for different image regions.

As shown in Figure 4, we first concatenate the enhanced low-/mid-/high-frequency $\widetilde{f}_l$, $\widetilde{f}_m$ and $\widetilde{f}_h$ to get $\widetilde{f}_{omni}$:

$$\widetilde{f}_{omni} = [(\widetilde{f}_l\uparrow_2)\uparrow_2, (\widetilde{f}_m)\uparrow_2, \widetilde{f}_h], \tag{3}$$

where $[\cdot]$ represents the 'concat' operation. Then we set up a learnable basis kernel pool $\mathcal{K} \in \mathbb{R}^{m\times in\times c\times k\times k}$, where m, in, c, and k represent the number of filters, input channel, output channel and kernel size (see Figure 4). After that, we pass $\widetilde{f}_{omni}$ through an embedding layer to obtain a coefficient tensor $\alpha \in \mathbb{R}^{h\times w\times m}$, where h and w represent the same height and width of omni-frequency $\widetilde{f}_{omni}$. Finally, we re-weight m filters in $\mathcal{K}$ with $\alpha(i,j,:)$ for region $(i,j,:)$ to get region-adaptive dynamic convolution filter $\widetilde{K}$:

$$\widetilde{K} = \sum_{n=1}^{m} \alpha_n(i,j,:)K_n. \tag{4}$$

With dynamic convolution kernel $\widetilde{K}$, we can get the region-adaptive attention map $A(i,j,:)$ for region $(i,j,:)$ as:

$$A(i,j,:) = \widetilde{f}_{omni}(i,j,:) * \widetilde{K}, \tag{5}$$

where $*$ represent the convolution operation. Finally, the aggregated omni-frequency feature f can be obtained by:

$$f(i,j,:) = A(i,j,:) \bullet \widetilde{f}_{omni}, \tag{6}$$

where $\bullet$ represent the dot multiplication.

Experiments

In this section, we first describe the datasets of RealSR and our implementation details in Section . And then, to verify the superiority of our method, we compare the proposed OR-Net with the current state-of-the-art RealSR methods and conventional SISR methods in Section . To validate the effectiveness of the proposed FD and RFA modules, we show the visualization analysis in Section and present a series of ablation studies for OR-Net in Section .

Dataset and Implementation Details

RealSR is now under-explored and few works focus on such new challenge. Hence, these are few datasets can be used for

