https://doi.org/10.2352/ISSN.2169-2629.2018.26.75

Deep Residual Network for Joint Demosaicing and Super-Resolution

Ruofan Zhou, Radhakrishna Achanta, Sabine Süsstrunk
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Abstract

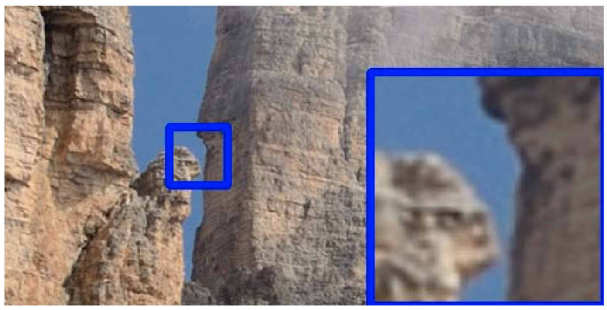
The two classic image restoration tasks, demosaicing and super-resolution, have traditionally always been studied independently. That is sub-optimal as sequential processing, demosaicing and then super-resolution, may lead to amplification of artifacts. In this paper, we show that such accumulation of errors can be easily averted by jointly performing demosaicing and super-resolution. To this end, we propose a deep residual network for learning an end-to-end mapping between Bayer images and high-resolution images. Our deep residual demosaicing and super-resolution network is able to recover high-quality super-resolved images from low-resolution Bayer mosaics in a single step without producing the artifacts common to such processing when the two operations are done separately. We perform extensive experiments to show that our deep residual network achieves demosaiced and super-resolved images that are superior to the state-of-the-art both qualitatively and quantitatively.

Introduction

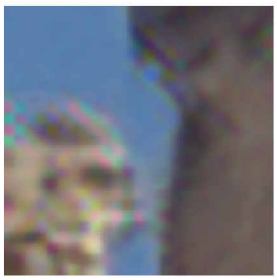
There is an evergrowing interest in capturing high-resolution images that follows the increasing quality of display devices. However, the most prevalent image capture devices are mobile phones, which are equipped with small lenses and compact sensors. Despite the large advancements made in improving the dynamic range and resolution of images captured by mobile devices, the inherent design choices limit the ability to capture very high-quality images.

The limitations result from two design decisions. Firstly, the single CMOS sensor in most of the cameras, including mobile cameras, measures at each spatial location only a limited range of wavelengths (red, green *or* blue) of the electromagnetic radiation instead of the full visible spectrum (red, green, *and* blue). This is achieved by placing a color filter array (CFA) in front of the sensor. The most common type of CFA is the Bayer pattern [1], which captures an image mosaic with twice as many green pixels as compared to red and blue pixels. Secondly, as the sensor needs to be compact to fit into the device, resolution is limited by the size of the photon wells. Small photon wells have a low well capacity, which limits the dynamic range of the image capture. Large photon wells limit the number of pixels and thus the resolution. To reconstruct full color from the CFA mosaiced image, demosaicing algorithms are applied, while low-resolution demosaiced images can only be dealt with using super-resolution algorithms in a post-processing step.

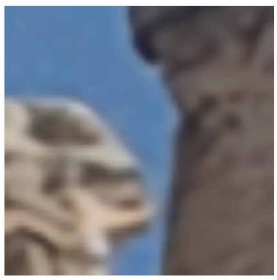
In the last decades, demosaicing and super-resolution have been independently studied and applied in sequential steps. However, the separate application of demosaicing and super-resolution is sub-optimal and usually leads to error accumulation. This is because artifacts such as color zippering introduced by demosaicing algorithms is treated as a valid signal of the input image by the super-resolution algorithms. As most of the super-resolution algorithms [3] rely on the assumption that the human visual system is more sensitive to the details in the luminance channel than the details in chroma channels, they only deal with noise in the luminance channel and neglect the artifacts in chroma channels caused by demosaicing algorithms. As a result, sequential application of super-resolution algorithm after demosaicing algorithm can lead to visually disturbing artifacts in the final output. An example is shown in Fig. 1b).



