Learning Reduced-Resolution and Super-Resolution Networks in Synch

Bolun Cai[1*] Xiangmin Xu[1†] Kailing Guo[1] Kui Jia[1] Dacheng Tao[2]
[1]School of Electronic and Information Engineering, South China University of Technology, China
[2]UBTECH Sydney AI Centre, School of IT, FEIT, The University of Sydney, Australia

Abstract

Recent studies have shown that deep convolutional neural networks achieve the excellent performance on image super-resolution. However, CNN-based methods restore the super-resolution results depending on interpolations a lot. In this paper, we present an end-to-end network (Reduced & Super-Resolution Network, RSRNet) for reduced-resolution and super-resolution without any cheap interpolation. 1) The reduced-resolution network is trained without supervision, which preserves more effective information and brings better visual effect in the LR image. 2) The super-resolution network learns a sub-pixel residual with dense connection to accelerate the convergence and improve the performance. For image super-resolution, we evaluate the proposed network on four benchmark datasets and set a new state of the art. Moreover, the reduced-resolution network in RSRNet is applicable to generate photo-realistic low-resolution image and improve image compression by using existing image codecs.

1. Introduction

Single-image super-resolution (SR) is a traditional computer vision problem, which aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) image. Single-image SR is widely used in computer vision applications such as HDTV [12], medical imaging [25], satellite imaging [32] and surveillance [36], where high-frequency details are required on demand. As mobile social network (*e.g*. Google+, WeChat and Twitter) continues to soar, thumbnail super-resolution is another important application in data storage and transmission over limited-capacity channels.

Single-image SR have been studied over decades. Early methods including bicubic interpolation [6], Lanczos resampling [9], gradient profile [27] and patch redundancy [11], are based on statistical image prior or internal patch representation. Recently, learning-based methods are proposed to model a mapping from LR to HR patches, such as neighbor embedding [4], sparse coding [33] and random forest [23]. However, SR is highly ill-posed since the process from HR to LR contains non-invertible downsampling.

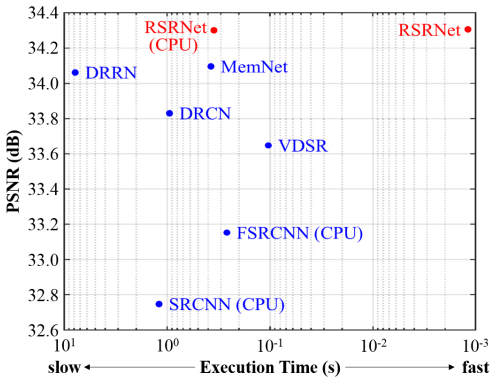


Figure 1. Plot of the trade-off between accuracy and speed for different methods on Set5 [2] with the scale factor $\times 3$. The proposed RSRNet achieves better restoration quality than existing methods, and is 100 times faster than MemNet [30].

Due to the powerful learning capability, deep convolutional neural networks (CNNs) have achieved state-of-the-art performance in many computer vision tasks, such as image classification [26], object detection [10] and image segmentation [5]. Recently, CNNs are introduced to address the ill-posed inverse problem of SR and have demonstrated superiority over traditional learning paradigms. Super-resolution convolutional neural network (SRCNN) [7] as the pioneer network predicts the nonlinear mapping from LR to HR in an end-to-end manner. To reduce the computational complexity, fast SRCNN (FSRCNN) [8] and efficient sub-pixel convolutional neural network (ESPCN) [24] upscale the resolution only at the output layer. Kim *et al*. [18] developed a very deep super-resolution (VDSR) network with 20 convolutional layers by residual learning, and Mao *et al*. [21] proposed a 30-layer residual encoder-decoder (RED) network with symmetric skip connections to help

*The early work was completed while the author was with Tencent Wechat AI.

†X. Xu is the corresponding author. (Email: xmxu@scut.edu.cn)

training. Deeply-recursive convolutional network (DRCN) [19] introduces a very deep recursive layer via a chain structure with 16 recursions, and deep recursive residual network (DRRN) [29] adopts recursive residual units to control the model parameters while increasing the depth.

Despite achieving excellent performance, the above CNN-based SR methods highly depend on interpolation based down/up-sampling (in Figure 2). Limitations problems mainly come from the following two aspects:

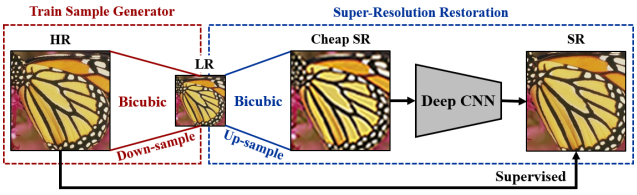


Figure 2. The classical framework based on CNNs for image SR [7, 18, 21, 19, 29]. The HR image is down-sampled to synthesize the train samples. Then, the deep model learns the mapping between the SR and the cheap SR upscaled by the same factor via bicubic interpolation.