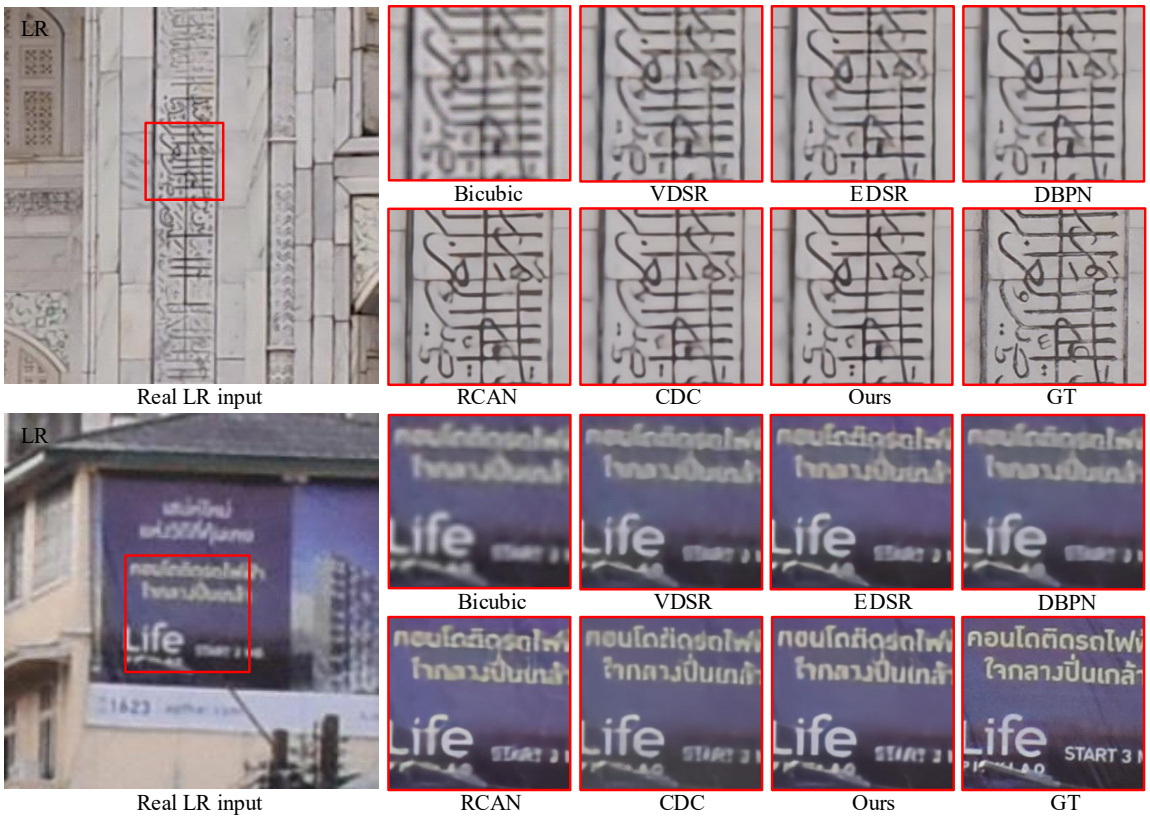

Figure 5: The qualitative comparison of our OR-Net with the state-of-the-art methods performed on DRealSR dataset ($\times 4$ scale). The VDSR, EDSR, DBPN, RCAN are designed for the general SISR and CDC is specifically designed for RealSR. Note that all methods are re-trained on the DRealSR dataset for fairness.

evaluation. Here we evaluate our OR-Net on DRealSR (Wei et al. 2020). DRealSR dataset is collected by (Wei et al. 2020). The training dataset contains 35,065, 26,118, and 30,502 image patches for scales of $\times 2$, $\times 3$ and $\times 4$, respectively. The size for patches of scale $\times 2$, $\times 3$ and $\times 4$ are 380×380, 272×272 and 192×192. The testing dataset contains 83, 84, and 93 images for $\times 2 \sim \times 4$, respectively.

The implementation of OR-Net is based on PyTorch framework. In the training process, we utilize Adam optimizer with an initial learning rate of 0.0001 and the learning rate decay by a factor of 0.5 each epoch. Batch size is 8 and we leverage random flip, random rotation and random cropping to achieve data augmentation. We randomly crop the training image as 192×192. For FD module, we set the channels of three frequency branches as 128, 128 and 64 from low-frequency to high-frequency components. For RFA module, we set the number of basis kernels K as 5.

L_1 loss has been verified effective and been widely used in many super-resolution works (Lim et al. 2017; Zhang et al. 2018a). In this paper, we also utilize the L_1 loss to optimize our OR-Net.

Comparison with State-of-the-Arts

We compare our OR-Net with the state-of-the-art traditional SISR models (including *Bicubic*, *VDSR* (Kim, Kwon Lee, and Mu Lee 2016), *EDSR* (Lim et al. 2017), *RDN* (Zhang et al. 2018b), *DDBPN* (Haris, Shakhnarovich, and Ukita 2018), *RCAN* (Zhang et al. 2018a)), and the RealSR models (including *LP-KPN* (Cai et al. 2019) and *CDC* (Wei et al. 2020)). As shown in Table. 1, our OR-Net achieves the best performance in terms of PSNR, SSIM and LPIPS compared to other general SISR and RealSR methods. We analyse that the general learning-based SISR methods that focus on solving synthetic degradation usually ignore the restoration of the full-frequency components of real LR, and cannot handle non-uniform distortions well. Besides, CDC and LP-KPN both ignore to leverage rich hierarchical information/features in frequency domains to comprehensively enhance texture details for real LR, which limits their practicality and scalability. In contrast, with the proposed frequency decomposition (FD) module and region-adaptive frequency aggregation (RFA) module, our OR-Net could enhance texture details for real LR images across all frequency scope while achieving content-aware super-resolution.

We provide the qualitative comparison of our OR-Net with state-of-the-art conventional SISR methods (including Bicubic, VDSR, EDSR, DBPN, and RCAN) and RealSR method (including CDC). We can see that those traditional SISR methods (including VDSR, EDSR, DBPN) fail to restore well some corrupted details of real LR, *e.g.*, the letters

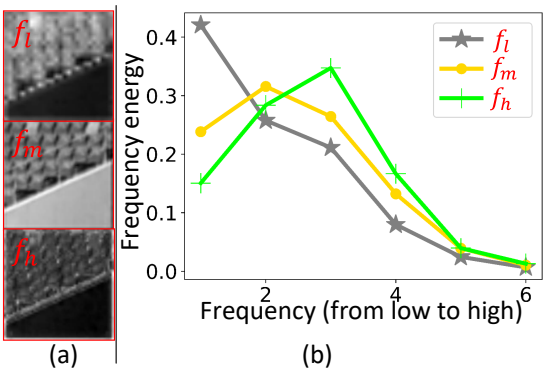

Figure 6: (a) Feature visualization and (b) frequency distribution analysis for FD module.

