Frequency-based Enhancement Network for Efficient Super-Resolution

PARICHEHR BEHJATI1, PAU RODRIGUEZ2, CARLES FERNÁNDEZ3, ARMIN MEHRI1, F. XAVIER ROCA1, SEIICHI OZAWA4, AND JORDI GONZÀLEZ1.

1Computer Vision Center, Univ. Autònoma de Barcelona, Bellaterra, Spain (e-mail: pbehjati@cvc.uab.cat, xavier.roca@uab.cat, Jordi.Gonzalez@uab.cat)
2ServiceNow Research Montreal, Canada (e-mail: pau.rodriguez@servicenow.com)
3Oxolo GmbH, Hamburg, Germany (e-mail: c.fernandeztena@oxolo.com)
4Center for Mathematical and Data Sciences, Kobe University, Kobe 657-8501, Japan (e-mail: ozawahsukokobe-u.ac.jp)

Corresponding author: Parichehr Behjati (e-mail: pbehjati@cvc.uab.cat).

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) and the European Regional Development Fund (ERDF) under Project PID2020-120311RB-I00/AEI/10.13039/501100011033

ABSTRACT Recently, deep convolutional neural networks (CNNs) have provided outstanding performance in single image super-resolution (SISR). Despite their remarkable performance, the lack of high-frequency information in the recovered images remains a core problem. Moreover, as the networks increase in depth and width, deep CNN-based SR methods are faced with the challenge of computational complexity in practice. A promising and under-explored solution is to adapt the amount of compute based on the different frequency bands of the input. To this end, we present a novel Frequency-based Enhancement Block (FEB) which explicitly enhances the information of high frequencies while forwarding low-frequencies to the output. In particular, this block efficiently decomposes features into low- and high-frequency and assigns more computation to high-frequency ones. Thus, it can help the network generate more discriminative representations by explicitly recovering finer details. Our FEB design is simple and generic and can be used as a direct replacement of commonly used SR blocks with no need to change network architectures. We experimentally show that when replacing SR blocks with FEB we consistently improve the reconstruction error, while reducing the number of parameters in the model. Moreover, we propose a lightweight SR model — Frequency-based Enhancement Network (FENet) — based on FEB that matches the performance of larger models. Extensive experiments demonstrate that our proposal performs favorably against the state-of-the-art SR algorithms in terms of visual quality, memory footprint, and inference time. The code is available at https://github.com/pbehjatii/FENet.

INDEX TERMS Deep learning, frequency-based methods, lightweight architectures, Single image super-resolution

I. INTRODUCTION

SINGLE image super-resolution (SISR) has recently received a considerable amount of attention from both academia and industry. The purpose of SISR is to reconstruct a high-resolution (HR) image from its low-resolution observation (LR). This offers an opportunity for overcoming resolution limitations in various computer vision applications such as medical imaging [49], security and surveillance [47]. In general, SISR is an inverse ill-posed problem since multiple HR images can map to the same LR input. To tackle such an inverse problem, numerous image SR methods have been proposed [16] based on deep neural architectures [9, 34, 36, 61] and shown prominent performance.

Convolutional Neural Networks (CNNs) have recently achieved unprecedented success in various problems [18, 57]. The powerful feature representation and end-to-end training paradigm of CNNs make them a promising approach to SISR. Recently, most CNN-based SR methods focus on elaborate architecture designs such as residual learning [2, 5, 31, 32] and dense connections [25, 67]. Although significant progress has been made, as discussed in [21, 46], texture details of the LR images often tend to be smoothed in the super-resolved results since most existing CNN-based SR methods do not pay enough attention to the limited high-frequency information in the LR images. In natural images, information is conveyed at different frequencies. The output feature maps...
of a convolutional layer can also be seen as a mixture of information at lower and higher frequencies. The lower frequency information is composed of global structures and textures that can directly be forwarded to the final HR output without substantial computations. The higher frequency information consists of fine details where more complex restoring functions are expected. At this point, leading CNN-based methods such as EDSR [37] and RDN [67] overlook the fact that most of the low-frequency information is already contained in the input. As a result, these models spend the same amount of computation treating low- and high-frequency information and lack flexible modulation ability in dealing with them, which ends up the representational ability of the network. Please note that, in this paper, the term frequency refers to low- and high-frequency features, and is not related to the frequency domain.

Previous works address this problem by incorporating attention mechanisms [9, 61, 66] into the networks to model interdependencies among spatial locations, channels, or both. The common idea behind attention-based SR methods is to adjust network architectures so that they produce rich feature representations. However, as SR networks are so diverse, the attention module is usually designed solely for a specific network structure [55]. Recently, various SR methods such as multi-branch networks [33, 60] and progressive reconstruction methods [35, 69] mainly focus on refining the high-frequency texture details. Although these methods delivered impressive results, they demand substantial memory and computational resources. Therefore, the efficient reconstruction of high-frequency details in SISR is still a challenge today.

In this paper, we address the aforementioned problems from a different perspective. Instead of designing deep and complex networks or adding various shortcut connections to strengthen feature representations, we introduce a novel Frequency-based Enhancement Block (FEB) which is able to separate features into low and high frequencies while also enabling efficient communication among them. Since low frequencies are preserved by downsampling operations and thus can be recovered directly from the input, FEB assigns more computational capacity to high frequencies. The proposed FEB gradually and iteratively enhances high-frequency feature maps during training while preserving low-frequency information, resulting in more accurate features that improve reconstruction quality.

The proposed FEB offers the following advantages. First, it is generic and can be easily applied to existing SR models without the need of modifying network architectures or requiring hyper-parameters tuning. Second, FEB reduces model parameters in the baseline SR models while simultaneously obtaining better SR performance. In Figure 1, we provide an example of visual quality of EDSR [37], which uses residual blocks [18] as its building module. It can be observed that, when we replace residual blocks with our blocks (EDSR-FEB), the network obtains better visual quality while reducing the number of parameters.

Based on FEB, we build a lightweight SR network named Frequency-based Enhancement Network (FENet), illustrated in Fig 2. Our network leads to significant improvements for single image SR, surpassing SR networks with complicated skip connections and concatenations. In summary, these are the main contributions of the paper:

- We propose a novel Frequency-based Enhancement Block (FEB) to perform frequency-based computation. Such a mechanism allocates more computation to high-frequency bands, allowing the network to focus on more informative features and improve its discriminative capabilities.
- The proposed block leads to the reduction of parameters by half in the baseline SR models while achieving better SR performance.
- We propose a lightweight Frequency-based Enhancement Network (FENet) for fast and accurate image super-resolution. Extensive experiments on a variety of public datasets demonstrate the superiority of the proposed architecture over state-of-the-art models, in terms of both quantitative and visual quality.

II. RELATED WORK

In recent years, the field of image SR has been dominated by CNNs, which achieve state-of-the-art performance [5, 36, 42, 43, 54, 61]. Here, we focus our discussion on the approaches that are most related to our work.