§ **Down-sampling problem.** For down-sampling, the model trained for a specific interpolation cannot work well with the other interpolations, and different interpolations bring significant difference pertaining to restoration accuracy. Therefore, all the solutions mentioned above generally assume that the degradation is bicubic interpolation (the default setting of $imresize()$ in *Matlab*) when shrinking an image. However, the bicubic function based on weighting transformation is a cheap reduced-resolution processing, which abandons the useful high-frequency details for HR image restoration.

§ **Up-sampling problem.** With cheap up-sampling, the networks [18, 21, 19, 29] increase the resolution in the preprocessing or at the first-layer to learn the residual of interpolation. However, the up-sampling do not bring additional information to solve the ill-posed restoration problem. To replace cheap interpolation, ESPCN [24] and FSRCNN [8] adopt sub-pixel shuffle and deconvolution layer for efficiency improvement, respectively. But without cheap up-sampling, they carry the input and restore the details as auto-encoder, and so they converge slowly.

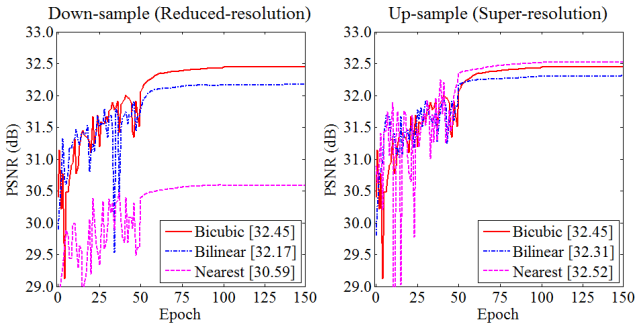
To address the above problems, we propose an end-to-end network (Reduced & Super-Resolution Network, **RSRNet**) without any cheap interpolation, which is trained for reduced-resolution and super-resolution in synch.

- A learnable reduced-resolution network (RRNet) is trained without supervision, which preserves more effective information and brings better visual effect in the LR image. For self-supervision, the super-pixel residual is adopted with a novel activation function, called quantized bilateral ReLU (Q-BReLU).
- A super-resolution network (SRNet) learns a sub-pixel residual with dense connection to accelerate the convergence and improve the performance: first, dense pixel representation trained with deep supervision extracts the multi-scale feature by long/short-term memory; second, sub-pixel residual restores the super-resolution result without up-sampling method.

To illustrate the effectiveness of the proposed method, Figure 1 shows the cost-performance of several state-of-the-art methods. Compared to existing CNN-based methods, RSRNet achieves the best performance with lower computational complexity.

2. Sampling Problem

We design an experiment to discuss the sampling problem of image SR. In this section, we re-implemente[1] the baseline model SRCNN [7] with Adam [20] optimizer. Learning rate decreases by the factor of 0.1 from 10^{-3} to 10^{-5} every 50 epochs.



(a) Down-sampling (b) Up-sampling

Figure 3. The convergence and accuracy analyses on different down/up-sampling methods.

2.1. Learn Restoration for Specified Down-sample

Almost all CNN-based SR methods [7, 8, 24, 21, 18, 19, 29, 30] are trained for the specified down-sampling. Shown in Table 1, the asymmetric of interpolations in the train/test phase will result in poor restoration, even worse than the direct-bicubic interpolation. The reason is that the CNNs learn the targeted mapping for the specified degradation. Moreover, different interpolations (nearest-neighbor, bilinear, bicubic) bring significant difference on convergence rate and restoration accuracy shown in Figure 3(a). The degradation with bicubic interpolation contains more useful information, so the model for bicubic converges faster and

[1]The SRCNN re-implemented by us achieves the better perform than the implementation (32.39dB) of authors [7].

achieves better performance. However, the bicubic interpolation is still a cheap reduced-resolution process, which abandons useful details for SR. In this paper, we train a joint network for reduced-resolution and corresponding super-resolution in synch.

Table 1. Compare with different down-sample degradations for scale factors $\times 3$ on Set5 [2] pertaining to PSNR (dB). (The PSNR of bicubic interpolation as a baseline is 30.39.)

Train \ Test	Nearest	Bilinear	Bicubic	Avg.
Nearest	**30.59**	29.76	30.56	30.30
Bilinear	25.38	**32.17**	31.19	29.58
Bicubic	27.59	31.92	**32.45**[1]	**30.70**

2.2. Learn Mapping from Cheap Up-sample

In the popular CNN-based methods (*e.g*. SRCNN [7], VDSR [18], DRCN [19], DRRN [29]), cheap up-sampling as preprocessing is used to increase the resolution before or at the first layer of the network. Shown in Figure 3(b), interpolations do not bring additional information to improve the restoration accuracy. Instead of improving, the complex preprocessing (bicubic) introduces smooth and inaccurate interpolation prematurely, resulting the network hard to be trained. Conversely, the nearest-neighbor interpolation achieves the best performance, because it selects the raw value of the nearest point and does not consider the values of neighboring points. To address this problem, ESPCN [24] and FSRCNN [8] carry the input and restore the details without cheap up-sampling, but which results the convergence rate decreasing. Therefore, we propose a sub-pixel residual learning to accelerate the convergence and improve the performance accordingly.