(a) Reference image from RAISE [2]



(b) FlexISP [15]+SRCNN [3] (28.2927 dB, 0.8420)



(c) Our output (30.9535 dB, 0.9118)

Figure 1: Comparison of our joint demosaicing and super-resolution output to the sequential application of demoisacing and super-resolution. The two numbers in the brackets are PSNR and SSIM respectively. Our method is able to faithfully reconstruct the original.

Although demosaicing and super-resolution have been investigated independently, it is reasonable to address them in a unified context, which is the aim of this paper. With the advent of deep learning, there are several methods for super-resolution [3, 12, 17] that successfully outperform traditional super-resolution meth-

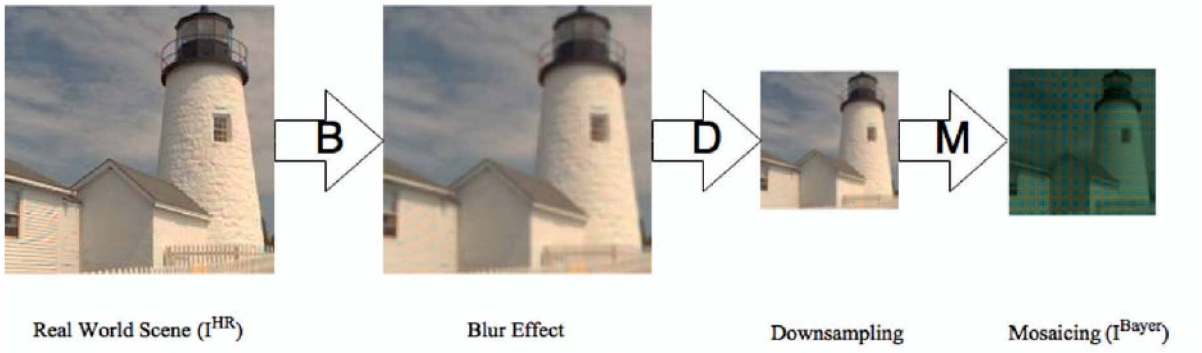


Figure 2: Diagram of the image formation in our model. I^{HR} is the intensity distribution of the real scene, B, D, M represent the blurring, downsampling, and mosaicing process, and I^{Bayer} is the observed Bayer mosaiced image. Our goal is to invert this process.

ods [4, 5, 6, 8, 18, 20]. Only recently, deep learning has also been successfully used for image demosaicing [14]. When using deep learning techniques, it is possible to address demosaicing and super-resolution simultaneously.

In this paper, we propose to use a deep residual network for joint demosaicing and super-resolution. More specifically, our network can learn an end-to-end mapping between RGGB Bayer mosaics and high-resolution color images. The main contributions of the paper are the following:

1) The first attempt to perform joint demosaicing and super-resolution on single Bayer image, to the best of our knowledge. Unlike existing super-resolution methods that usually super-resolve only the luminance channel while resorting to interpolation of the chroma channels, we directly generate full-color three channel super-resolution output. 2) Both demosaicing and super-resolution are jointly optimized through the network, therefore conventional artifacts such as moiré and zippering, which pose a post-processing challenge, are nearly eliminated. 3) We demonstrate both quantitatively and qualitatively that our approach generates higher quality results than the state-of-the-art. In addition, our method is computationally more efficient because of the joint operation.

Joint Demosaicing and Super-Resolution

A common image formation model for imaging systems is illustrated in Fig. 2. In this model, the real world scene I^{HR} is smoothed by a blur kernel representing the point spread function of the camera. It is downsampled by a factor r and mosaiced by the CFA to get the observed Bayer mosaic I_{Bayer}. Our goal is to provide an approximate inverse operation estimating a high-resolution image $I^{SR} \approx I^{HR}$ given such a low-resolution Bayer image I^{Bayer}. In general, I^{Bayer} is a real-valued tensor of size $h \times w \times 1$, I^{HR} is a tensor of $r \cdot h \times r \cdot w \times 3$. This problem is highly ill-posed as the downsampling and mosaicing are non-invertible.

To solve this problem, traditional methods usually design nonlinear filters that incorporate prior heuristics about inter-channel and intra-channel correlation. A deep CNN is a better substitute for such methods, as convolutional layers can automatically learn to exploit inter-channel and intra-channel correlation through a large dataset of training images. Moreover, the exclusive use of a set of convolutional layers enables joint optimization of all the parameters to minimize a single objective as is the case in joint demosaicing and super-resolution.