A. EVOLUTION OF ARCHITECTURES FOR SR

Recently, CNN-based methods have dramatically boosted the performance of image SR, due to their strong nonlinear representational power. They learn mappings between LR and HR images from large-scale paired datasets. Since the advent of SRCNN [13], a three layer CNN, a great number of CNN-based methods have been proposed to improve model representation ability by using more elaborate neural network architecture designs. Kim et al. [26] first pushed the depth of SR networks to 20 with the help of residual learning, outperforming SRCNN by a large margin. Ledig
et al. [31] employed residual blocks proposed in [18] to construct deeper network (SRResNet) for image SR, which was further improved by EDSR [37] and MDSR [37] by removing unnecessary modules (e.g., batch normalization) from the residual blocks. By using effective building modules, image SR networks became deeper and yielded better performance. Later, in order to employ hierarchical features from all the convolutional layers in deep networks, dense blocks started being employed in SR architectures [3, 22, 25, 53, 63]. More recently, Zhang et al. [67] and Liu et al. [39] also used dense and residual connections in RDN and RFANet to utilize information from the whole feature hierarchy. In addition to residual and dense blocks, Li et al. [32] and Lan et al. [30] proposed a multi-scale block to explore the multi-scale information of LR images. Although these existing CNN-based SR approaches have provided outstanding performance, they devoted to designing deeper and wider network to enhance their representational learning capacity. Increase in depth and width has also raised computational demands and memory consumption. This makes modern architectures less applicable in practice.

Numerous lightweight models have been proposed to alleviate the aforementioned computational burden. For example, DRCN [27] was the first to apply recursive algorithm to SISR to reduce the number of parameters by reusing them multiple times. Tai et al. [51] and Ahn et al. [2] improved DRCN by combining the recursive and residual network schemes in order to achieve better performance with even fewer parameters. Likewise, Behjati et al. [5] and Jiang et al. [25] also joined residual connections and recursive layers to reduce the computational cost. On the other hand, LapSRN [28] employed a pyramidal framework to increase the image size gradually. By doing so, LapSRN effectively performed SISR on extremely low-resolution cases. Chu et al. [11] and Ahn and Cho [1] introduced neural architecture search strategies to automatically build an SR model given certain constraints. Meanwhile, Hui et al. [24] proposed an information multi-distillation block (IMDB) that extracted features at a granular level with the channel splitting strategy. More recently, Luo et al. [40] proposed lattice blocks that applied so-called butterfly structures to combine residual layers. Later, Xuehui Wang and Chen. [58] proposed an attentive feature block to utilize auxiliary features of previous layers for facilitating features learning of the current layer. Li et al. [34] proposed a linearly-assembled pixel-adaptive regression network, which casts the direct LR to HR mapping learning into a linear coefficient regression task. Recently, to simplify the challenges of directly super-resolving details, some authors adopted the progressive structure to reconstruct HR images in a stage-by-stage upscaling manner [36, 38, 69].

By considering that there are different types of information within and across feature maps which have a different contribution for image SR, the aforementioned SR approaches cannot capture low- and high-frequency feature representations separately in the process of feature embedding, thus hindering their representational ability [21, 46, 66].

B. FREQUENCY BASED SR METHODS

It is well known that high-frequency information (e.g., texture, edges) is significant for SISR. Li et al. [35] proposed a super-resolution feedback network (SRFBN) based on a recurrent architecture design. The network is based on a feedback block that consists of several projection groups. Each projection group first finds high-resolution features (via deconvolution) and then generates low-resolution features (via convolution). As a result, this network is able to gradually recover high-frequency components. Later, Haris et al. [17] proposed a method to refine high-frequency texture details with a series of up and downsampling layers that are densely connected with each other to combine HR images from multiple depths in the network. More recently, Qiu et al. [46] and Yang and Lu [60] proposed multi-branch architectures. In these methods, one branch is responsible for capturing high-frequency features such as texture and edge, and another is to learn low-frequency features such as image outline and contour. Similarly, Li et al. [33] introduced the octave convolution to image SR which uses two branches to perform information update and frequency communication between low- and high-frequency features.

Although these existing SR approaches have made good efforts to improve SR performance, they tend to increase the amount of compute on high-frequency information by increasing the overall number of operations of the model, without paying attention to model complexity. The increase in complexity due to the independent treatment of multiple frequencies is a key issue that limits the performance of these deep CNN-based methods.

C. ATTENTION BASED SR METHODS

Attention mechanism has demonstrated great superiority in improving performance of CNNs for various computer vision tasks [20, 57]. Hu et al. [20] introduced squeeze-and-excitation (SE) block that models channel-wise relationships in a computationally efficient manner and enhances the representational ability of the network, showing its effectiveness on image classification. CBAM [57] modified the SE block to exploit both spatial and channel-wise attention. Zhang et al. [66] first incorporated SE [20] with SR and pushed the state-of-the-art performance of SISR. More recent works, such as [9, 21, 29, 43, 44, 58, 59, 61], extend this idea by adopting different spatial attention mechanisms or designing advanced attention blocks.

All above-mentioned approaches improve CNNs for image SR by either refining architectural designs or adding complexity to hand-designed blocks. Conversely, our proposal is able to efficiently restore textures at different frequencies. Such mechanism helps the network to explicitly allocate computation to high-frequency features, thus improving the discriminative capabilities of the network.

III. FREQUENCY-BASED ENHANCEMENT NETWORK

In this section, We first describe the overall network architecture. Next, we detail the proposed Frequency-based En-
enhancement Block (FEB). Finally, we discuss the differences between the proposed method and similar related works.

A. NETWORK OVERVIEW

As shown in Fig 2, the overall network architecture of Frequency-based Enhancement Network (FENet) consists of a non-linear mapping module and a reconstruction module. Let’s denote as $I_{LR}$ and $I_{SR}$ the input and output of FENet, respectively. Following [2, 5, 40, 58, 68], we apply only one $3 \times 3$ convolutional layer ($H$) to extract the initial features $H_0$ from the LR input image:

$$H_0 = \mathcal{H}(I_{LR}). \quad (1)$$

It is worth noting that only one convolutional layer is used here for lightweight design.

Then, we use the non-linear mapping module, which consists of several stacked FEBs to generate new powerful representations, which can be formulated as

$$H_k = B_k(H_{k-1}), \quad k = 1, \ldots, M, \quad (2)$$

where $B_k$ denotes mapping function of the $k$-th FEB. $H_{k-1}$ represents the features from the previous adjacent FEB, and $M$ is the total number of FEBs.

Inspired by [24, 32, 58, 67], we apply a feature fusion strategy to integrate the features from all the FEBs. This strategy helps to extract more hierarchical contextual information. The fusion operation is formulated as

$$H = F([H_1, H_2, \ldots, H_M]) \quad (3)$$

where $[H_1, H_2, \ldots, H_M]$ refers to the concatenation of feature maps produced by FEBs and $F$ is a $1 \times 1$ convolutional operation.

Finally, we utilize the reconstruction module that contains convolutional layers and pixelshuffle layers [50] to upsample the features to the HR size. In addition, we incorporate a global connection path (green line in the Fig 2) to grant access to the original LR information and facilitate the back-propagation of the gradients, in which only a bicubic interpolation is applied to the input $I_{LR}$. Therefore, we obtain:

$$I_{SR} = \mathcal{R}(H) + \text{Bicubic}^\dagger(I_{LR}) \quad (4)$$

where $\mathcal{R}$ is the reconstruction module, and $I_{SR}$ is the final output of the network.

To optimize the network parameters, we adopt $L_1$ loss as a cost function for training. Given a training set with $N$ pairs of LR images and HR counterparts, denoted by $\{I_{LR}^i, I_{HR}^i\}_{i=1}^N$, the network is optimized to minimize the $L_1$ loss function:

$$L_1(\theta) = \frac{1}{N} \sum_{i=1}^{N} \|I_{SR} - I_{HR}\|_1, \quad (5)$$

where $\theta$ denotes the parameter set.