3. Reduced&Super-Resolution Network

In this paper, we propose a reduced & super-resolution network (RSRNet), an end-to-end system that simultaneously learns the mappings between reduced-resolution and super-resolution. In Figure 4, we present the architecture designs of RSRNet.

3.1. Unsupervised Reduced-Resolution Network

A learnable reduced-resolution network (RRNet) is trained without supervision, which learns the super-pixel residual with a novel activation function, called quantized bilateral ReLU (Q-BReLU).

3.1.1 Super-pixel Residual Learning

Without supervised signal, RRNet is capable to retain useful information and discard redundant information proactively. However, the multi-layer network is an end-to-end relation requiring very long-term memory. For this reason, the LR image generated from the learned features contains irrational artifacts. We can solve this problem simply by a super-pixel residual-learning.

In RRNet, the pixel of LR output $\mathbf{L}_{x,y}$ of scale factors $1/d$ is largely similar to the super-pixel of HR input $\mathbf{H}$. Therefore, we define a super-pixel residual image $\mathbf{R}^r_{x,y} = \mathbf{L}_{x,y} - \frac{1}{|\Omega|}\sum_{\{m,n\}\in\Omega_{d\cdot(x,y)-\lfloor d/2\rfloor}} \mathbf{H}_{m,n}$, where $\lfloor\cdot\rfloor$ takes the integer downwardly and Ω is a neighborhood with the size of $d\times d$. In $\mathbf{R}^r$, most values are likely to be very low and even close to zero. Formally, the reduced-resolution mapping is denoted as $\mathcal{F}^r[\mathbf{H}] \rightarrow \mathbf{R}^r$, which includes a inference model (three convolution layers with the size of 3×3 and the stride of 1) and a down-sample layer (a convolution layer with the size of $d\times d$ and the stride of d). The original mapping is recast into

$$\mathbf{L}_{x,y} = \mathcal{F}^r[\mathbf{H}]_{x,y} + \frac{1}{|\Omega|}\sum\nolimits_{\{m,n\}\in\Omega_{d\cdot(x,y)-\lfloor d/2\rfloor}} \mathbf{H}_{m,n}, \quad (1)$$

which can be implemented by feedforward neural networks with shortcut connections [14] and average pooling.

3.1.2 Quantized Bilateral ReLU (Q-BReLU)

Standard choices of nonlinear activation function such as rectified linear unit (ReLU) offers local linear to overcome the problem of vanishing gradient. However, ReLU is designed for classification problems and not preferred for image restoration. In particular, ReLU inhibits values only when they are less than zero, which might lead to response overflow especially without supervision. Moveover, the general digital image is quantified to integers between 0 and 255.

To overcome this limitation, quantized bilateral rectified linear unit (Q-BReLU) is proposed to keep bilateral restraint and response quantization shown in Figure 5. Q-BReLU is a variation of BReLU [3], which is adopted for haze transmission restoration. BReLU is defined as $f_{brelu} = \max\left(\min\left(x, t_{\max}\right), t_{\min}\right)$, where $t_{\min,\max}$ is the marginal value. Denoting $\Delta t = t_{\max} - t_{\min}$ for terse expression, Q-BReLU is defined as

$$f_{qbrelu}(x) = \frac{\Delta t}{Q-1}\left\lfloor \frac{Q-1}{\Delta t}\left(f_{brelu}(x) - t_{\min}\right) + 0.5 \right\rfloor + t_{\min}, \quad (2)$$

where Q is the number of quantities.

However, the gradient of Q-BReLU alternates between 0 and ∞ according to (2). We exploit an approximate gradient with local continuous for back-propagation learning. To retain center quantization and zero-mean deviation, BReLU as a spline function is adopted to fit Q-BReLU shown in Figure 5(b). Therefore, the approximate gradient of Q-

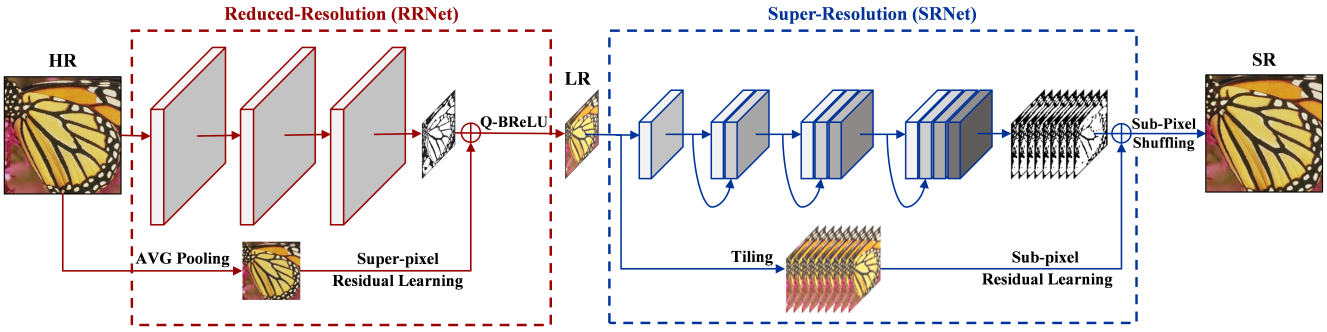
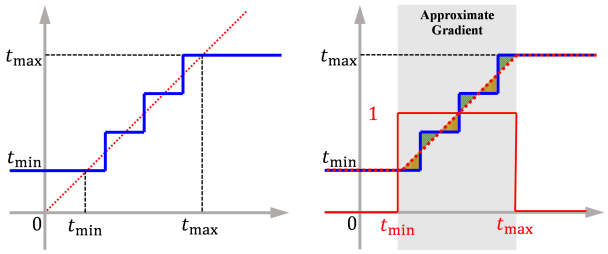


