Scale-Recurrent Multi-Residual Dense Network for Image Super-Resolution

Kuldeep Purohit[0000−0002−6709−1627], Srimanta Mandal[0000−0003−3871−6621], and A. N. Rajagopalan[0000−0002−0006−6961]

IPCV Lab, Department of Electrical Engineering, IIT Madras, Chennai, India
kuldeeppurohit3@gmail.com, in.srimanta.mandal@ieee.org,
raju@ee.iitm.ac.in

Abstract. Recent advances in the design of convolutional neural network (CNN) have yielded significant improvements in the performance of image super-resolution (SR). The boost in performance can be attributed to the presence of residual or dense connections within the intermediate layers of these networks. The efficient combination of such connections can reduce the number of parameters drastically while maintaining the restoration quality. In this paper, we propose a scale recurrent SR architecture built upon units containing series of dense connections within a residual block (Residual Dense Blocks (RDBs)) that allow extraction of abundant local features from the image. Our scale recurrent design delivers competitive performance for higher scale factors while being parametrically more efficient as compared to current state-of-the-art approaches. To further improve the performance of our network, we employ multiple residual connections in intermediate layers (referred to as Multi-Residual Dense Blocks), which improves gradient propagation in existing layers. Recent works have discovered that conventional loss functions can guide a network to produce results which have high PSNRs but are perceptually inferior. We mitigate this issue by utilizing a Generative Adversarial Network (GAN) based framework and deep feature (VGG) losses to train our network. We experimentally demonstrate that different weighted combinations of the VGG loss and the adversarial loss enable our network outputs to traverse along the perception-distortion curve. The proposed networks perform favorably against existing methods, both perceptually and objectively (PSNR-based) with fewer parameters.

Keywords: Super-resolution · Deep Learning · Residual Networks · Dense connections.

1 Introduction

Super-resolution (SR) techniques are devised to cope up with the issue of limited-resolution while imaging by generating a high resolution (HR) image from a low resolution (LR) image. However, the possibility of multiple HR images leading to the same LR image makes the problem ill-posed. This can be addressed by regularized mapping of LR image patches to HR counterparts, which are generally extracted from some example images. However, a constrained linear mapping may not be able to represent complex textures of natural images. Deep learning based techniques can behave better in this case by learning a non-linear mapping function.

Convolutional neural networks (CNNs) have played an important role in deep learning based techniques by learning efficient features of images. Deeper CNN architectures can represent an image better than shallower frameworks. However, *deeper the better* assumption does not work often due to vanishing or exploding gradient issue. Thus, gradient flow became an important issue in deep learning based methods. The residual connection [14, 24] helps in this aspect by allowing deeper models to learn. Deeper networks with residual mapping are generally used for higher level vision tasks such as classification. Hence, effective employment of such framework in SR requires some modifications such as removal of batch normalization [25]. Yet, most of these architectures are not able to learn hierarchical features across layers from the LR image. Such features can boost performance, as has been demonstrated by a residual dense network using a sequence of residual dense blocks [45]. However, the number of parameters for such deeper dense networks often becomes a bottleneck when limited computational resources are available.

SR for different scale factors requires separate training of the network. Joint training for different scale factors can address the issue, as has been attempted by VDSR [19], which needs a bicubic interpolated LR image as input. However, this strategy can come in the way of exploiting hierarchical features from the original LR image, and crucial details may be lost. Further, processing such a high dimensional image for a large number of layers demands higher computational resources. Another way to deal with the situation is to learn the model for lower scale factor such as 2 and use it to initialize the learning for higher factors such as 3, 4, etc [25]. However, this strategy is parametrically inefficient and does not work well for higher scale factors (e.g., 8).

In order to accommodate different up-sampling factors while keeping a check on the number of parameters, we propose a scale-recurrent strategy that helps in transferring learned filters from lower scale factors to higher ones. We use our scale-recurrent strategy in conjunction with a smaller version of Residual Dense Network (RDN) [45], where we use fewer Residual Dense Blocks (RDBs) to reduce the number of parameters as compared to the original RDN. We choose RDBs as building blocks since the combination of residual and dense connections can help in overcoming their individual limitations. This combination allows for efficient flow of information throughout the layers while eliminating the vanishing gradient issue. We refer to this scale-recurrent residual dense network as SRRDN.

Motivated by the recent developments in network designs based on dense connections, we introduce multiple residual connections within an RDB using 1×1 convolutions that results in superior performance with marginal parametric cost. The proposed units are termed as *Multi-Residual Dense Blocks (MRDB)*. Our proposed scale-recurrent network with MRDBs is termed as multi-residual dense network (MRDN).

We demonstrate that training our network with a pixel-reconstruction loss (L1 loss) produces results with good PSNR/SSIM performance. Recent findings suggest that although these metrics measure the objective quality of HR reconstruction, they are not necessarily correlated with perceptual quality [3]. To improve perceptual performance (for photo-realistic image super-resolution), we include a GAN-based framework along with VGG loss function into our model. Different weighting schemes for adversarial loss and VGG losses produce different quality of results, which allows us to traverse the perception-distortion curve [3]. Specifically, VGG loss along with pixel-reconstruction loss is used to train a network (MRDN), which leads to good PSNR values (albeit with lower perceptual quality). Also, this network is further trained with only VGG loss and adversarial loss to obtain a network (MRDN-GAN) that generates better perceptual quality than MRDN (but with lower PSNR). During test-time, a soft-thresholding based strategy is further utilized to reach a desirable trade-off between PSNR and perceptual quality.

2 Related Works and Contributions

Super-resolving a single image generally requires some example HR images to import relevant information for generating the HR image. Two streams of approaches make use of the HR example images in their frameworks: i) Conventional, and ii) deep learning based. The functioning of conventional SR approaches depends on finding patches, similar to the target patch in the database of patches. Since there could be many similarities, one needs to regularize the problem. Thus, most of the conventional approaches focus on discovering regularization techniques in SR such as Tikhinov [44], total-variation [29], Markov random field [18], non-local-mean [27, 11, 28], sparsity-based prior [41, 42, 10], and so on [9, 28].