B. Frequency-based Enhancement Block (FEB)

A natural image can be decomposed into a low frequency component that describes smoothly changing structures and a high-frequency component that describes the rapidly changing fine details [8, 48]. Similarly, we argue that the output feature maps of a convolutional layer can also be decomposed into features of different frequencies, and propose an efficient Frequency-based Enhancement Block (FEB) which naturally decomposes low and high frequencies at feature level. The high-frequency information part is processed by higher-complexity operations (in number of parameters and non-linearities), whereas the lower-frequency part is processed by lower-complexity operations to compensate for the increase of computation. As a result, the proposed approach learns discriminative representations in order to efficiently achieve more accurate reconstructions.

As demonstrated in Fig 3, the proposed FEB contains two pathways, each of which is responsible for a different functionality. Each pathway has a $1 \times 1$ convolutional layer at the beginning. Given the input $X \in \mathbb{R}^{C \times H \times W}$, where $C$ denotes the number of channels and $H \times W$ the spatial dimensions, we have

$$X_1 = F'_{\text{split}}(X) \quad (6)$$
$$X_2 = F''_{\text{split}}(X) \quad (7)$$

where $\{X_1, X_2\}$ only have half of the channel number of $X$. $F'_{\text{split}}$ and $F''_{\text{split}}$ are two $1 \times 1$ convolutional operations, respectively.
Then, the described operations are separately sent into a dedicated pathway for collecting different types of information (i.e., low- and high-frequency information). The first pathway targets at retaining the original information (low-frequency). To save computation, we perform only a simple $3 \times 3$ convolutional operation to capture the global layout and coarse details as follows:

$$Y_1 = F_1(X_1), \quad \text{(8)}$$

where $Y_1$ is the output of the $3 \times 3$ convolutional layer ($F_1$).

In the second pathway, we first apply an average pooling layer upon $X_2$, yielding $T_1$

$$T_1 = \text{AvgPool}^k(X_2, k), \quad \text{(9)}$$

where $k$ denotes the kernel size of the pooling layer and the size of the intermediate feature map $T_1$ is $C \times \frac{H}{k} \times \frac{W}{k}$. Each value in $T_1$ can be viewed as the average intensity of each specified small area of $X_2$. After that, $T_1$ is upsampled via a bicubic interpolation operator to produce a new tensor $T_2$ of the same size as $X_2$

$$T_2 = \text{Bicubic}^{+}(T_1, k), \quad \text{(10)}$$

where $T_2$ contains averaged information and it can be regarded as a smoother version of the original $X_2$. Then, in order to obtain the high-frequency information, $T_2$ is element-wise subtracted from $X_2$

$$T_3 = X_2 - T_2, \quad \text{(11)}$$

The visual activation maps of $X_2$, $T_2$ and high-frequency information ($T_3$) are also shown in Fig. 4. It can be observed that $T_2$ is smoother than $X_2$ as it is the average information of $X_2$. Meanwhile, $T_3$ retains the details and edges. Now, the high-frequency enhancement operation can be formulated as follows

$$Y'_2 = \sigma(F_2(T_3) + X_2) \cdot F_3(X_2), \quad \text{(12)}$$

where $\sigma$ is the sigmoid function, and $F_2$ and $F_3$ are two $3 \times 3$ convolutional layers, respectively. As shown in (12), we use $X_2$ as residuals to form the weights, which is found beneficial. Then the output of the second pathway can be written as

$$Y_2 = F_4(Y'_2), \quad \text{(13)}$$

where $F_4$ is a $3 \times 3$ convolutional operation. Finally, both intermediate outputs of the first and second pathways $\{Y_1, Y_2\}$ are concatenated together as the output $Y \in \mathbb{R}^{C \times H \times W}$ to obtain a rich feature representation.

Compared to other works such as [33, 60], which require a considerably large amount of computations for decomposing features of different frequencies, FEB can separate the low- and high-frequency feature representations in an efficient way and focus on reconstructing the high-frequency ones.

C. DISCUSSION

1) Difference to Prominent SR Blocks

Prominent SR blocks such as residual blocks [37] or dense blocks [53] process low- and high-frequency information simultaneously by the same convolution operations and do not discriminate the computation of features by their frequency components. Therefore, some local details of LR images cannot be effectively utilized for HR reconstruction, leading to blurry super-resolved results [33]. In contrast, our proposal treats different frequencies in a heterogeneous way and also models inter-channel dependencies, which consequently enrich the output feature. Moreover, FEB benefits...
SR approaches by reducing the number of parameters while achieving superior SR performance.

2) Difference to Attention-Based Methods

Our work is quite different from existing methods such as [12, 21, 43, 66] which rely on supplementary attention blocks and require additional learnable parameters. In contrast, our approach internally changes the way of exploiting convolutional filters of convolutional layers, and hence require no additional learnable parameters. In the following experiment section, we will demonstrate without any extra learnable parameters, FEB can yield significant improvements over baselines and other attention-based SR approaches. Moreover, it is complementary to attention mechanisms, and also benefit from their inclusion into the pipeline.

IV. EXPERIMENTAL RESULTS

In this section, we first conduct an ablation study to validate the effectiveness of the proposed FEB. Then, we systematically compare FENet with state-of-the-art SISR algorithms on five commonly used benchmark datasets.

A. SETTINGS

Datasets and Metrics. Following [67], we use 800 high-quality images from the DIV2K dataset [52] for training. We evaluate our model on several benchmark datasets: Set5 [6], Set14 [62], B100 [4], Urban100 [23], and Manga109 [41], each with diverse characteristics. To keep the consistency with previous works [11, 25, 29, 33, 34, 58, 61], we use Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) [56] as the assessment methods to evaluate image reconstruction accuracy. PSNR evaluates the image by statistically measuring distortion values between the reconstructed image and the ground-truth image. The higher the PSNR, the better the quality of the reconstructed image. SSIM measures the structural similarity between two images based on luminance, contrast, and structure. The SSIM values range between 0 to 1, 1 means perfect matching the reconstructed image with the original one. All results are evaluated on the luminance channel (Y). In addition to PSNR and SSIM, we adopt Perceptual Index (PI) [7] to evaluate reconstructed image perceptual quality accurately. PI has a high correlation with human-opinion scores and can avoid the situation where over-smoothed images may present a higher PSNR and SSIM when the performances of the two methods are similar. The lower PI value denotes the better perceptual quality.

Degradation models. To fairly compare against existing works, we adopt bicubic down sampling (denoted as BI) as our standard degradation model for generating LR images from ground truth HR images, at ×2, ×3, and ×4 scales. Moreover, to comprehensively illustrate the efficacy of the proposed FEB, we further adopt two other multi-degradation models as in [67]. We define BD as a degradation model that performs bicubic down sampling on HR images at ×3 scale, and then blurs them with a Gaussian kernel of size 7×7 and standard deviation 1.6. Additionally, we further produce LR images in a more challenging way: we first bicubic downsample HR images with scaling factor ×3 and then add Gaussian noise with noise level 30 (denoted as DN).