Scheme	bran.=1	bran.=2	bran.=3 (ours)	bran.=4
PSNR	32.25	32.40	**32.59**	32.52
SSIM	0.856	0.859	**0.877**	0.861
LPIPS	0.310	0.299	**0.230**	0.296

Table 2: Performance of different settings in OR-Net, where "bran. = i" means that there are i decomposed frequency branches in FD module of OR-Net.

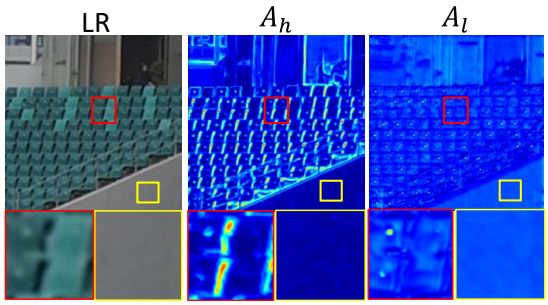

Figure 7: Visualization of the synthesized attention maps in RFA module, and $A_h, A_l \in A$.

RFA	FEU	PSNR	SSIM	LPIPS
✓	✓	**32.59**	**0.863**	**0.292**
×	✓	32.34	0.858	0.302
✓	×	32.23	0.859	0.294
×	×	32.11	0.856	0.302
SA	✓	32.46	0.860	0.295

Table 3: Ablation experiments conducted on DRealSR to study the effectiveness of the proposed RFA module and FEU design in our OR-Net.

in the second rows. RealSR method CDC cannot effectively remove blur artifacts. However, our OR-Net can achieve better restoration of textures and details for all regions.

We also evaluate our OR-Net on RealSR dataset (Cai et al. 2019) and traditional SISR datasets in **Supplementary**.

Visualization

To study the influence of each module in our OR-Net, we visualize their inner features to understand how they work. For FD module, we first visualize the features for three frequency scales in Figure 6(a), we see that the feature in high-frequency branch contains more details and texture information. We then analyse the low-/mid-/high-frequency features according to wavelet transform in Figure 6(b). From left to right, with the frequency scale increases from 1 to 6, we find that the energy of low-frequency features f_l is almost concentrated in the low frequency domain, and the energy of mid-/high-frequency features f_m/f_h are concentrated in the relatively higher frequency domains, which validate the effectiveness of our frequency decomposition module.

For RFA module, we visualize the generated attention maps that used to adaptively aggregate frequency features in Figure 7. We see that the attention map A_h for high-frequency features (Note that $A_h \in A$) tends to enhance the high-frequency details at the object edges, and the low-frequency attention map A_l ($A_l \in A$) tends to select the low-frequency features at the flat region.

Ablation Studies

Effectiveness of omni-frequency. To study the influence of omni-frequency, we set the number of decomposed frequency branches in OR-Net as 1, 2, 3, 4 to get several schemes (denoted as $bran.$ 1, 2,3, 4). As shown in Table 2, we observe that the performance of OR-Net is improved as the number of frequency branch increases, but begins to decay when the number of branch exceeds 3. To balance the complexity and performance, we employ 3 frequency branches (corresponding to the low-/mid-/high-frequency components) in this paper.

Effectiveness of our RFA module and FEU design. As shown in Table. 3, the proposed RFA module could bring 0.25dB gain in PSNR, and the FEU design also could bring 0.36dB gain in PSNR. Moreover, when we replace our RFA module with a general spatial attention (SA) to implement frequency aggregation, we see that our RFA module outperforms SA by 0.13dB in PSNR.

Conclusion

In this paper, we propose an **O**mni-frequency **R**egion-adaptive **Net**work (OR-Net) to enable effective real image super-resolution. To efficiently promise the learned feature representations of OR-Net are both *informative* and *content-aware*, we first start from the frequency perspective and design a Frequency Decomposition (FD) module to fully leverage the omni-frequency features to comprehensively enhance texture details for LR images. Then, to adaptively aggregate such omni-frequency features to achieve the content-aware super-resolution, we further introduce a Region-adaptive Frequency Aggregation (RFA) module. Extensive experiments on several benchmarks demonstrate the superiority of OR-Net, and comprehensive ablation analysis verify the effectiveness of FD and RFA modules.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China 2018AAA0101400, and NSFC under Grant U1908209, 61632001.