Figure 4. The deep convolutional network for reducing in synch with improving resolution. 1) The reduced-resolution network (RRNet) is trained unsupervisedly with super-pixel residual learning and Q-BReLU function. 2) The super-resolution network (SRNet) learns the dense pixel representation by sub-pixel residual learning.



(a) Q-BReLU (b) The gradient of Q-BReLU

Figure 5. The quantized bilateral rectified linear unit (Q-BReLU) with 2-bit quantities $Q = 2^2$. (a) Q-BReLU is denoted in ***solid blue***. (b) The approximate gradient of Q-BReLU is denoted in ***dashed red***, and ***solid red*** denotes the spline fitting Q-BReLU. **Center Quantization**: the high-precision value is rounded at the nearest quantization interval (between neighboring blue dasheds); **Zero-mean Deviation**: the positive/negative (green/yellow area) deviation balance out the approximate bias.

BReLU is defined as

$$\frac{\partial f_{qbrelu}(x)}{\partial x} = \begin{cases} 1, & t_{\min} < x < t_{\max} \\ 0, & otherwise \end{cases} . \quad (3)$$

To verify the impact of Q-BReLU, we illustrate an example for $\times 2$ down-sampling in Figure 6. The LR image generated without Q-BReLU contains irrational noises, while the result with Q-BReLU appears naturally.

(a) HR (b) Bicubic (c) w/o Q-BReLU (d) Q-BReLU

Figure 6. The images *baby* from Set5 [2] generated by RRNet with/without Q-BReLU.

3.2. Residual Super-Resolution Network

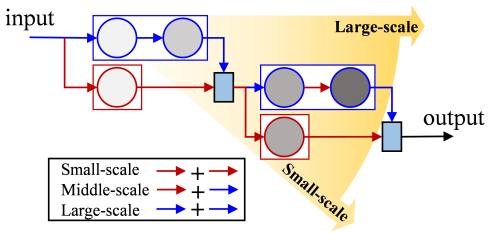
The super-resolution network (SRNet) is a dense architecture to learn the sub-pixel residual, which achieves a high performance and efficiency without any cheap interpolation.

3.2.1 Dense Pixel Representation

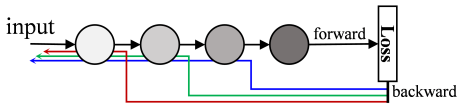
SRNet connects pixel representation densely [15] for pixel-wise prediction. Each layer produces k feature maps, then it follows that the l'th layer has $k \times l$ input feature-maps. In the deep connection layers, a large number of feature maps increase the computational cost and the model size. It has been demonstrated in [28] that a 1×1 convolution as a bottleneck improves the computational efficiency and keeps model compact. Figure 7 shows how the dense pixel representation improves the performance.

Multi-scale feature. For SR, different kinds of components may be relevant to different scales of neighbourhood in the LR image. In [35], multi-scale neighborhood has been proven effective for SR. SRNet is a kind of multi-scale architectures to improve the performance: the receptive field is getting the larger when network stacks more layers as illustrated in Figure 7(a). Given fixed kernel of size 3×3, there are three streams of multi-scales (small/middle/large-scale) corresponding to $\{5, 7, 9\}$, respectively.

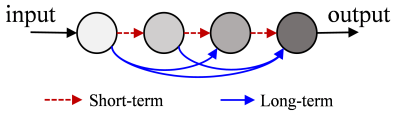
Deeply-supervised learning. Compared to image classification, image SR is a low-level vision task, where different layers complete similar task to restore the details progressively. The kernels in shallow layers can be shared to boost the performance recursively. However, recursions are hard to train due to exploding/vanishing gradients. Skip-connection, similar to deeply-supervised learning shown in Figure 7(b), alleviates the vanishing-gradient problem and enhances the feature propagation in the network. Back-propagation goes through a small number of layers when supervision information goes directly from loss function to



(a) Multi-scale feature



(b) Deeply-supervised learning



(c) Long/short-term memory

Figure 7. The analysis to the architecture of SRNet.

early layers.

Long/short-term memory. Sparse-coding is a representative method for example-based SR. Sparse coefficients are passed into a dictionary for restoring HR patches. SRNet can be viewed as a kind of sparse coding: the convolutional kernels with size 3×3 are equivalent to dictionaries; the bottlenecks with nonlinear activation function are equivalent to sparse coefficients. With dense connections, the neural unit learns multi-level dictionaries under long/short-term memory shown in Figure 7(c). The short-term memory is generated from the previous unit, and the long-term memory is generated from the memory in earlier stage. The bottleneck as a gate unit adaptively controls the forget coefficients of the long/short-term memory.