Although, the sparsity-based prior works quite efficiently, the linear mapping of information may fail to represent complex structures of an image. Here, deep-learning based approaches have an upper hand as they can learn a non-linear mapping between LR and corresponding HR image [6, 8, 40, 19, 36, 22, 23, 34, 37, 25, 43, 16]. Deep learning stepped into the field of SR via SRCNN [7] by extending the notion of sparse representation using CNN. The non-linearity involved in CNN is able to better represent complex structures than conventional approaches to yield superior results. However, going to deeper architectures increases the difficulty in training such networks. Employing a residual network into the frame along with skip connections and recursive convolution can mitigate this issue [19, 20]. Following such an approach, VDSR [19] and DRCN [20] methods have demonstrated performance improvement. The power of recursive blocks involving residual units to create a deeper network was explored in [36]. Recursive unit in conjunction with a gate unit can act as a memory unit that adaptively combines the previous states with the current state to produce a super-resolved image [37]. However, these approaches interpolate the LR image to the HR grid and feed it to the network. But this increases the computational requirement due to the higher dimension.

To circumvent the dimension issue, networks exists that are tailored to extract features from the LR image which are then processed in subsequent layers. At the end layer, up-sampling is performed to match with the HR dimension [8, 24]. This process can be made faster by reducing the dimension of the features going to the layers that map from LR to HR and is known as FSRCNN [8]. ResNet [14] based deeper network with generative adversarial network (GAN) [12] can produce photo-realistic HR results by including perceptual loss [17] in the network, as devised in SRResNet [24]. The perceptual loss is further used with a texture synthesis mechanism in GAN based model to improve SR performance [34]. Though these approaches are able to add textures in the image, sometimes the results contain artifacts. The model architecture of SRResNet [24] has been simplified and optimized to achieve further improvements in EDSR [25]. This was later modified in MDSR [25], which performs joint training for different scale factors by introducing scale-specific feature extraction and pixel-shuffle layers.

2.1 Contributions

The contributions of the presented work are listed below:

- We present a scale recurrent SR framework, which works in conjunction with Residual Dense Blocks. The scale recurrent design helps in producing better results for higher scale factors while eliminating the requirement of large number of parameters.
- The multi-Residual Dense Blocks, we propose involve a series of multiple residual and dense connections within a block. This leads to effective gradient propagation by mitigating feature redundancy.
- To achieve perceptually attractive results, our network is also trained with deep feature loss, and adversarial loss alongside pixel reconstruction loss. We experimentally demonstrate that different weights on these losses produce results that traverse along the perception-distortion curve. The two complementary outputs are effectively fused during test time using a soft-thresholding based technique to achieve perception-distortion trade-off.

3 Architecture Design

The success of recent approaches has emphasized the importance of network design. Specifically, most recent image and video SR approaches are built upon two popular image classification networks: residual networks [14] and densely connected networks [15]. These network designs have also enjoyed success and achieved state-of-the-art performance in other image restoration tasks such as image denoising, dehazing, and deblurring. Motivated by the generalization capability of such advances in network designs, the recent work of RDN [45] proposed a super-resolution network which involves a mixture of both residual and dense connections and yields state-of-the-art results. The fundamental block of this network is RDB, which we too adopt in our work.

While DenseNet was proposed for high-level computer vision tasks (e.g., object recognition), RDN adopted and improved upon this design to address image SR. Specifically, batch-normalization (BN) layers were removed as they hinder the performance of the network by increasing computational complexity and memory requirements. The pooling layers are removed too since they could discard important pixel-level information. To enable a higher growth rate, each dense block is terminated with a 1×1 conv layer (Local Feature Fusion) and its output is added to the input of the block using Local Residual Learning. This strategy has been demonstrated to be very effective for SR [45].

Our network contains a sequence of 6 RDBs which extract deep hierarchical features from the input LR image. The outputs of each RDB are concatenated and fed into a set of 1×1 and 3×3 layers, which results in reduced number of feature maps. This strategy helps in the efficient propagation of hierarchical features through the network by adaptive fusion of shallow and deep features extracted in LR space [45]. These features are fed into a pixel-shuffle layer, followed by a convolution layer that yields the HR image. We also add the bilinear up-sampled image to the output layer of the network that enforces the network to focus on learning high-frequency details.

3.1 Scale-Recurrent Design

Most of the existing SR approaches handle different scale factors independently, hence neglecting inter-scale relationships. They need to be trained independently for different scale factors. However, VDSR [19] can address the issue by jointly training a network for multiple scales. This kind of training requires LR images of different resolutions to be up-sampled by bi-cubic interpolation prior to feeding to the network. Interpolation by a large factor causes loss of information and requires higher computational resources as compared to scale-specific networks.