We build our framework in a data-driven fashion: we create the training set from a large set of high-quality images I^{HR}, and produce the input measurements I^{Bayer} using the same process as the image formation model illustrated in Fig. 2. We train our deep convolutional network on this dataset.

Deep Residual Network Design

We use a standard feed-forward network architecture to implement our joint demosaicing and super-resolution, which is presented in Fig. 3 and Tab. 1. The goal of the network is to recover from I^{Bayer} an image $I^{SR} = F(I^{Bayer})$ that is as similar as possible to the ground truth high-resolution color image I^{HR}. We wish to learn a mapping F from a large corpus of images, which conceptually consists of three stages:

1. **Color Extraction:** this operation separates the color pixels into different channels from the mono-channel Bayer mosaic. With this operation, no hand-crafted rearrangement of the Bayer input is needed unlike other demosaicing algorithms [15, 14]. This operation gives a set of color features from the Bayer image.
2. **Feature Extraction & Non-linear Mapping:** following the intuition of the first deep neural network for super-resolution [3], this operation extracts overlapping patches from the color features to use high-dimensional vectors to represent the Bayer image in a low-resolution feature space, which is then mapped to the high-resolution feature space.
3. **Reconstruction:** this operation aggregates high-resolution representations to generate the final high-resolution color image I^{SR}.

Color Extraction

The Bayer mosaic is a matrix with the three color samples arranged in a regular pattern in a single channel. To make the spatial pattern translation-invariant and reduce the computational cost in latter steps, it is essential to separate the colors in the Bayer image into different channels at the beginning. The Bayer pattern is regular and has a spatial size of $s \times s$, where $s = 2$. Since the neighboring colors may also affect the result, we build our first convolutional layer L_1 with a spatial size of $2 \cdot s$ and a stride of s:

$$I^1 = L_1(I^{Bayer})_{(x,y)} = (W_1 * I^{Bayer} + b_1)_{(2 \cdot x, 2 \cdot y)}, \quad (1)$$

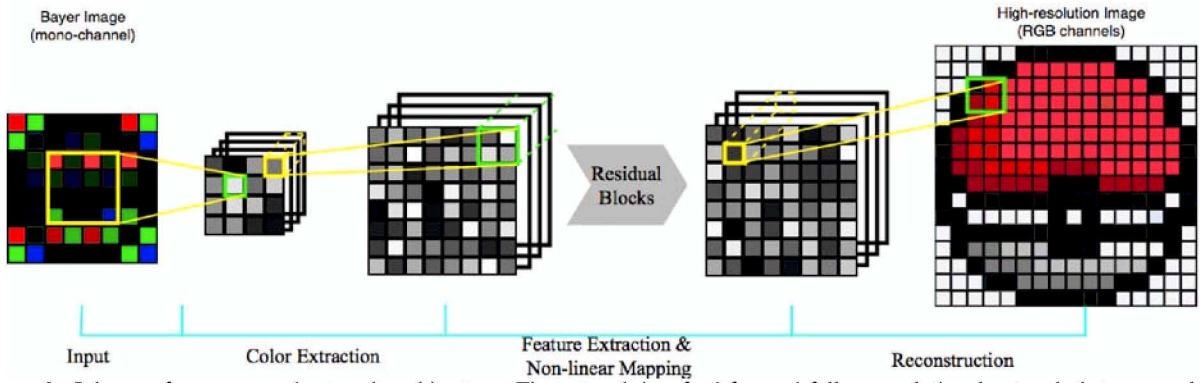


Figure 3: Scheme of our proposed network architecture. The network is a feed-forward fully convolutional network that maps a low-resolution Bayer mosaic to a high-resolution color image. Conceptually, the network has three components: color extraction, feature extraction and reconstruction.

Table 1: Summary of our network architecture. Stage number 1,2,3 of the first column correspond to the three stages of color extraction, feature extraction and non-linear mapping, and reconstruction, respectively, illustrated in Fig. 3. We set the number of filters $C = 256$ and use 24 residual blocks in Stage 2.