Implementation details. During training, data augmentation is carried out by means of random horizontal flips and 90° rotation. At each training mini-batch, 64 LR RGB patches of size 64 × 64 are provided as inputs. We train FENet using an ADAM optimizer with learning rate $10^{-3}$. The learning rate is halved every $2 \times 10^5$ iterations. We set the number of FEB to 12 in our FENet. Our network has been implemented using PyTorch, and trained on a NVIDIA RTX 3090 GPU.

| Configurations | 1 | 2 | 3 |
|----------------|---|---|---|
| Low-frequency path | ✓ | × | ✓ |
| High-frequency path | × | ✓ | ✓ |
| Params | 706K | 771K | 675K |
| Set5 | 32.02 | 32.06 | 32.24 |
| Set14 | 28.46 | 28.43 | 28.61 |
| B100 | 27.42 | 27.44 | 27.63 |
| Urban100 | 25.84 | 25.86 | 26.20 |
| Manga109 | 30.12 | 30.20 | 30.46 |

B. ABLATION STUDY

1) The importance of FEB

In this section, we conduct ablation experiments to explore the influence of each pathway (low- and high-frequency paths) inside the proposed FEB on the reconstruction performance. Therefore, we use FENet as the basic network and run the following experiments: (1) deactivating low-frequency path ($Y_1$) in FEB; (2) deactivating high-frequency path ($Y_2$) in FEB, and (3) activating both low- and high-frequency paths. To keep the number of parameters similar, we use 8 and 6 FEBs in the first two experiments respectively without channel reduction.

As reported in Table 1, we observe a significant performance drop when either low- or high-frequency path is deactivated in FEB. This is mainly because: 1) when low-frequency path ($Y_1$) is deactivated in FEB the high-frequency path ($Y_2$) focuses too strongly on high-frequency details, smoothing other important aspects of the input that should be preserved by the low-frequency path; 2) the network without high-frequency path ($Y_2$) processes low- and high-frequency information simultaneously by the same convolution operations, and do not explicitly extrude the high frequencies from image features. Thus, some local details of the LR image cannot be effectively utilized for HR reconstruction.

In Fig 5, we additionally visualize the average feature maps of low-frequency ($Y_1$ in (8)) and high-frequency ($Y_2$ in (13)) paths within the first four FEBs. It can be observed that low-frequency feature maps describe the overall outline.
of the butterfly, while high-frequency ones represent the edges and textures of the butterfly. This visualization shows how FEB is able to efficiently restore textures at different frequencies and can potentially improve performance.

2) The effectiveness of FEB
To demonstrate the effectiveness of our proposed FEB scheme, we use FENet as the basic network. To keep the number of parameters similar, we replace the 12 FEBs with 8 residual blocks (RB) [37], 5 dense blocks (DB) [53], 6 information multi-distillation blocks (IMDB) [24], or 4 multi-scale residual blocks (MSRB) [32]. In Table 2, we compare the number of parameters and the performance in PSNR for all methods for scale factor $\times 4$.

As reported in Table 2, the method with FEB outperforms all the methods with different SR blocks with fewer number of parameters. The reason is the proposed block treats different frequencies in a heterogeneous way and thus it improves the performance of super-resolution. This experiments justify that the proposed FEB results more helpful for image SR. We additionally provide visual comparisons (Fig 6) of FENet using different SR blocks for scale factor $\times 4$. It can be observed that the network using FEB obtains better visual quality and represents more diverse structure patterns.

3) Attention mechanisms vs FEB
To further verify the effectiveness of FEB, we use a ResNet architecture, i.e., a regular architecture composed of 8 stacked residual blocks. Then, we integrate two commonly used attention mechanism namely channel attention (CA) [66] (ResNet-CA) and channel-wise and spatial attention residual [21] (ResNet-CSAR) into residual blocks as done in [66], respectively. Furthermore, we replace 8 residual blocks with 12 FEBs (FENet) and integrated the two mentioned attention mechanisms into FEBs and named them as FENet-CA and FENet-CSAR.

As reported in Table 3, ResNet-CSAR and ResNet-CA obtain better performance than ResNet but they require additional learnable parameters. Quite differently, FENet does not rely on any extra learnable parameters since it heterogeneously exploits the convolutional filters and thus achieves better performance than ResNet-CSAR and ResNet-CA. It should also be mentioned that the proposed FEB is also compatible with the above mentioned attention mechanisms. For example, when adding CA blocks to each FEB of FENet (FENet-CA), we can further gain another 0.07dB in average. This also indicates that our approach is orthogonal to this kind of supplementary attention modules.

To dig deeper into difference between the proposed block and attention-based approaches, we visualize the average feature map of the output $Y_2$ in (12), the outputs of CA and CSAR attentions in the residual blocks in Fig 7. Our network should focus on high-frequency components (i.e. frequencies and can potentially improve performance.

![Average feature maps of low-frequency ($Y_1$ in (8)) and high-frequency ($Y_2$ in (13)) paths.](image1)

![Average feature maps of the output $Y_2$ in (12), the outputs of CA and CSAR in residual blocks, respectively.](image2)

![Visual comparisons of SR results using FENet with different SR blocks (scale factor $\times 4$).](image3)

![Average PSNR obtained with FENet when using different attention mechanisms on five benchmark datasets (scale factor $\times 4$).](image4)
edges and contours) and suppress the smooth area of the original input image. Compared with the CA and CSAR, feature maps acquired from $Y_2$ in (12) contain more negative values, showing a stronger effect of suppressing the smooth area of the input image as well as directing computations towards edges and details. This visualization indicates that the network with FEB can generate richer and more discriminative feature representations than the different attention mechanisms.

4) Generalization ability
To demonstrate the generalization ability of the proposed structure, we select two state-of-the-art SR networks with different model sizes, called EDSR [37] and RCAN [66]. The EDSR contains 32 stacked residual blocks with $256 \times 256$ filters. The RCAN consists of 200 residual channel attention blocks with $64 \times 64$ filter sizes. We replace their building blocks with FEBs. The corresponding networks with FEB are named as $EDSR$-FEB and $RCAN$-FEB, respectively. For fair comparison, all networks are trained on their default settings.

As shown in Table 4, EDSR-FEB has an improvement of 0.08dB in average with almost $\times 2$ fewer number of parameters (parameters: 28M) compared to the original EDSR (parameters: 43M). Moreover, the improvement by RCAN-FEB is also higher than RCAN with approximately half amount of parameters. From these comparisons, we can easily find that (1) the proposed FEB perform much better than channel attention, (2) for deeper networks, a similar phenomenon can also be observed, (3) FEB reduces the number of parameters by half while achieving better performance.

Fig 1 and 8 additionally show visual comparisons for scale factor $\times 4$. It can be observed that EDSR-FEB and RCAN-FEB can reconstruct sharper and more natural-looking images. This is mainly because FEB can extract high-frequency features and use them for reconstruction.