Our network's global design is a multi-scale pyramid which recursively uses the same convolutional filters across scales. This is motivated by the fact that a network capable of super-resolving an image by a factor of 2 can be recursively used to super-resolve the image by a factor $2s, s = 1, 2, 3 \ldots$. Even with the same training data, the recurrent exploitation of shared weights works in a way similar to using data multiple times to learn parameters, which actually amounts to data augmentation with respect to scales. We design the network to reconstruct HR images in intermediate steps by progressively performing a 2× upsampling of the input from the previous level. Specifically, we first train a network to perform SR by a factor of 2 and then re-utilize the same weights to take the output of 2× as input and result into an output at resolution 4×. This architecture is then fine-tuned to perform 4× SR. We experimentally found that such initialization (training for the task of 2× SR) leads to better convergence for larger scale factors. Ours is one of the first approaches to re-utilize the parameters across scales, which significantly reduces the number of trainable

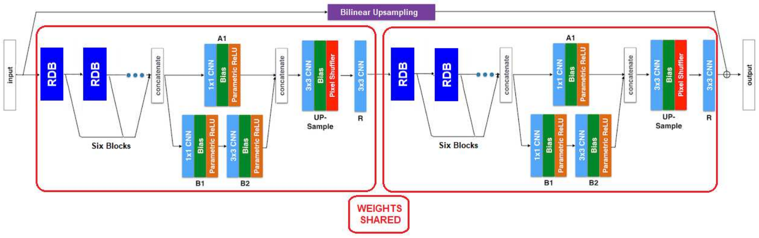


Fig. 1: Network architecture of the proposed Scale-Recurrent Residual Dense Network for 4× SR.

parameters while yielding performance gains for higher scale factors. We term our network SRRDN, whose 4× SR version is shown in Fig. 1.

3.2 Multi-Residual Dense Blocks

We also propose improvements in the structure of RDB for efficient extraction of high-resolution features from low-resolution images. The effectiveness of residual and dense connections has been proved in various vision tasks; yet, they cannot be considered as optimum topology. For example, too many additions on the same feature space may impede information flow in ResNet [15]. The possibility of same type of raw features from different layers can lead to redundancies in DenseNet [4]. Some of these issues are addressed in recent image classification networks [4, 39]. However, these designs are optimized for image classification tasks and their applicability to image restoration has not been explored yet.

Dual Path Networks (DPN) [4] bridge the densely connected network [15] with higher order recurrent neural networks [35] to provide new interpretation of dense connections. Mixed Link Networks [39] have also shown that both dense connections and residual connections belong to a common topology. These methods utilize these interpretations to design hybrid networks that incorporate the core idea of DenseNet with that of ResNet. These works demonstrate that inclusion of addition and concatenation-based connections improves classification accuracy, and is more effective than going deeper or wider. Essentially, DenseNet connects each layer to every other layer in a feed-forward fashion. Such connections alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters. ResNet and its variants enable feature re-usage while DenseNet enables new feature exploration; both being important for learning good representations. By carefully incorporating these two network designs into dual-path topologies, DPN shares common features while maintaining the flexibility to explore new features through dual path architectures. Inspired by the DPN network that was originally designed for the task of image classification, we propose a design change specially tailored for super-resolution.

An RDB of SRRDN already contains multiple paths connecting the current layer to previous network layers. One connection is present in the form of a concatenation of features, which is similar to the connections in DenseNet. Al-

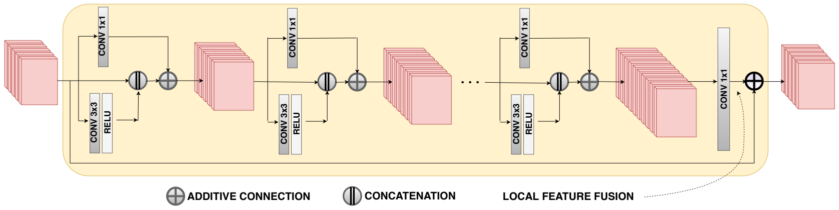


Fig. 2: Structure of our Multi-Residual Dense Block. Within each module, concatenation operation is performed using the features estimated by the conv 3×3 layer, and addition operation is performed on the features estimated by 1×1 conv layer (which continuously increase the number of feature maps to match the size of concatenated output). At the end of the block, a 1×1 conv layer performs local feature fusion to adaptively control the output information.

though growth rates affect the performance positively, it is harder to train a large number of dense blocks which possess a higher growth rate, as has been experimentally demonstrated in [45]. This can be addressed by Local Feature Fusion (see Fig. 2), by including a second connection that stabilizes the training of wide network. This brings down the number of output feature-maps to the number of input feature-maps and enables introduction of a single residual connection between the input and the output of the block (Local Residual Learning).

In order to further improve the gradient flow during training, we introduce a third connection: Multi-Residual connections. Essentially, at each intermediate layer of the block, we convolve the input features using a 1×1 conv layer and add them to the output obtained after the concatenation operation. This type of connection has two properties: Firstly, existing feature channels get modified, which helps in deeper and hierarchical feature extraction. Secondly, it enables learning of equally meaningful features even with a lower growth-rate during feature concatenation. This strategy promotes new feature exploration with a moderate growth rate and avoids learning of redundant features. These two features enable improved error gradient propagation during training. Our scale-recurrent framework built using MRDBs as basic blocks is termed as MRDN.

4 Perceptual and Objective Quality Trade-off

Conventional pixel reconstruction based loss functions such as L1 loss encourage a network to produce results with better objective quality but it could be perceptually inferior. In contrast, VGG/GAN-based loss functions enforce the network to produce perceptually better results [3]. Most of the existing methods, once trained, cannot be altered to produce results with different objective quality and/or perceptual quality, during test time. We propose to use two networks to overcome this issue. Our first network (MRDN) is trained with a weighted combination of L1 and VGG54 losses so that it results in outputs with better objective quality. Our second network has the same architecture as the first but it is trained with a combination of perceptually motivated losses such as VGG54 feature-based loss and adversarial loss. The adversarial loss pushes the network output to the manifold of natural high-resolution images using a discriminator network that is trained to differentiate between the super-resolved images and original photo-realistic images. We refer to this network as MRDN-GAN.