Stage	Layer	Output Dimension
Input (Bayer image)		$h \times w \times 1$
1	Conv with a stride of 2	$\frac{h}{2} \times \frac{w}{2} \times C$
	Sub-pixel Conv	$h \times w \times \frac{C}{4}$
	Conv, PReLU	$h \times w \times C$
2	Residual Block	$h \times w \times C$
	...	$h \times w \times C$
	Residual Block	$h \times w \times C$
3	Sub-pixel Conv	$2 \cdot h \times 2 \cdot w \times \frac{C}{4}$
	Conv, PReLU	$2 \cdot h \times 2 \cdot w \times C$
	Conv	$2 \cdot h \times 2 \cdot w \times 3$
Output (color image)		$2 \cdot h \times 2 \cdot w \times 3$

where I^1 represents the output from the first layer, W_1 and b_1 represent the filters and biases of the first convolutional layer, and $*$ denotes the convolution operation. Here, W_1 corresponds to $C = 256$ filters of support $2 \cdot s \times 2 \cdot s$.

We build an efficient sub-pixel convolutional Layer [17] L_2 to upsample the color features back to the original resolution:

$$\begin{aligned} L_2(I^1)_{(x,y,c)} &= I^1_{(x',y',c')}, \\ x' &= \left\lfloor \frac{x}{s} \right\rfloor, \\ y' &= \left\lfloor \frac{y}{s} \right\rfloor, \\ c' &= \frac{C \cdot mod(y,s)}{s} + \frac{C \cdot mod(x,s)}{s^2} + c, \end{aligned} \tag{2}$$

here, the sub-pixel convolutional layer is equivalent to a shuffling operation which reshapes a tensor of size $H \times W \times C$ into a tensor of size $s \cdot H \times s \cdot W \times \frac{C}{s^2}$. We find that applying this sub-pixel convolutional layer helps reduce checkerboard artifacts in the output.

This color extraction operation can be generalized to other CFAs by modifying s with respect to the spatial size and arrangement of the specific CFA. We set $s = 2$ for the Bayer CFAs, CYGM CFA or RGBE CFA, and $s = 6$ for the X-trans pattern.

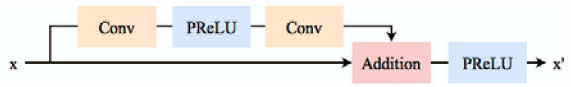


Figure 4: Illustration of the architecture of our residual blocks. We remove the batch normalization layer in the original residual blocks [7] and replace the ReLU with Parametric ReLU. This structure enables faster convergence and better performance.

Feature Extraction & Non-linear Mapping

Inspired by Dong *et al.* [3], to explore relationships within each color channel and between channels, as well as to represent the Bayer image in a high-resolution feature space, we exploit a group of convolutional layers.

Previous work [7] has demonstrated that residual networks exhibit excellent performance both in accuracy and training speed in computer vision problems ranging from low-level to high-level tasks. We build a set of 24 residual blocks, each having a similar architecture as Lim *et al.* [12], which is demonstrated in Fig. 4. We remove the batch normalization layers in the original residual blocks [7] since these layers get rid of range flexibility from networks by normalizing the features [12], while the scale of the features may be useful for image restoration problems. We also replace the activation functions ReLU with Parametric ReLU (PReLU) to prevent dead neurons and vanishing gradients caused by ReLU. For convenience, we set all residual network blocks to have the same number of filters, $C = 256$.

Reconstruction

In the reconstruction stage, we apply another sub-pixel convolutional layer [17] to upsample the extracted features to the desired resolution. This is followed by a final convolutional layer to reconstruct the high-resolution color image.

Experiments

Datasets

For training and evaluation of the network, we use the publicly available dataset RAISE [2] which provides 8,162 uncompressed raw images as well as their demosaiced counterparts in TIFF format.