TABLE 4: Average PSNR obtained with state-of-the-art SR methods when using FEB on five benchmark datasets (scale factor $\times 4$).

| Name      | EDSR | EDSR-FEB | RCAN | RCAN-FEB |
|-----------|------|----------|------|----------|
| Set5      | 32.50| 32.58 (+0.08dB) | 32.63| 32.70 (+0.07dB) |
| Set14     | 28.72| 28.80 (+0.08dB) | 28.87| 28.96 (+0.06dB) |
| B100      | 27.72| 27.81 (+0.09dB) | 27.77| 27.85 (+0.08dB) |
| Urban100  | 26.67| 26.76 (+0.09dB) | 26.82| 26.89 (+0.07dB) |
| Manga109  | 31.02| 31.09 (+0.07dB) | 31.22| 31.30 (+0.08dB) |

5) Comparing pooling methods
In this section, we investigate the influence of different pooling types on the performance. The proposed block adopts average pooling for downsampling and bicubic interpolation for upsampling. In our experiments, we use FENet as the basic network and then replace average pooling operators in all FEBs with maximum pooling operators. As shown in Table 5, using the average pooling operator while keeping the rest of the configurations unchanged yields a performance increase of about 0.08dB in average. We argue that this may be due to the fact that, unlike maximum pooling, average pooling builds connections among locations within the whole pooling window, which can better capture local contextual information.

In addition, we further investigate the behavior of the proposed block (FEB), when the average pooling for downsampling in (9) and bicubic interpolation used for upsampling in (10) are replaced with a convolutional layer and a deconvolutional layer, respectively. As reported in Table 5, it can be observed that the performance as well as the number of parameters of the network increase, when we replace average pooling and bicubic interpolation with learnable operations. Although the performance of the network increases by 0.11dB on average, this leads to a more complex network with more parameters. While weighing the network performance and network complexity, we finally use average pooling and bicubic interpolation for the rest of the experiments, the results are close to the network with conv-deconv operations, but the number of model parameters is only one fourth of it.
6) The effect of Downsampling rate

We also investigate how the downsampling rate in FEB influences the image SR performance. In Table 6, we show the performance with different downsampling rates used in FEB. It can be observed that as the downsampling rate increases, slightly better performance is achieved. However, we do not use larger downsampling rates due to two reasons: (1) the resolution of the input features is already very small; (2) higher downsampling rates lead to performance improvements at the expense of more computations due to bicubic operation. Therefore, for the rest of experiments, we set the downsampling rate to 4 for all scale factors, as it still provides significant improvements with a lower computational cost than \( \times 5 \).

7) The effect of increasing the number of FEB

As discussed in [37], increasing the depth of the network can effectively improve the performance. In this work, adding the number of FEBs is the simplest way to gain excellent result. For better balancing the model size and performance, we compare the proposed model with the different numbers of FEBs, i.e., 6, 8, 10, and 12. As shown in Table 7, our FENet performance improves rapidly with the growth in number of FEBs. Although the performance of the network would further improve by using more FEBs, we found it leads to diminishing returns with respect to the number of parameters. Therefore, we use 12 FEBs in our experiments. Furthermore, we find by adding a global connection path (green line in Fig 2) to grant the output access to the original LR input is beneficial for reconstruction performance. Discarding this connection decreases performance (-0.04dB on average).

C. COMPARISON WITH STATE-OF-THE-ART METHODS

1) Results with BI degradation model

In this section, we compare the proposed FENet with state-of-the-art lightweight models: VDSR [26], DRCN [27], SR-DenseNet [53], SEINet [10], SRResNet [31], CARN [2], IMDN [24], SRFBN-S [35], A2F-S [58], CBPN [69], LAPAR-A [34], MADNet [29], FALSR-A [11], DPN [36], HDRN [25], and OISR-RK2 [19]. We have also listed the performance of state-of-the-art large SR methods including EDSR [37], FSN [33], and CASGCN [61] for reference.

Table 8 shows quantitative results when evaluating PSNR and SSIM on five benchmark dataset with different algorithms for scale factors \( \times 2, \times 3, \) and \( \times 4 \). For a more informative comparison, the number of parameters is also given. From Table 8, we find that FENet only has less than 0.7M parameters but performs favorably against other compared approaches on most datasets. For example, in comparison with SRDenseNet [53] and OISR-RK2 [19], FENet achieves better or competitive results, while only needing 30% and 40% of their parameters, respectively. On the other hand, thanks to the FEB, FENet achieves competitive or better results when compared to the large SR methods. Specifically, FENet outperforms FSN [33] by a large margin at all scales in all datasets with 18× fewer parameters.

In Fig 9, we present some qualitative visual comparisons for the \( \times 4 \) scale factor. It can be observed that SR images reconstructed by FENet have more refined details, especially in the edges and lines. This further validates the effectiveness of the proposed FEB.

2) Results with BD and DN degradation models

Following [67], we also show the SR results with BD degradation model and further introduce DN degradation model. The proposed FENet is compared with state-of-the-art methods including SPMSR [45], SRCNN [14], FSR-CNN [15], VDSR [26], IRCNN_G [64], IRCNN_C [64], and
TABLE 8: Average PSNR/SSIM values for models with the same order of magnitude of parameters. Performance is shown for scale factors \( \times 2, \times 3 \) and \( \times 4 \) with bold degradation model. The best and second best results are highlighted in red and blue respectively.

| Scale | Method                  | Params | PSNR  | SSIM  | Scale5 | Set14 | B100  | Urban100 | Manga109 |
|-------|-------------------------|--------|-------|-------|--------|-------|-------|----------|----------|
|       |                         |        | PSNR  | SSIM  |        |       |       |          |          |
|       |                         |        |       |       |        |       |       |          |          |
| \( \times 4 \) | CBPN [69]              | 1M     | 37.90 | 0.9591 | 33.60 | 0.9171 | 32.17 | 0.8898 | 31.94 | 0.9256 | 38.36 | 0.9765 | –      | –      |
| \( \times 2 \) | MADNet [29]             | 0.9M   | 37.94 | 0.9604 | 31.46 | 0.9167 | 32.10 | 0.8988 | 31.74 | 0.9246 | –      | –      |
| \( \times 3 \) | DPN [36]                | 0.8M   | 37.52 | 0.9558 | 33.55 | 0.9166 | 32.12 | 0.8987 | 31.93 | 0.9256 | –      | –      |
|         |                        |        |       |       |        |       |       |          |          |
| \( \times 2 \) | MADNet [29]             | 0.9M   | 37.75 | 0.9590 | 33.49 | 0.9150 | 32.03 | 0.8890 | 31.87 | 0.9250 | 38.07 | 0.9770 | –      | –      |
| \( \times 3 \) | HDRN [25]               | 0.9M   | 38.01 | 0.9605 | 33.62 | 0.9183 | 32.19 | 0.8999 | 32.10 | 0.9283 | 38.67 | 0.9772 | –      | –      |
| \( \times 4 \) | HDRN [25]               | 1.0M   | 38.00 | 0.9605 | 33.63 | 0.9177 | 32.19 | 0.8996 | 32.17 | 0.9283 | 38.88 | 0.9774 | 32.20 | 0.9290 |
| \( \times 3 \) | DPN [36]                | 0.8M   | 38.00 | 0.9605 | 33.63 | 0.9177 | 32.19 | 0.8996 | 32.17 | 0.9283 | 38.88 | 0.9774 | 32.20 | 0.9290 |
| \( \times 2 \) | MADNet [29]             | 0.9M   | 38.01 | 0.9605 | 33.62 | 0.9183 | 32.19 | 0.8999 | 32.10 | 0.9283 | 38.67 | 0.9772 | –      | –      |
| \( \times 3 \) | HDRN [25]               | 0.9M   | 38.00 | 0.9605 | 33.63 | 0.9177 | 32.19 | 0.8996 | 32.17 | 0.9283 | 38.88 | 0.9774 | –      | –      |
| \( \times 4 \) | HDRN [25]               | 1.0M   | 38.00 | 0.9605 | 33.63 | 0.9177 | 32.19 | 0.8996 | 32.17 | 0.9283 | 38.88 | 0.9774 | 32.20 | 0.9290 |