Let θ represent the weights and biases in the network i.e., $\theta = \{W, B\}$. Given a set of training image pairs ${I_k}^L, {I_k}^H$, we minimize the following Mean Absolute Error (MAE) to obtain results with better objective quality.

$$l_{MAE}(\theta) = ||F({I_k}^L, \theta) - {I_k}^H||_1 \tag{1}$$

To obtain perceptually superior results (photo-realistic appearance), the following loss function is used:

$$l_{VGG/i.j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}({I_k}^H)_{x,y} - \phi_{i,j}(F({I_k}^L, \theta))_{x,y})^2. \tag{2}$$

Here $W_{i,j}$ and $H_{i,j}$ describe the dimensions of the respective feature maps within the VGG network. Additionally, a conditional adversarial loss is also adopted that encourages sharper texture in the images generated by the network. The objective function for minimization becomes:

$$l_{CGAN}(F, D) = \mathbf{E}[\log D(U({I_k}^L), {I_k}^H)] + \mathbf{E}[\log(1 - D(U({I_k}^L), F({I_k}^L, \theta)))], \tag{3}$$

where $\mathbf{E}$ represents the expectation operation, and $U(\cdot)$ bi-linearly up-samples ${I_k}^L$ to match the resolution of ${I_k}^H$. Here, D represent discriminator network, whose architecture is similar to [24], except that we feed two images to the network by concatenating them along channel dimension.

Once the two networks are trained, we pass each test image through them, separately. The outputs are expected to have complementary properties. MRDN returns an HR image (I_{HR1}) which is as close as possible to the ground-truth (in terms of mean-squared error (MSE)). However, as explained in [3], such objectively superior output would be perceptually inferior. On the other hand, MRDN-GAN leads to a perceptually superior image (I_{HR2}), while compromising on objective quality (in terms of PSNR). To obtain results which lie in between these two images on the plane, we need to preserve the sharpness features from I_{HR2}, while bringing the intensities closer to I_{HR1}. To enable this flexibility, we adopt a soft-thresholding based approach as described in [5]. The adjusted image I can be obtained through the following formulation:

$$I = I_{HR2} + S_\lambda(I_{HR1} - I_{HR2}). \tag{4}$$

where $S_\lambda(\cdot)$ is a pixel-wise soft-thresholding operation that depends on λ which controls the amount of information to be combined from the two images. λ is calculated as $\lambda = S_v(\mathcal{R}(K * \gamma))$, where S_v is a vector that contains sorted non-zero entries of the matrix $(I_{HR1} - I_{HR2})$, $\mathcal{R}$ is the rounding-off operation and K is the number of elements of S_v. The parameter $\gamma \in (0, 1]$ needs to be controlled in our approach.

Generally, when increasing the value of threshold γ, the resultant image tends to have higher objective quality and lower perceptual quality. This is because a larger γ can remove more high-frequency details and, thus, decrease the perceptual quality. Since some of these high-frequency details can negatively affect the objective quality, removing them leads to better PSNR. Different values for the threshold γ leads to different trade-offs between I_{HR1} and I_{HR2}.

5 Experimental Results

5.1 Experimental Setup

Here, we specify the details of training setup, test data and evaluation metrics. **Datasets and degradation models.** Following [38, 25, 45, 43], we use 800 training images from DIV2K dataset [38] as training set. For testing, we use five standard benchmark datasets: Set5 [1], Set14 [42], B100 [30], Urban100 [16], Manga109 [31], and PIRM-self [2]. We consider bicubic(BI) down-sampling to generate the LR images.
Evaluation metrics. The SR results are evaluated with two metrics: PSNR and perceptual score. For a given image I, the perceptual metric is defined as

$$P(I) = \frac{1}{2}((10 - M(I) + N(I)) \tag{5}$$

where $M(I)$ and $N(I)$ are estimated using [26] and [32], respectively. These metrics have been used to evaluate different approaches in the PIRM SR Challenge. **Training settings.** Data augmentation is performed on the 800 training images, which are randomly rotated by 90°, 180°, 270° and flipped horizontally. Our model is trained by ADAM optimizer [21] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The initial leaning rate is set to 10^{-4} and is then decreased by half every 2×10^5 iterations of back-propagation.
Implementation Details The network is implemented using Pytorch library. For training the first network, we used a weighted sum of VGG54 loss and L1 Loss. For the second network, we used a weighted sum of VGG54 loss and conditional-GAN loss. The experiments have been conducted on a machine with i7-4790K CPU, 64GB RAM and 1 NVIDIA Titan X GPU using PyTorch [33]. During training, we considered a batch of randomly extracted 16 LR RGB patches of size 32×32 pixels. Training the first network (MRDN) took approximately 40 hours. The second network (MRDN-GAN) was then trained for 26 hours.

5.2 Perceptually Motivated Results

This work has been used for the purpose of participating in the PIRM 2018 SR Challenge, which focuses on photo-realistic results (measured using perceptually motivated metric) while maintaining certain levels of tolerance in terms of root mean squared error (RMSE). In this challenge, there exist three tracks corresponding to different ranges of RMSE for scale factor of ×4. Track 1 corresponds to RMSE ≤ 11.5. Track 2: $11.5 \leq$ RMSE < 12.5, while Track 3 included results

Table 1: Quantitative results (PSNR & P-Score) for factor 4 (for region 3 of PIRM challenge). Bold indicates best performance.

Method	Set5		Set14		B100		Urban100		PIRM-self	
	PSNR	P-Score	PSNR	P-Score	PSNR	P-Score	PSNR	P-Score	PSNR	P-Score
SRGAN [24]	29.40	3.61	26.02	2.91	25.16	2.59	22.79	**3.45**	**26.23**	2.35
ENET-PAT [34]	28.56	**2.93**	25.75	3.01	25.38	2.93	23.68	3.47	25.06	2.69
MRDN-GAN	**30.08**	3.43	**26.67**	**2.82**	**25.74**	**2.37**	**24.54**	3.55	25.79	**2.19**

with RMSE ≥ 12.5. Perceptually attractive images are generally rich in various high-frequency (HF) image details. Thus, the objective is to bring out HF details while super-resolving the given LR images such that the resultant images yield better perceptual score. We employed our networks to generate results with scores suitable for each track and proved that our technique can elegantly facilitate quality control during test time. Our team *REC-SR* secured the 7^{th}, 7^{th} and 10^{th} ranks in Tracks 1, 2 and 3, respectively.