3.2.2 Sub-pixel Residual Learning

The HR image can be decomposed into a low-frequency information (low-resolution image) and high-frequency information (residual image). In SRNet, the input image $\mathbf{L}$ and output image $\mathbf{S}$ share the same low-frequency information. Without any cheap interpolation, we adopt a sub-pixel residual learning to carry the LR input to the HR result.

Depending on different sub-pixel location in HR space, the residual patterns containing d^2 channels are activated by a convolution of size 1×1. Sub-pixel shuffle [24] is a periodic operator that rearranges the elements of an $H \times W \times d^2$ tensor to a tensor of shape $d \cdot H \times d \cdot W$. In the mathematica formula, the sub-pixel residual image is written by $\mathbf{R}^s_{\lfloor x/d \rfloor, \lfloor y/d \rfloor, d\cdot(y \setminus d)+(x \setminus d)} = \mathbf{S}_{x,y} - \mathbf{L}_{\lfloor x/d \rfloor, \lfloor y/d \rfloor}$, where $\setminus$ denotes the remainder operator. To learn the sub-pixel residual image similarity to (1), the restoration result is defined by

$$\mathbf{S}_{x,y} = \mathcal{F}^s[\mathbf{L}]_{\lfloor x/d \rfloor, \lfloor y/d \rfloor, d\cdot(y \setminus d)+(x \setminus d)} + \mathbf{L}_{\lfloor x/d \rfloor, \lfloor y/d \rfloor}, \quad (4)$$

where $\mathcal{F}^s[\cdot]$ is the sub-pixel residual prediction. It is effective to be implemented by a tile layer and an element-wise sum layer.

4. Experiments

4.1. Implementation Details

The model is trained on 91 images from [33] and 200 images from the training set of [22], which are widely used for SR [18, 19, 29, 30]. Following [7], the luminance channel is only considered in YCbCr colour space, because humans are more sensitive to luminance changes. We train a specific network for each upscaling factor ($\times 2, 3, 4$).

Detailed configurations and parameter settings of RSRNet shown in Figure 4 are summarized in Table 2. Motivated by the experiment, a leaky ReLU (LReLU) $f_{lrelu}(x) = \max(x, 0.05x)$ instead of ReLU is used as the activation function except the output layer. The layers with residual learning are initialized by drawing randomly from a Gaussian distribution ($\mu = 0, \sigma = 0.001$), because most values in the residual images are likely to be zero or small. The other filter weights are initialized according to [13].

Table 2. The detailed configurations of RSRNet.

	Shortcut	Trunk	pad	stride	initialize
Reduced-Resolution	[$d \times d$ AVG Pool]	[3 × 3,64 LReLU] × 3	1	1	MSRA
		[$d \times d$, 1]	0	d	Gaussian
	[SUM Eltwise Q-BReLU]		–	–	–
Super-Resolution	[d^2 Tile]	[3 × 3,64 LReLU]	1	1	MSRA
		[1 × 1,64 LReLU 3 × 3,64 LReLU] × 3	1	1	MSRA
		[1 × 1, d^2]	0	1	Gaussian
	[SUM Eltwise Sub-pixel]		–	–	–

In the training phase, we rotate the images with the degree of $90°, 180°, 270°$ for data augmentation. Sub-images are extracted to ensure that all pixels in the original image appear once and only once as the ground truth of the training data. For $\times 2, 3, 4$, we set the size of training sub-images to be 60, 69, 72, respectively. The model is trained with L1 loss using an Adam [20] optimizer in the *Caffe* [17] package. Learning rate decreases by half from 10^{-3} to 10^{-5}

every 50 epochs. The final layer learns 10 times slower as in [7]. Based on the parameters above, training RSRNet with a batch-size of 256 roughly takes one day using one Nvidia GeForce GTX 1080 GPU.

4.2. Comparisons with Image Super-Resolution

To assess SR quantitatively, RSRNet is evaluated of three different scale factors ($\times 2, 3, 4$) on four benchmarks [2, 34, 22, 16]. We compute the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) of the results to compare 6 recent methods, including SRCNN [7], FSRCNN [8], VDSR [18], DRCN [19], DRRN [29] and MemNet [30]. As shown in Table 3, the proposed RSRNet outperforms the state-of-the-art methods. Qualitative comparisons among SRCNN [7], FSRCNN [8] and VDSR [18] are illustrated in Figure 10 with their public codes. As we can see, our method can produce relatively sharper edges and contours while others generate blurry results. In addition, severe distortions are found in some reconstructed results using existing methods, whereas RSRNet can reconstruct the texture patterns and avoid the distortions.