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
TABLE 9: Quantitative results with BD and DN degradation models. The best and second best results are highlighted in red and blue respectively.

| Methods       | Degrad | Set5  | Set14 | B100  | Urban100 | Manga109 |
|---------------|--------|-------|-------|-------|----------|----------|
|               | PSNR   | SSIM  | PSNR  | SSIM  | PSNR     | SSIM     | PSNR     | SSIM     |
| SPMSR [45]    | BD     | 32.21 | 0.9001| 28.89 | 0.8105   | 28.13    | 0.7740   | 25.84    | 0.7856   | 29.64    | 0.9003   |
|               | DN     | 25.01 | 0.6950| 23.78 | 0.5898   | 23.76    | 0.5558   | 21.19    | 0.5373   | 23.75    | 0.7148   |
| SRCNN [14]    | BD     | 32.05 | 0.8944| 28.80 | 0.8074   | 28.13    | 0.7736   | 25.70    | 0.7770   | 29.47    | 0.8924   |
|               | DN     | 25.01 | 0.6950| 23.78 | 0.5898   | 23.76    | 0.5558   | 21.19    | 0.5373   | 23.75    | 0.7148   |
| FSRCNN [15]   | BD     | 26.23 | 0.8124| 24.44 | 0.7106   | 24.86    | 0.6832   | 24.14    | 0.5556   | 22.99    | 0.7111   |
|               | DN     | 25.20 | 0.7183| 24.00 | 0.6112   | 24.00    | 0.5749   | 22.22    | 0.6096   | 24.20    | 0.7525   |
| VDSR [26]     | BD     | 33.38 | 0.9182| 29.63 | 0.8281   | 28.65    | 0.7922   | 26.77    | 0.8154   | 31.15    | 0.9245   |
|               | DN     | 25.70 | 0.7379| 24.45 | 0.6305   | 24.28    | 0.5900   | 22.90    | 0.6429   | 24.88    | 0.7765   |
| IRCNN_G [64]  | BD     | 29.55 | 0.8124| 24.44 | 0.7106   | 24.86    | 0.6832   | 24.14    | 0.5556   | 22.99    | 0.7111   |
|               | DN     | 25.20 | 0.7183| 24.00 | 0.6112   | 24.00    | 0.5749   | 22.22    | 0.6096   | 24.20    | 0.7525   |
| IRCNN_C [64]  | BD     | 33.38 | 0.9182| 29.63 | 0.8281   | 28.65    | 0.7922   | 26.77    | 0.8154   | 31.15    | 0.9245   |
|               | DN     | 25.70 | 0.7379| 24.45 | 0.6305   | 24.28    | 0.5900   | 22.90    | 0.6429   | 24.88    | 0.7765   |
| SRMDNF [65]   | BD     | 34.09 | 0.9242| 30.11 | 0.8364   | 28.98    | 0.8009   | 27.50    | 0.8370   | 32.97    | 0.9391   |
|               | DN     | 27.74 | 0.8026| 26.13 | 0.6924   | 25.64    | 0.6495   | 24.28    | 0.7092   | 26.72    | 0.8590   |
| FENet (Ours)  | BD     | 34.60 | 0.9277| 30.57 | 0.8433   | 29.22    | 0.8060   | 28.39    | 0.8539   | 34.03    | 0.9459   |
|               | DN     | 28.57 | 0.8164| 26.29 | 0.6945   | 26.01    | 0.6611   | 24.99    | 0.7369   | 28.26    | 0.8611   |

SRMDNF [65]. As shown in Table 9, FENet performs the best on all datasets with BD and DN degradation models. The significantly better results of our method indicate that FENet adapts well to scenarios with multiple degradation models.

In Fig 10, we show two sets of visual results with BD and DN degradation models from the standard benchmark datasets. For BD degradation model, the proposed FENet suppresses the blurring artifacts and recovers sharper edges. For DN degradation model, FENet can not only handle the noise efficiently, but also recover details more accurately. These comparisons further showcase the robustness and effectiveness of our method in handling BD and DN degradation models.

FIGURE 10: Visual results of BD and DN degradation models (×3).

3) Model complexity analysis

In this section, we compare the trade-off between performance, number of parameters and the number of multiplications and additions (Multi-Adds) for our method and existing lightweight SR networks. The Multi-Adds are calculated corresponding to a 1280 × 720 HR image.

Fig 11 shows the PSNR performances of several existing lightweight models, namely VDSR [26], DRCN [27], SR-DenseNet [53], SEINet [10], SRRResNet [31], CARN [2], IMDN [24], SRFBN-S [35], AF-S [58], CBPN [69], LAPAR-A [34], MADNet [29], FALS-A [11], DPN [36], HDRN [25], and OISR-RK2 [19] versus the number of parameters and Multi-Adds for results evaluated on Urban100 for ×4. As shown in Fig 11, FENet achieves state-of-the-art results with less parameters and Multi-Adds operations. This demonstrates that our proposal achieves a better trade-off between model size and reconstruction performance.

FIGURE 11: Comparing capacity vs performance for lightweight state-of-the-art SISR models on Urban100 (×4). Circle sizes are set proportional to the number of multiplications and additions (Multi-Adds).
TABLE 10: Average running time (s) and memory consumption (MB) comparison on Urban100 for \( \times 4 \).

| Methods      | Params | Memory | Running Time(s) | PSNR      |
|--------------|--------|--------|-----------------|-----------|
| CARN [2]     | 1.5M   | 1.116  | 0.032           | 26.07     |
| SRBNet-S [35] | 0.5M   | 2.154  | 0.031           | 25.71     |
| SRFDenseNet [53] | 2.0M | 5.531  | 0.221           | 26.05     |
| IMDN [24]    | 0.7M   | 871    | 0.028           | 26.04     |
| A2F-S [58]   | 0.3M   | 915    | 0.032           | 25.58     |
| LAPAR-A [34] | 0.7M   | 1.240  | 0.053           | 26.14     |
| MSRN [32]    | 1.0M   | 2.108  | 0.031           | 26.06     |
| RCAN [66]    | 0.4M   | 1.131  | 0.087           | 26.82     |
| EDSR [37]    | 0.6M   | 1.128  | 0.035           | 26.64     |
| FENet (Ours) | 0.6M   | 1.531  | 0.018           | 26.20     |

4) Memory complexity and running time analysis

Table 10 illustrates the superiority of the proposed FENet in terms of Inference Time (s) and Memory Consumption (MB), when compared to recent light- and heavy-weight state-of-the-art approaches on Urban100 with scale factor \( \times 4 \). For a fair comparison, we use a single NVIDIA RTX 3090 GPU for evaluation, and their official source code implementations. It can be observed that our model achieves dominant performance in terms of memory usage and time consumption, reflecting its efficiency.

5) Perceptual Metrics

Perceptual metrics better reflect the human judgment of image quality. In this paper, Perceptual Index (PI) [7] is chosen as the perceptual metric. Table 11 shows the PI for those works with publicly available source code, and the same order of magnitude in terms of parameters. We observe that our proposed model obtains better results than all the compared baselines. This demonstrates the ability of the proposed FENet for generating realistic images.