Quantitative Results: Meeting the Perception-Distortion Curve As explained in Section 4, we analyze the effect of different loss configurations on the performance of the network for single image super-resolution. Our networks are trained for ×4 SR and tested on 100 images from the PIRM-self set. We have plotted the trade-off between the mean-perceptual score and mean square error in Fig. 3(a). The points labeled in blue represent loss configurations which contained higher weights for pixel-reconstruction loss, thus leading to superior objective quality. Specifically, we trained our network with different weighted combinations of L1 loss and VGG loss. The slight variation in the performance is due to small differences in the duration of training as well as the relative coefficient of the L1 loss. This relative coefficient was varied in the range $(0.05, 1)$ to obtain various models.

The points labeled in red represent loss configurations which contained higher weight to adversarial loss, leading to better perceptual quality. Specifically, we trained our network using various weighted combinations of VGG loss and conditional GAN loss. The variation in the performance is due to differences in the

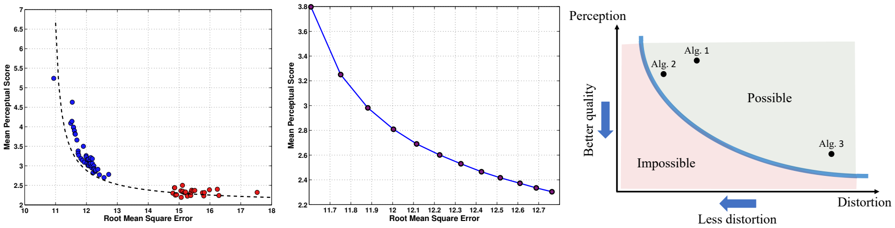


Fig. 3: (a) Perceptual and RMSE scores of various trained instances of our network. The blue points correspond to our network trained with VGG+L1 loss while the red points correspond to training with VGG loss+adversarial loss. Results are evaluated for 4× SR on PIRM-self validation dataset; (b) Results for Track 2 using soft-thresholding on the output of our two networks for various thresholds; (c) The expected behavior of an SR algorithm in perception-distortion plane.

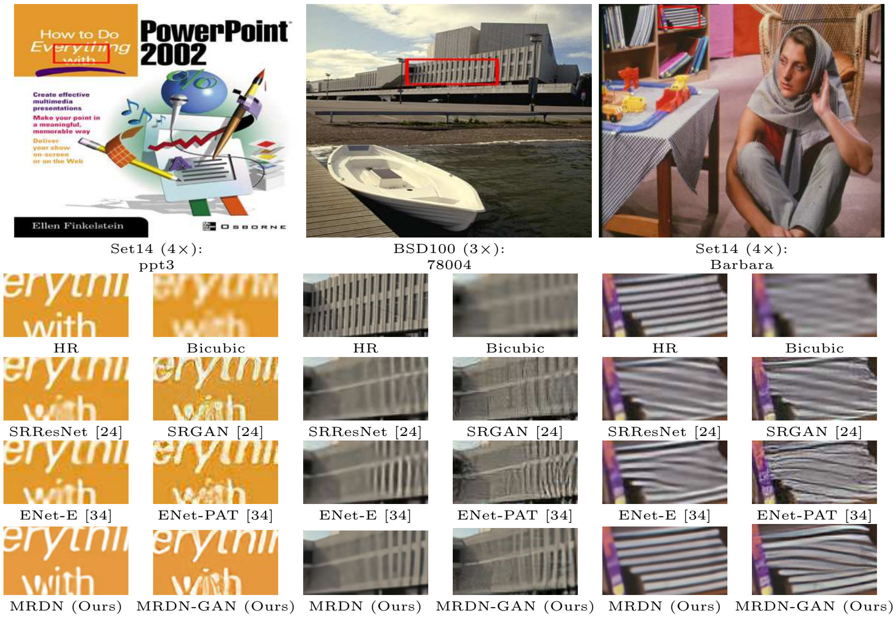


Fig. 4: Visual comparison for 4× SR on images from Set14 and BSD100 datasets.

duration of training as well as the relative coefficient of the adversarial loss. This relative coefficient was varied between $(0.02, 0.005)$. Results of our two networks are combined using a soft-thresholding strategy and plotted in Fig. 3(b).

Note that the distribution of these evaluations follows the curve (shown in Fig. 3(c)) as explained in [3]. Specifically, the point at the left extreme corresponds to the network purely trained using L1 loss from scratch. Consistent with the findings of [3], it leads to the lowest MSE but a very poor perceptual score. On the other hand, the right-most point corresponds to a network fine-tuned purely using the adversarial loss (no VGG or L1 loss). This yields one of the best perceptual performance but fares poorly in terms of MSE. Our results show strong agreement with the argument that an algorithm can be potentially improved only in terms of its distortion or in terms of its perceptual quality, one at the expense of the other. We observed that a balanced combination of these loss functions is more appropriate in practice.

The results are further quantitatively compared with the perceptual SR benchmarks in terms of PSNR and perceptual score (P-Score) in Table 1. One can note that our network produces results with better PSNR values and P-scores than existing approaches on almost all the datasets.

Qualitative Results With the help of adversarial training, image SR methods such as SRGAN [24] and ENet [24] propose networks that can produce perceptually superior (photo-realistic) results (while being objectively inferior). They also present their objectively superior counterparts: SResNet and ENetE, which

Table 2: Ablation studies (PSNR & SSIM) for factor 4.

Method	Set5		Set14		B100		Urban100	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
MRDN†	32.27	0.8961	28.64	0.7833	27.62	0.7378	26.14	0.7883
SRRDN	32.34	0.8968	28.68	0.7839	27.65	0.7381	26.27	0.7920
MRDN	**32.48**	**0.8983**	**28.77**	**0.7855**	**27.71**	**0.7396**	**26.45**	**0.7956**

are not trained using adversarial loss. We visually compare the results of these approaches with our networks: MRDN and MRDN-GAN for the task of 4× SR.