As for effectiveness, we evaluate the execution time using the public code of these methods. The experiments are conducted with an Intel CPU (Xeon E5-2620, 2.1 GHz) and an NVIDIA GPU (GeForce GTX 1080). Figure 1 shows the PSNR performance of the state-of-the-art methods versus the execution time. The super-resolution phase of RSRNet achieves higher cost-performance than existing methods. Even on the mobile CPU platform (A9 of iPhone 6S), our method for scale factors $\times 3$ implemented by *ncnn*[2] library processes a 150×150 image with approximately 200 ms.

4.3. Comparisons with Image Reduced-Resolution

Existing image down-sampling (*e.g.* nearest-neighbor, bilinear, bicubic) is based on locally weighting. Interpolation transformation struggles to find pixel-wise weights of plausible solutions, which are typically over-smooth and have poor perceptual quality – that is, they will lose valiable high-frequency details such as texture. We illustrate this problem in Figure 8, where multiple potential solutions with high texture details are weighted to create the smooth results of bilinear or bicubic.

The proposed method provides a powerful RRNet for generating photo-realistic LR images with high perceptual quality. RRNet encourages the LR image to move towards regions of potential manifold with high probability of containing photo-realistic texture. We show two standard test images (*lena* and *baboon*) with RSRNet compared to other down-sampling methods in Figure 9. RSRNet generates relatively sharper and richer texture patterns.

[2]https://github.com/Tencent/ncnn

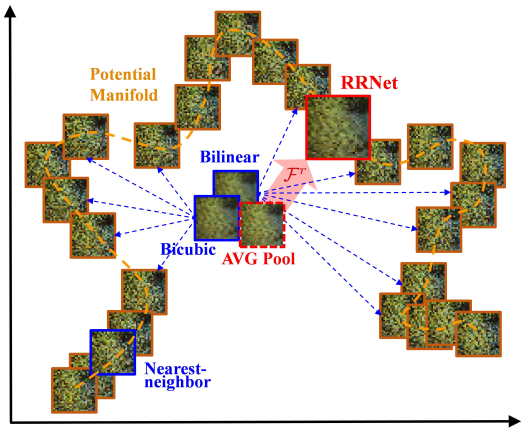
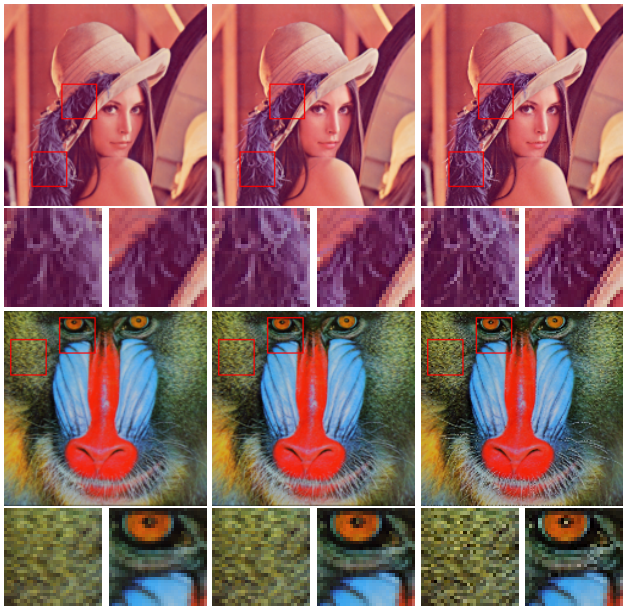


Figure 8. Illustration of potential solutions and LR results obtained with existing interpolations and RSRNet. The solutions based on linear-function (bilinear and bicubice) appear overly smooth due to the pixel-wise weighing of possible solutions; the solution based on nearest-neighbor optionally select a sample at the manifold space. RRNet learns the residual $\mathcal{F}^r$ from pixel-wise average (average pooling) towards the potential manifold, and produces perceptually more convincing solutions.



(a) Bilinear (b) Bicubic (c) RSRNet

Figure 9. Reduced-resolution results with scale factor $3\times$. RSRNet's result looks slightly sharper than bilinear and bicubic interpolation. (1) The first row shows image *lena*. RSRNet generates more realistic textures of hair that are significantly sharper. (2) The second row shows image *baboon*. RSRNet produces high-frequency patterns missing in the results of bicubic and bilinear, *e.g.* the fur texture and the lightspot in the baboon's eyeball.

Table 3. Average PSNR/SSIM for scale factors $\times 2$, $\times 3$ and $\times 4$ on datasets Set5 [2], Set14 [34], B100 [22] and Urban [16]. Red color indicates the best performance and blue color indicates the second best performance.