VI. LIMITATIONS AND FUTURE WORK

Although our method is the fastest compared to other SR approaches we have identified the bicubic interpolation operation in (10) as one of the main computational bottlenecks. Thus, we hypothesize that substituting it for a more efficient operation or implementation would effectively speed up our model. Furthermore, the loss function adopted by our method is the distortion-oriented rather than perception-oriented metric, which also limits obtaining better perceptual quality HR images.

In future work, we will explore the extensions of the proposed framework on other image restoration applications, such as deblocking, inpainting, and low-light image enhancement. We also wish to further develop this work by applying our technique to video data. Many streaming services require a large storage to provide high-quality videos. In conjunction with our approach, one may devise a service that stores low-quality videos that go through our SR system to produce high-quality videos on the fly.

VI. CONCLUSION

This paper presents a novel Frequency-based Enhancement Block (FEB). This block is able to naturally decompose features into low and high frequencies and explicitly allocate more computational capacity to high-frequency ones thus improving the discriminative capabilities of the network. The proposed FEB can be easily replaced with commonly used SR blocks. We proved that when replacing SR blocks with FEB we consistently improve the reconstruction error (PSNR: +0.08dB on average) while reducing the number of parameters by half in the model. Furthermore, We showed that the proposed block is orthogonal and complementary to attention-based SR methods. Based on FEB, we proposed a lightweight Frequency-based Enhancement Network (FENet) for accurate image SR. Experimental results on several benchmark datasets demonstrate that our method can achieve superior performance at a moderate size. We hope that the idea of decomposing low- and high-frequency information at the feature level for adaptive computation can provide the computer vision community with a different perspective on network architecture design.

APPENDIX

A. ABBREVIATIONS

The abbreviations and acronyms used in this paper are first introduced in the text, and for convenience, the list of abbreviation is summarized in Table 12.

B. DATA DESCRIPTION

In Table 13, we list a number of image datasets commonly used by the SR community and this work. We specifically indicate their amount of HR images, average resolution, image formats, and category keywords.

REFERENCES

[1] J. Y. Ahn and N. I. Cho. Multi-branch neural architecture search for lightweight image super-resolution. IEEE Access, 9:153633–153646, 2021.
[2] N. Ahn, B. Kang, and K.-A. Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In Proceedings of the European Conference on Computer Vision (ECCV), pages 252–268, 2018.
TABLE 13: List of public image datasets for super-resolution

| Dataset     | Amount | Avg. Resolution | Format | Category Keywords                  |
|-------------|--------|-----------------|--------|------------------------------------|
| Set5        | 5      | (313, 336)      | PNG    | Baby, bird, butterfly, head, woman |
| Urban100    | 14     | (492, 446)      | PNG    | Humans, animals, insects, flowers, vegetables, comic, slides, etc. |
| B100        | 100    | (435, 367)      | JPG    | Animal, building, food, landscape, people, plant, etc. |
| Urban100    | 100    | (984, 797)      | PNG    | Architecture, city, structure, urban, etc. |
| Manga109    | 109    | (826, 1169)     | PNG    | Manga volume                                   |
| DIV2K       | 1000   | (1972, 1437)    | PNG    | Environment, flora, fauna, handmade object, people, scenery, etc. |

[3] Z. An, J. Zhang, Z. Sheng, X. Er, and J. Lv. Rbdn: Residual bottleneck dense network for image super-resolution. *IEEE Access*, 9:103440–103451, 2021.

[4] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

[5] P. Behjati, P. Rodriguez, A. Mehri, I. Hupont, C. F. Tena, and J. Gonzalez. Overnet: Lightweight multi-scale super-resolution with overscaling network. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2694–2703, 2021.

[6] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

[7] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[8] F. W. Campbell and J. G. Robson. Application of Fourier analysis to the visibility of gratings. *The Journal of physiology*, 197(3):551–566, 1968.

[9] R. Chen, H. Zhang, and J. Liu. Multi-attention augmented network for single image super-resolution. *Pattern Recognition*, 122:108349, 2022.

[10] J.-S. Choi and M. Kim. A deep convolutional neural network with selection units for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition Workshops*, pages 154–160, 2017.

[11] X. Chu, B. Zhang, H. Ma, R. Xu, and Q. Li. Fast, accurate and lightweight super-resolution with neural architecture search. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 59–64, IEEE, 2021.

[12] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019.

[13] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

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

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

[16] G. Freedman and R. Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):1–11, 2011.

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

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] X. He, Z. Mo, P. Wang, Y. Liu, M. Yang, and J. Cheng. Ode-inspired network design for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1732–1741, 2019.

[20] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[21] Y. Hu, J. Li, Y. Huang, and X. Gao. Channel-wise and spatial feature modulation network for single image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):3911–3927, 2019.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[23] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015.

[24] Z. Hui, X. Gao, Y. Yang, and X. Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2024–2032, 2019.

[25] K. Jiang, Z. Wang, P. Yi, and J. Jiang. Hierarchical dense recursive network for image super-resolution. *Pattern Recognition*, 107:107475, 2020.

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

[27] J. Kim, J. K. Lee, and K. M. Lee. Deeply-recursive convolu-
tional network for image super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1637–1645, 2016.

[28] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 624–632, 2017.

[29] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo. Madnet: A fast and lightweight network for single-image super-resolution. IEEE transactions on cybernetics, 51(3):1443–1453, 2020.

[30] R. Lan, L. Sun, Z. Liu, H. Lu, Z. Su, C. Pang, and X. Luo. Cascading and enhanced residual networks for accurate single-image super-resolution. IEEE transactions on cybernetics, 51(1):115–125, 2020.

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

[32] J. Li, F. Fang, K. Mei, and G. Zhang. Multi-scale residual network for image super-resolution. In Proceedings of the European Conference on Computer Vision (ECCV), pages 517–532, 2018.

[33] S. Li, Q. Cai, H. Li, J. Cao, L. Wang, and Z. Li. Frequency separation network for image super-resolution. IEEE Access, 8:33768–33777, 2020.

[34] W. Li, K. Zhou, L. Qi, N. Jiang, J. Lu, and J. Jia. Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. Advances in Neural Information Processing Systems, 33, 2020.

[35] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu. Feedback network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3867–3876, 2019.

[36] Y. Liang, R. Timofte, J. Wang, S. Zhou, Y. Gong, and N. Zheng. Single-image super-resolution-when model adaptation matters. Pattern Recognition, 116:107931, 2021.

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

[38] H. Lin and J. Yang. Light weight ibp deep residual network for image super resolution. IEEE Access, 9:93399–93408, 2021.

[39] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu. Residual feature aggregation network for image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2359–2368, 2020.

[40] X. Luo, Y. Xie, Y. Zhang, Y. Qu, C. Li, and Y. Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16, pages 272–289. Springer, 2020.

[41] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yasumura, and K. Aizawa. Sketch-based manga retrieval using deep convolutional neural network. In Proceedings of the 15th International Conference on Pattern Recognition (ICPR), pages 7196–7202. IEEE, 2021.

[42] A. Mehri, P. B. Ardakani, and A. D. Sappa. Mprnet: Multi-path residual network for lightweight image super resolution. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2704–2713, 2021.

[43] Y. K. Ooi, H. Ibrahim, and M. N. Mahyuddin. Enhanced dense space attention network for super-resolution construction from single input image. IEEE Access, 9:126837–126855, 2021.