Visual comparisons of the results of our networks MRDN and MRDN-GAN with these techniques on images from standard SR benchmarks are given in Fig. 4. In all the images, it can be seen that the results of SRResNet and ENetE suffer from blurring artifacts. This demonstrates the insufficiency of only pixel-reconstruction losses. However, the efficient design of our MRDN leads to improved recovery of scene texture in challenging regions. For example, in image "ppt3", all the compared methods fail to recover the letters 'i' and 't'. However, our proposed MRDN recovers them. On the other hand, GAN-based methods of SRGAN, and ENetPAT produce distorted scene textures. The results of ENet-PAT are sharper than SRGAN but it generates unwanted artifacts and arbitrary edges (e.g., the result for the image "78004"). In contrast, our proposed MRDN-GAN leads to textures which are closer to that of the ground-truth HR image too. Similar observations can be found in other images. These comparisons show that the design of SR network plays an important role in both objective and perceptual quality of SR.

In Fig. 5, we compare the results of our model on Urban100 dataset with state-of-the-art SR approaches which are not perceptually motivated for a scale factor of 4. For such texture-rich scenes, a major challenge is to bring out high frequency image details. One can observe that most of the existing approaches fail in this aspect and their results are blurred (see Fig. 5). However, our MRDN-GAN is capable of generating sufficiently detailed textures.

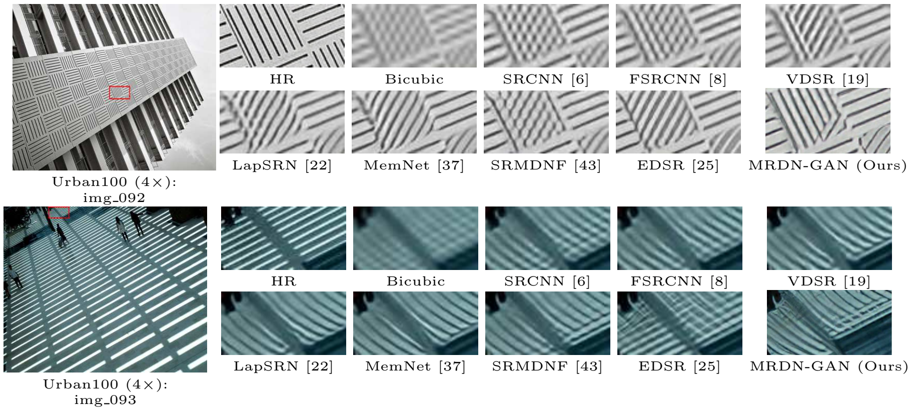


Fig. 5: Visual comparisons for 4× SR on Urban100 dataset.

Table 3: Quantitative results with bicubic degradation model. Bold indicates best performance, red color second best, and blue color the third best performance.

Method	Scale	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×4	28.42	0.8104	26.00	0.7027	25.96	0.6675	23.14	0.6577	24.89	0.7866
SRCNN [6]	×4	30.48	0.8628	27.50	0.7513	26.90	0.7101	24.52	0.7221	27.58	0.8555
FSRCNN [8]	×4	30.72	0.8660	27.61	0.7550	26.98	0.7150	24.62	0.7280	27.90	0.8610
VDSR [19]	×4	31.35	0.8830	28.02	0.7680	27.29	0.0726	25.18	0.7540	28.83	0.8870
LapSRN [22]	×4	31.54	0.8850	28.19	0.7720	27.32	0.7270	25.21	0.7560	29.09	0.8900
MemNet [37]	×4	31.74	0.8893	28.26	0.7723	27.40	0.7281	25.50	0.7630	29.42	0.8942
EDSR [25]	×4	32.46	0.8968	28.80	**0.7876**	27.71	**0.7420**	**26.64**	**0.8033**	**31.02**	0.9148
SRMDNF [43]	×4	31.96	0.8925	28.35	0.7787	27.49	0.7337	25.68	0.7731	30.09	0.9024
D-DBPN [13]	×4	32.47	0.8980	**28.82**	0.7860	**27.72**	0.7400	26.38	0.7946	30.91	0.9137
RDN [45]	×4	32.47	**0.8990**	28.81	0.7871	**27.72**	0.7419	26.61	0.8028	31.00	**0.9151**
MRDN (ours)	×4	**32.48**	0.8983	28.77	0.7855	27.71	0.7396	26.45	0.7956	30.92	0.9137
Bicubic	×8	24.40	0.6580	23.10	0.5660	23.67	0.5480	20.74	0.5160	21.47	0.6500
SRCNN [6]	×8	25.33	0.6900	23.76	0.5910	24.13	0.5660	21.29	0.5440	22.46	0.6950
FSRCNN [8]	×8	20.13	0.5520	19.75	0.4820	24.21	0.5680	21.32	0.5380	22.39	0.6730
SCN [40]	×8	25.59	0.7071	24.02	0.6028	24.30	0.5698	21.52	0.5571	22.68	0.6963
VDSR [19]	×8	25.93	0.7240	24.26	0.6140	24.49	0.5830	21.70	0.5710	23.16	0.7250
LapSRN [22]	×8	26.15	0.7380	24.35	0.6200	24.54	0.5860	21.81	0.5810	23.39	0.7350
MemNet [37]	×8	26.16	0.7414	24.38	0.6199	24.58	0.5842	21.89	0.5825	23.56	0.7387
MSLapSRN [23]	×8	26.34	0.7558	24.57	0.6273	24.65	0.5895	22.06	0.5963	23.90	0.7564
EDSR [25]	×8	26.96	0.7762	24.91	0.6420	24.81	0.5985	22.51	0.6221	24.69	0.7841
D-DBPN [13]	×8	27.21	0.7840	25.13	0.6480	24.88	0.6010	22.73	0.6312	**25.14**	**0.7987**
MRDN (ours)	×8	**27.27**	**0.7860**	**25.15**	**0.6511**	**24.95**	**0.6020**	**22.82**	**0.6340**	24.99	0.7950