Dataset	Scale	Bicubic	SRCNN [7]	FSRCNN [8]	VDSR [18]	DRCN [19]	DRRN [29]	MemNet [30]	RSRNet
Set5	$\times 2$	33.66/0.9299	36.66/0.9542	37.00/0.9558	37.53/0.9587	37.63/0.9588	37.74/0.9591	37.78/**0.9597**	**37.92**/0.9549
	$\times 3$	30.39/0.8682	32.75/0.9090	33.16/0.9140	33.66/0.9213	33.82/0.9226	34.03/0.9244	34.09/0.9248	**34.29**/**0.9300**
	$\times 4$	28.42/0.8104	30.48/0.8628	30.71/0.8657	31.35/0.8838	31.53/0.8854	31.68/0.8888	31.74/0.8893	**31.92**/**0.9032**
Set14	$\times 2$	30.24/0.8688	32.45/0.9067	32.63/0.9088	33.03/0.9124	33.04/0.9118	33.23/0.9136	33.28/0.9142	**34.11**/**0.9286**
	$\times 3$	27.55/0.7742	29.30/0.8215	29.43/0.8242	29.77/0.8314	29.76/0.8311	29.96/0.8349	30.00/0.8350	**30.30**/**0.8578**
	$\times 4$	26.00/0.7027	27.50/0.7513	27.59/0.7535	28.01/0.7674	28.02/0.7670	28.21//0.7721	28.26/**0.7723**	**28.34**/0.7539
B100	$\times 2$	29.56/0.8431	31.36/0.8879	31.50/0.8906	31.90/0.8960	31.85/0.8942	32.05/0.8973	32.08/0.8978	**32.52**/**0.9074**
	$\times 3$	27.21/0.7385	28.41/0.7863	28.52/0.7893	28.82/0.7976	28.80/0.7963	28.95/**0.8004**	28.96/0.8001	**28.99**/0.7969
	$\times 4$	25.96/0.6675	26.90/0.7101	26.96/0.7128	27.29/0.7251	27.23/0.7233	27.38/**0.7284**	**27.40**/0.7281	27.22/0.7010
Urban	$\times 2$	26.88/0.8403	29.51/0.8946	29.85/0.9009	30.76/0.9140	30.75/0.133	31.23/0.9188	31.31/0.9195	**32.27**/**0.9305**
	$\times 3$	24.46/0.7349	26.24/0.7991	26.42/0.8064	27.14/0.8279	27.15/0.8276	27.53/**0.8378**	27.56/0.8376	**28.03**/0.8346
	$\times 4$	23.14/0.6577	24.52/0.7226	24.60/0.7258	25.18/0.7524	25.14/0.7510	25.44/**0.7638**	25.50/0.7630	**25.66**/0.7145

Figure 10. Super-resolution results with scale factor $\times 3$ and average PSNR/SSIM of each sub-figure. (1) The first row shows image *253027* from B100 [22]. RSRNet accurately reconstruct the original pattern, while severe distortions are found in the results using other methods. (2) The second row shows image *ppt3* from Set14 [34]. Texts in RSRNet are sharp and identified, while others are blurry. (3) The last row shows image *img002* from Urban [16]. RSRNet can well reconstruct the lines while other methods generate blurry results.

Table 4. Size(bytes)/bpp/SSIM result of RSRNet + (7-Zip/JPEG2000), JPEG and JPEG2000 on datasets Set5 [2].

Image	JPEG		JPEG2000	
	RSRNet + 7-Zip	JPEG ($q = 44$)	RSRNet + JPEG2000	JPEG2000 ($r = 13\%$)
baby (510×510)	19,674/0.6051/**0.9829**	**17,450/0.5367**/0.9614	**16,870/0.5189/0.9829**	19,503/0.6000/0.9713
bird (288×288)	6,974/0.6726/**0.9656**	**6,883/0.6639**/0.9609	6,233/0.6012/**0.9656**	**6,191/0.5971**/0.9581
butterfly (255×255)	**6,031/0.7420/0.9608**	9,305/1.1448/0.9411	6,116/0.7524/**0.9608**	**4,778/0.5878**/0.9160
head (279×279)	**5,887/0.6050/0.8376**	5,894/0.6057/0.8082	**5,254/0.5400**/0.8376	5,546/0.5700/**0.8470**
woman (228×342)	**6,520/0.6689/0.9439**	7,069/0.7252/0.9326	6,001/0.6157/**0.9439**	**5,809/0.5960**/0.9436
Average	**9,017/0.6587/0.9300**	9,320/0.7353/0.9290	**8,095/0.5800/0.9300**	8,366/0.5902/0.9272

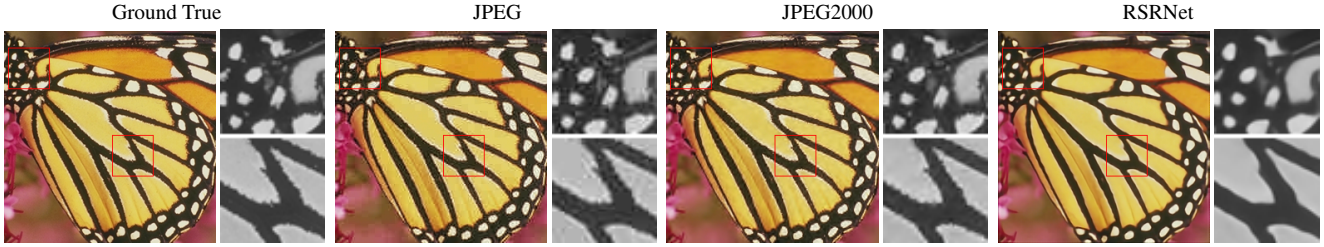


Figure 11. Subjective performance comparison between JPEG, JPEG2000 and RSRNet when the bit rate 0.74bpp. The visually disturbing blocking, aliasing, or ringing artifacts commonly seen in images compressed with JPEG, JPEG2000. For the proposed method, we see that compressed images preserves the smoothness and sharpness of many contours and edges, giving them a more natural appearance.