[44] T. Feleg and M. Elad. A statistical prediction model based on sparse representations for single image super-resolution. IEEE transactions on image processing, 23(6):2569–2582, 2014.

[45] Y. Qiu, R. Wang, D. Tao, and J. Cheng. Embedded block residual network: A recursive restoration model for single-image super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4180–4189, 2019.

[46] P. Rasti, T. Uiboupin, S. Escalera, and G. Anbarjafari. Convolutional neural network super resolution for face recognition in surveillance monitoring. In International conference on articulated motion and deformable objects, pages 175–184. Springer, 2016.

[47] L. Russell. De valois and karen k. de valois. Spatial Vision, (14), 1988.

[48] W. Shi, J. Caballero, C. Ledig, X. Zhuang, W. Bai, K. Bhatia, A. M. S. M. de Marvao, T. Dawes, D. O’Regan, and D. Rueckert. Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In International conference on medical image computing and computer-assisted intervention, pages 9–16. Springer, 2013.

[49] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, W. Bishop, D. Rueckert, and Z. Wang. 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, pages 1874–1883, 2016.

[50] Y. Tai, J. Yang, and X. Liu. Image super-resolution via deep recursive residual network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3147–3155, 2017.

[51] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 114–125, 2017.

[52] T. Tong, G. Li, X. Liu, and Q. Gao. Image super-resolution using dense skip connections. In Proceedings of the IEEE international conference on computer vision, pages 4799–4807, 2017.

[53] F. Wang and M. Gong. Single image super-resolution by residual recovery based on an independent deep convolutional network. IEEE Access, 9:43701–43710, 2020.

[54] F. Wang, H. Hu, and C. Shen. Bam: A lightweight and efficient balanced attention mechanism for single image super resolution. arXiv preprint arXiv:2104.07566, 2021.

[55] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004.

[56] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cham: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018.

[57] Y. Z. J. Y. L. F. Xuehui Wang, Qing Wang and L. Chen. Spatial Vision, 517–532, 2018.

[58] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 114–125, 2017.

[59] T. Tong, G. Li, X. Liu, and Q. Gao. Image super-resolution using dense skip connections. In Proceedings of the IEEE international conference on computer vision, pages 4799–4807, 2017.

[60] F. Wang and M. Gong. Single image super-resolution by residual recovery based on an independent deep convolutional network. IEEE Access, 9:43701–43710, 2020.

[61] F. Wang, H. Hu, and C. Shen. Bam: A lightweight and efficient balanced attention mechanism for single image super resolution. arXiv preprint arXiv:2104.07566, 2021.
[62] R. Zeyde, M. Elad, and M. Prott. On single image scale-up using sparse-representations. In International conference on curves and surfaces, pages 711–730. Springer, 2010.

[63] H. Zhang and V. M. Patel. Densely connected pyramid dehazing network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3194–3203, 2018.

[64] K. Zhang, W. Zuo, S. Gu, and L. Zhang. Learning deep cnn denoiser prior for image restoration. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3929–3938, 2017.

[65] K. Zhang, W. Zuo, and L. Zhang. Learning a single convolutional super-resolution network for multiple degradations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3262–3271, 2018.

[66] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European conference on computer vision (ECCV), pages 286–301, 2018.

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

[68] H. Zhao, X. Kong, J. He, Y. Qiao, and C. Dong. Efficient image super-resolution using pixel attention. In European Conference on Computer Vision, pages 56–72. Springer, 2020.

[69] F. Zhu and Q. Zhao. Efficient single image super-resolution via hybrid residual feature learning with compact back-projection network. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pages 0–0, 2019.

PARICHEHR BEHJATI received the B.Sc. and M.Sc. degree in computer science from Eastern Mediterranean University. She had previously worked as a research assistant in the computer science department, at the same University from 2014 to 2016. She is currently a Ph.D. candidate on deep learning and computer vision at Computer Vision Center, Universitat Autònoma de Barcelona. Her research interests include Deep Learning and Computer Vision. ORCID: 0000-0003-4266-545X

PAU RODRIGUEZ is a Research Scientist at ServiceNow Research, adjunct professor at Universitat Autonoma de Barcelona, and ELLIS member. His main research interest is on machine learning methods that generalize with fewer labeled data, closer to how humans learn. Previously, he did a Ph.D. in Deep Learning and Computer Vision at CVC-UAB. Before, he finished the Master of Artificial Intelligence at KU Leuven. He would like AI to solve the most important problems of humanity. ORCID: 0000-0002-1689-8084

CARLES FERNÁNDEZ TENA received a PhD Cum Laude in Computer Vision and AI from Universitat Autònoma de Barcelona in 2010, also receiving the 2010 Extraordinary Ph.D. Award. He worked for the face recognition company Herta from 2010 to 2021, where he led the R&D department as Director of Research and CTO. He currently is Computer Vision Lead at Oxolo, working on AI-based realistic face and body synthesis. He has published more than 60 scientific articles in peer-reviewed international journals and conferences and participated in several FP7 and H2020 projects. His research interests include Deep Learning, Computer Vision, and HPC for face and video analytics. ORCID: 0000-0001-6185-3427

ARMIN MEHRI received my B.S. degree in Computer Engineering from Eastern Mediterranean University, in 2014 and received my M.S. degree in 2017 from the same university. Currently, He is currently a Ph.D. candidate in Computer Vision Center at the Universitat Autònoma de Barcelona. His main research interests are in the areas of Computer Vision, Image Processing, Image Enhancement, and Image Colorization under cross-modal frameworks resulting in cross-spectral domains. ORCID: 0000-0003-3472-2530

F. XAVIER ROCA received his Ph.D. degree in Computer Science from Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain, in 1990. He is an Associate Professor and the Director of the Department of Computer Science, UAB. He is also a Research Fellow with the Computer Vision Center. He has been a Principal Researcher in several projects (public and private funds). He is working in technological transfer computer vision. The topics of his research are active vision, biometrics and tracking. ORCID: 0000-0002-7043-7334.

SEICHI OZAWA (Senior Member, IEEE) received Dr. Eng. in computer science from Kobe University. He is currently the deputy director of The Center for Mathematical and Data Sciences and full professor with Department of Electrical and Electronic Engineering, Graduate School of Engineering, Kobe University, Japan. His current research interests are deep learning, machine learning, pattern recognition, incremental learning, big data analytics, cybersecurity, text mining, computer vision, and privacy preserving data mining. He published more than 160 journal and conference papers, and book chapters/monographs. He is currently an associate editor of IEEE Trans. on Neural Networks and Learning Systems, IEEE Trans. on Cybernetics and 2 international journals. He is the vice-president for Membership of International Neural Network Society, vice-president for finance of Asia Pacific Neural Network Society, and board of governor of Japan Neural Network Society. He is a member of Neural Networks TC and Smart World TC of IEEE CI Society. ORCID: 0000-0002-0965-0064
JORDI GONZÁLEZ received his PhD degree in Computer Engineering from Universitat Autònoma de Barcelona (UAB) in 2004. At present he is an Associate Professor in Computer Science at the Computer Science Department, UAB. He is also a research fellow at the Computer Vision Center, where he has cofounded three spin-offs (Cloud Size Services, Visual Tagging, and Care Respite) and the Image Sequence Evaluation (ISE Lab) research group. His research interests include machine learning techniques for the computational interpretation of social images, or visual hermeneutics. ORCID: 0000-0001-8033-0306

***