References

Ahmed, N.; Natarajan, T.; and Rao, K. R. 1974. Discrete cosine transform. *IEEE transactions on Computers* 100(1): 90–93.

Akbari, M.; Liang, J.; Han, J.; and Tu, C. 2020. Generalized Octave Convolutions for Learned Multi-Frequency Image Compression. *arXiv preprint arXiv:2002.10032* .

Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; and Zhang, L. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE International Conference on Computer Vision*, 3086–3095.

Chen, C.; Xiong, Z.; Tian, X.; Zha, Z.-J.; and Wu, F. 2019a. Camera lens super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1652–1660.

Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; and Liu, Z. 2020. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11030–11039.

Chen, Y.; Fan, H.; Xu, B.; Yan, Z.; Kalantidis, Y.; Rohrbach, M.; Yan, S.; and Feng, J. 2019b. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, 3435–3444.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11065–11074.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2015. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence* 38(2): 295–307.

Dong, C.; Loy, C. C.; and Tang, X. 2016. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, 391–407. Springer.

Fritsche, M.; Gu, S.; and Timofte, R. 2019. Frequency separation for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 3599–3608. IEEE.

Fu, X.; Huang, J.; Ding, X.; Liao, Y.; and Paisley, J. 2017. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Transactions on Image Processing* 26(6): 2944–2956.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2018. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1664–1673.

Hou, Q.; Zhang, L.; Cheng, M.-M.; and Feng, J. 2020. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4003–4012.

Jia, X.; De Brabandere, B.; Tuytelaars, T.; and Gool, L. V. 2016. Dynamic filter networks. In *Advances in neural information processing systems*, 667–675.

Kim, D.-W.; Ryun Chung, J.; and Jung, S.-W. 2019. Grdn: Grouped residual dense network for real image denoising and gan-based real-world noise modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Kim, J.; Kwon Lee, J.; and Mu Lee, K. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1646–1654.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Li, X.; Jin, X.; Lin, J.; Liu, S.; Wu, Y.; Yu, T.; Zhou, W.; and Chen, Z. 2020. Learning Disentangled Feature Representation for Hybrid-distorted Image Restoration. In *European Conference on Computer Vision*, 313–329. Springer.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 136–144.

Lin, X.; Ma, L.; Liu, W.; and Chang, S.-F. 2019. Context-Gated Convolution. *arXiv preprint arXiv:1910.05577* .

Mei, Y.; Fan, Y.; Zhou, Y.; Huang, L.; Huang, T. S.; and Shi, H. 2020. Image Super-Resolution With Cross-Scale Non-Local Attention and Exhaustive Self-Exemplars Mining. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mildenhall, B.; Barron, J. T.; Chen, J.; Sharlet, D.; Ng, R.; and Carroll, R. 2018. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2502–2510.

Niklaus, S.; Mai, L.; and Liu, F. 2017. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, 261–270.

Pan, J.; Liu, Y.; Sun, D.; Ren, J. S.; Cheng, M.-M.; Yang, J.; and Tang, J. 2020. Image Formation Model Guided Deep Image Super-Resolution. In *AAAI*, 11807–11814.

Pang, Y.; Li, X.; Jin, X.; Wu, Y.; Liu, J.; Liu, S.; and Chen, Z. 2020. FAN: Frequency Aggregation Network for Real Image Super-resolution. *arXiv preprint arXiv:2009.14547* .

Sajjadi, M. S.; Scholkopf, B.; and Hirsch, M. 2017. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, 4491–4500.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wei, P.; Xie, Z.; Lu, H.; Zhan, Z.; Ye, Q.; Zuo, W.; and Lin, L. 2020. Component Divide-and-Conquer for Real-World Image Super-Resolution. *arXiv preprint arXiv:2008.01928* .

Woo, S.; Park, J.; Lee, J.-Y.; and So Kweon, I. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xiao, M.; Zheng, S.; Liu, C.; Wang, Y.; He, D.; Ke, G.; Bian, J.; Lin, Z.; and Liu, T.-Y. 2020. Invertible Image Rescaling. *ECCV* .

Yu, T.; Guo, Z.; Jin, X.; Wu, S.; Chen, Z.; Li, W.; Zhang, Z.; and Liu, S. 2020. Region Normalization for Image Inpainting. In *AAAI*, 12733–12740.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018a. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 286–301.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2018b. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2472–2481.

Zhang, Z.; Lan, C.; Zeng, W.; Jin, X.; and Chen, Z. 2020. Relation-Aware Global Attention for Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3186–3195.

Zhao, F.; Zhao, J.; Yan, S.; and Feng, J. 2018. Dynamic conditional networks for few-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 19–35.