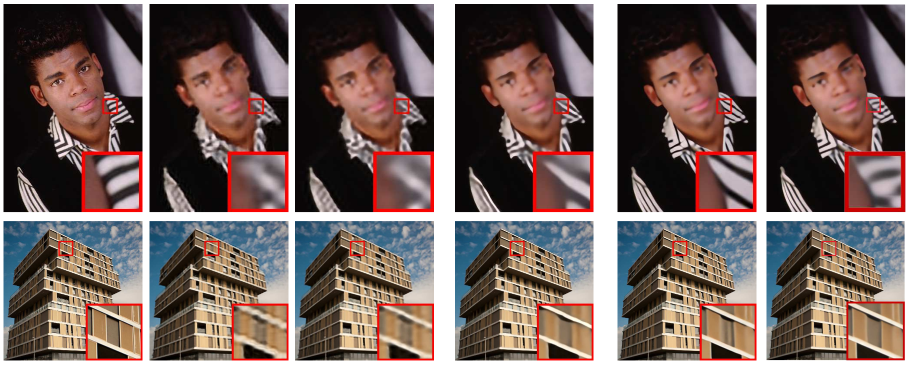
5.3 Ablation Study

In Table 2, we evaluate the role of different components of our network. We have compared the performance of MRDN with its counterpart that does not share the weights across scales (MRDN†) and matches the number of parameters in MRDN (by reducing the number of blocks). Further, in Table 2, we compare the performance of MRDN with SRRDN to demonstrate the effectiveness of multi-residual dense connections. One can observe that MRDN† performs inferior to its scale-recurrent version (i.e., MRDN). This clearly explains the advantages of scale-recurrent strategy. The performance improvement of MRDN over SRRDN underlines the efficiency of multi-residual dense connection.

We further evaluate the performance of our network with state-of-the-art SR approaches on standard SR benchmarks in terms of PSNR and SSIM in Table 3. One can observe that our network MRDN performs comparably to the best performing approaches such as RDN, DBPN, EDSR etc., although our network has significantly fewer parameters. Moreover, we are able to produce best results using Set5 dataset for factor 4. Our scale recurrent strategy reveals its benefits for scale factor 8 leading to state-of-the-art results for most of the datasets. The quantitative improvements can be further verified through the qualitative results given in Fig. 6. This demonstrates that our network with appropriate loss functions can not only produce perceptually better results but also it has the ability to generate HR results that are objectively superior.

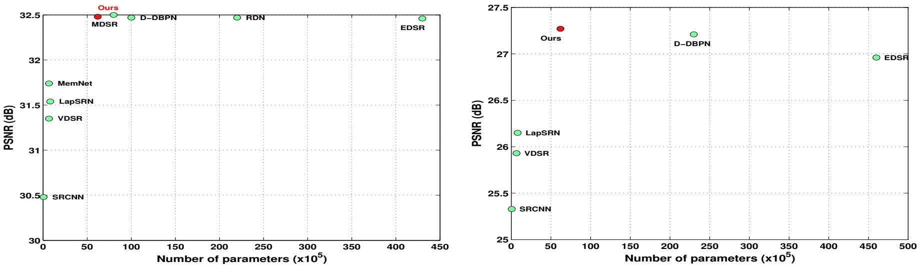
5.4 Parametric Analysis

We analyze performance with respect to size of models for different approaches in Fig. 7. Our MRDN has fewer parameters than that of state-of-the-art approaches EDSR, MDSR, DDBPN and RDN, leading to a better trade-off between model size and performance.



HR SRCNN [6] VDSR [19] MSLapSRN[22] D-DBPN [13] Ours

Fig. 6: Visual comparisons with existing approaches for super-resolution by a factor of 8 on 302008.png from BSD100, and img 087.png from Urban100.



(a) Approaches for 4× SR. (b) Approaches for 8× SR.

Fig. 7: Comparison with existing approaches in terms of PSNR and number of parameters required for scaling factors 4 and 8. Results are evaluated on Set5.

6 Conclusions

We proposed a scale-recurrent deep architecture, which enables transfer of weights from lower scale factors to the higher ones, in order to reduce the number of parameters as compared to state-of-the-art approaches. We experimentally demonstrated that our scale-recurrent design is well-suited for higher up-sampling factors. The error gradient flow was improved by elegantly including multiple residual units (MRDN) within the Residual Dense Blocks. To produce perceptually better results, VGG-based loss functions were utilized along with a GAN framework. Different weights were assigned to the loss functions to obtain networks focused on improving either perceptual quality or objective quality during super-resolution. The perception-distortion trade-off was addressed by a soft-thresholding technique during test time. We demonstrated the effectiveness of our parametrically efficient model on various datasets.