4.4. Comparisons with Image Compression

Image compression is a fundamental and well-studied problem in engineering, which aims to reduce irrelevance and redundancy for storage and transmission. With the decreasing of bits per pixel (bpp), high compression ratio causes the decoded image to have blocking artifacts or noises. Existing image codecs usually consists of transformation, quantization and entropy-coding. Recently deep-learning based image compression methods [31, 1] achieve competitive performance. However, they ignore the compatibility with existing image codecs, which limits their wide applications in engineering. RSRNet is compatible with existing image coding standards to improve image compression. In RSRNet, RRNet produces a compact transformation for encoding using existing codecs, and SRNet reconstructs the decoded image to avoid blocking artifacts.

To evaluate the performance of RSRNet for image compression, we conduct experimental comparisons with standard compression methods including JPEG and JPEG2000. For compression evaluation, luminance values usually is considered in YCbCr colour space. Bits for header information of compressed files were counted towards the bit rate of compared methods. Since JPEG is without lossless compression. Compared with JPEG, we used RSRNet for compression transforming and a common file compressor for quantization coding. The raw image is down-sampled by RRNet and stored as *.pgm* file, an uncompressed format. Then the *.pgm* file is coded by 7-Zip [3] with solid compression. Compared with JPEG2000, we simply adopted OpenJPEG [4] with lossless compression to code the LR image generated by RRNet.

In Table 4, we evaluate the compression ratio on Set5 with a similar distortion factor SSIM. We use RSRNet trained with scale factor $\times 3$ as the transformation. For JPEG and JPEG2000, we test the codecs at quality parameter $q = 44$ and compression ratio $r = 13\%$, respectively. The comparisons show that the proposed method significantly outperforms JPEG and JPEG2000 on bbp. To demonstrate the qualitative nature of compression artifacts, we show a representative example of a compressed image *butterfly* with bpp ≈ 0.74 in Figure 11.

5. Conclusion

For image super-resolution, we demonstrate that down-sampling will lose useful information, and the up-sampling at the first layer does not provide any extra information. To address the sampling problems, we proposed Reduced & Super-Resolution Network (RSRNet) in this paper. RSRNet is an end-to-end system without any cheap interpolation to learn mappings for resolution reduction and improvement simultaneously. Experimental results reveal that the proposed method achieves state-of-the-art results on standard benchmarks with a higher speed. Moreover, the reduced-resolution network in RSRNet can also be applied to generate photo-realistic LR image and improve image compression with existing image coding standards.

[3] http://www.7-zip.org/

[4] http://www.openjpeg.org/

References

[1] J. Ballé, V. Laparra, and E. P. Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016.

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

[3] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016.

[4] H. Chang, D.-Y. Yeung, and Y. Xiong. Super-resolution through neighbor embedding. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2004.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[6] C. De Boor. Bicubic spline interpolation. *Studies in Applied Mathematics*, 41(1-4):212–218, 1962.

[7] 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.

[8] 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.

[9] C. E. Duchon. Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology*, 18(8):1016–1022, 1979.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[11] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 349–356. IEEE, 2009.

[12] T. Goto, T. Fukuoka, F. Nagashima, S. Hirano, and M. Sakurai. Super-resolution system for 4k-hdtv. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4453–4458. IEEE, 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[14] 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.

[15] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

[16] 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.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[18] J. Kim, J. Kwon Lee, and K. Mu 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.

[19] J. Kim, J. Kwon Lee, and K. Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1645, 2016.

[20] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] X. Mao, C. Shen, and Y.-B. Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in Neural Information Processing Systems*, pages 2802–2810, 2016.

[22] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 416–423. IEEE, 2001.

[23] S. Schulter, C. Leistner, and H. Bischof. Fast and accurate image upscaling with super-resolution forests. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3799, 2015.

[24] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. 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.

[25] W. Shi, J. Caballero, C. Ledig, X. Zhuang, W. Bai, K. Bhatia, A. M. S. M. de Marvao, T. Dawes, D. ORegan, 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.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] J. Sun, Z. Xu, and H.-Y. Shum. Image super-resolution using gradient profile prior. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[29] 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*, 2017.

[30] Y. Tai, J. Yang, X. Liu, and C. Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4539–4547, 2017.

[31] L. Theis, W. Shi, A. Cunningham, and F. Huszár. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, 2017.

[32] M. Thornton, P. M. Atkinson, and D. Holland. Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. *International Journal of Remote Sensing*, 27(3):473–491, 2006.

[33] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.

[34] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.

[35] K. Zhang, X. Gao, D. Tao, and X. Li. Multi-scale dictionary for single image super-resolution. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1114–1121. IEEE, 2012.

[36] L. Zhang, H. Zhang, H. Shen, and P. Li. A super-resolution reconstruction algorithm for surveillance images. *Signal Processing*, 90(3):848–859, 2010.