References

1. Bevilacqua, M., Roumy, A., Guillemot, C., line Alberi Morel, M.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Proceedings of the British Machine Vision Conference. pp. 135.1–135.10. BMVA Press (2012). https://doi.org/http://dx.doi.org/10.5244/C.26.135
2. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L.: 2018 PIRM Challenge on Perceptual Image Super-resolution. ArXiv e-prints pp. 1–22 (Sep 2018)
3. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. arXiv preprint arXiv:1711.06077 (2017)
4. Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., Feng, J.: Dual path networks. In: Advances in Neural Information Processing Systems. pp. 4467–4475 (2017)
5. Deng, X.: Enhancing image quality via style transfer for single image super-resolution. IEEE Signal Processing Letters **25**(4), 571–575 (2018)
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2), 295–307 (Feb 2016). https://doi.org/10.1109/TPAMI.2015.2439281
7. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 184–199. Springer International Publishing, Cham (2014)
8. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 391–407. Springer International Publishing, Cham (2016)
9. Dong, W., Zhang, L., Shi, G., Li, X.: Nonlocally centralized sparse representation for image restoration. IEEE Transactions on Image Processing **22**(4), 1620–1630 (April 2013). https://doi.org/10.1109/TIP.2012.2235847
10. Dong, W., Zhang, L., Shi, G., Wu, X.: Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. IEEE Transactions on Image Processing **20**(7), 1838 –1857 (Jul 2011). https://doi.org/10.1109/TIP.2011.2108306
11. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: IEEE International Conference on Computer Vision (ICCV). pp. 349–356 (Sept 2009). https://doi.org/10.1109/ICCV.2009.5459271
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014)
13. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1664–1673 (June 2018)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR. vol. 1, p. 3 (2017)

16. Huang, J., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5197–5206 (June 2015). https://doi.org/10.1109/CVPR.2015.7299156
17. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 694–711. Springer International Publishing, Cham (2016)
18. Kanemura, A., ichi Maeda, S., Ishii, S.: Superresolution with compound markov random fields via the variational {EM} algorithm. Neural Networks **22**(7), 1025 – 1034 (2009). https://doi.org/10.1016/j.neunet.2008.12.005
19. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1646–1654 (June 2016). https://doi.org/10.1109/CVPR.2016.182
20. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1637–1645 (June 2016). https://doi.org/10.1109/CVPR.2016.181
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), http://arxiv.org/abs/1412.6980
22. Lai, W., Huang, J., Ahuja, N., Yang, M.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5835–5843 (July 2017). https://doi.org/10.1109/CVPR.2017.618
23. Lai, W., Huang, J., Ahuja, N., Yang, M.: Fast and accurate image super-resolution with deep laplacian pyramid networks. CoRR **abs/1710.01992** (2017), http://arxiv.org/abs/1710.01992
24. Ledig, C., Theis, L., Huszr, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 105–114 (July 2017). https://doi.org/10.1109/CVPR.2017.19
25. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1132–1140 (July 2017). https://doi.org/10.1109/CVPRW.2017.151
26. Ma, C., Yang, C.Y., Yang, X., Yang, M.H.: Learning a no-reference quality metric for single-image super-resolution. Computer Vision and Image Understanding **158**, 1–16 (2017)
27. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Non-local sparse models for image restoration. In: IEEE 12th International Conference on Computer Vision. pp. 2272 –2279 (29 2009-oct 2 2009). https://doi.org/10.1109/ICCV.2009.5459452
28. Mandal, S., Bhavsar, A., Sao, A.K.: Noise adaptive super-resolution from single image via non-local mean and sparse representation. Signal Processing **132**, 134 – 149 (2017). https://doi.org/http://dx.doi.org/10.1016/j.sigpro.2016.09.017
29. Marquina, A., Osher, S.J.: Image super-resolution by TV-regularization and bregman iteration. Journal of Scientific Computing **37**, 367–382 (2008). https://doi.org/10.1007/s10915-008-9214-8
30. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and

measuring ecological statistics. In: Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001. vol. 2, pp. 416–423 vol.2 (July 2001). https://doi.org/10.1109/ICCV.2001.937655

31. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications **76**(20), 21811–21838 (Oct 2017). https://doi.org/10.1007/s11042-016-4020-z, https://doi.org/10.1007/s11042-016-4020-z
32. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a" completely blind" image quality analyzer. IEEE Signal Process. Lett. **20**(3), 209–212 (2013)
33. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
34. Sajjadi, M.S.M., Schlkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 4501–4510 (Oct 2017). https://doi.org/10.1109/ICCV.2017.481
35. Soltani, R., Jiang, H.: Higher order recurrent neural networks. arXiv preprint arXiv:1605.00064 (2016)
36. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2790–2798 (July 2017). https://doi.org/10.1109/CVPR.2017.298
37. Tai, Y., Yang, J., Liu, X., Xu, C.: Memnet: A persistent memory network for image restoration. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 4549–4557 (Oct 2017). https://doi.org/10.1109/ICCV.2017.486
38. Timofte, R., Agustsson, E., Gool, L.V., Yang, M., Zhang, L., et al.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1110–1121 (July 2017). https://doi.org/10.1109/CVPRW.2017.149
39. Wang, W., Li, X., Yang, J., Lu, T.: Mixed link networks. arXiv preprint arXiv:1802.01808 (2018)
40. Wang, Z., Liu, D., Yang, J., Han, W., Huang, T.: Deep networks for image super-resolution with sparse prior. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 370–378 (Dec 2015). https://doi.org/10.1109/ICCV.2015.50
41. Yang, J., Wright, J., Huang, T., Ma, Y.: Image super-resolution via sparse representation. IEEE Transactions on Image Processing **19**(11), 2861–2873 (Nov 2010). https://doi.org/10.1109/TIP.2010.2050625
42. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Boissonnat, J.D., Chenin, P., Cohen, A., Gout, C., Lyche, T., Mazure, M.L., Schumaker, L. (eds.) Curves and Surfaces. pp. 711–730. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
43. Zhang, K., Zuo, W., Zhang, L.: Learning a single convolutional super-resolution network for multiple degradations. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3262–3271 (June 2018)
44. Zhang, X., Lam, E., Wu, E., Wong, K.: Application of Tikhonov regularization to super-resolution reconstruction of brain MRI images. In: Gao, X., Mller, H., Loomes, M., Comley, R., Luo, S. (eds.) Medical Imaging and Informatics, Lecture Notes in Computer Science, vol. 4987, pp. 51–56. Springer Berlin Heidelberg (2008). https://doi.org/10.1007/978-3-540-79490-5_8
45. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)