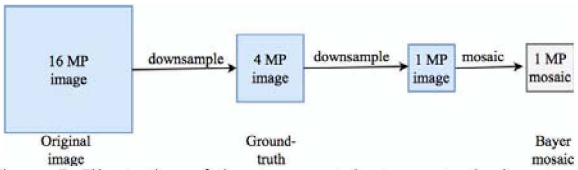


Figure 5: Illustration of the steps we take to create the input and output images of our training and testing dataset. The original 16 megapixel images are downsampled to 4 megapixel to eliminate demosaicing errors and noise. The 4 megapixel images serve as reference super-resolution images, whose downsampled 1 megapixel version provides the the single-channel Bayer CFA images used as input to our network.

Note that if we use images that are already demosaiced by a given algorithm, our network will learn to generate any artifacts introduced by the demosaicing algorithm. Moreover, we only deal with demosaicing and super-resolution, we have the assumption that other image restoration tasks such as denoising would be resolved in other steps in the image processing pipeline, thus noise should not be modeled in our image sampling pipeline. We circumvent the problem as follows. We use the demosaiced images of RAISE that are larger than 16 megapixels in size. We then perform a progressive downsizing of the image in steps by a factor of 1.25 each time until we obtain one-fourth of the original image size (*i.e.*about 4 megapixels). This is done to eliminate artifacts that have potentially been introduced by the demosaicing algorithm as well as by other factors in the camera processing pipeline, such as sensor noise. This way we obtain the high-quality ground-truth I^{HR} that serves as the super-resolved images.

To create input Bayer mosaics I^{Bayer} from these ground-truth images, we further downsample the previously downsample images to one-fourth of the size (to about 1 megapixel) also using the progressive downsizing. We follow the assumed image formation demonstrated in Fig. 2. As required for the Bayer pattern, we set the downsampling factor $r = 2$, and sample pixels from the three channels in the Bayer CFA pattern to obtain single-channel mosaiced images as low-resolution input images for training. Thus for a $H \times W \times 1$ Bayer image input, the desired color image output is of size $2 \cdot H \times 2 \cdot W \times 3$. These steps are illustrated in Fig. 5.

To train our network, we use a subset of RAISE of 6,000 images. In particular, we randomly selected 4,000 photos from the Landscape category and randomly selected 2,000 photos from other categories. We also randomly select 50 images from the rest of the RAISE dataset to build the testing set. We ensure that there is no duplicate image in the training and testing set.

Training Details

For training, we use $64 \times 64 \times 1$ sized patches from the created Bayer mosaics as input. As output images we use color image patches of size $128 \times 128 \times 3$ from the high-resolution (4 megapixel) images. We train our network with the ADAM optimizer [9] by setting the learning rate $= 1e-4$ and $\varepsilon = 10^{-8}$. We set the size of mini-batch as 16. For better convergence of the network, we halve the learning rate after every $10,000$ mini-batch updates. We use L1 loss as the loss function.

Results

Since we are not aware of any other joint demosaicing and super-resolution algorithms in the existing literature, there is no similar existing method for us to compare to. Instead, to illustrate the performance of our proposed end-to-end network, we designed experiments to simulate the conventional image processing pipeline for comparison. We compare our method with the sequential application of different state-of-the-art demosaicing algorithms (FlexISP [15], SEM [10] and DemosaicNet [14]) and the state-of-the-art super-resolution algorithm (SRCNN [3] and MDSR [12]). We use the published code from these method. As in the conventional image processing pipeline, demosaicing and super-resolution are two different components which are supposed to be resolved independently, we don't fine-tune the super-resolution algorithms on the demosaicing algorithm results.

Note that SEM [10] and DemosaicNet [14] perform joint demosaicing and denoising, for fair comparison, we set the noise-level $\sigma = 0$ for these methods. As SRCNN only provides upsampling in the luminance channel, we upsample the chroma channels using bicubic interpolation. The process is shown in Fig. 7.

Quantitative Results

In Tab. 2 we report the PSNR values of our approach in comparison to other methods on the testing dataset. Our approach outperforms the PSNR scores of the next best combination of state-of-the-art techniques of demosaicing and super-resolution by a significant PSNR difference of $1.3dB$ on average computed over the 50 images of the test-set.

Table 2: Mean PSNR and SSIM of different methods evaluated on our testing dataset. For faire comparsion with methods(*) that perform joint demosaicing and denoising, we set their noise-level to 0.

Method	PSNR	SSIM
FlexISP [15]+SRCNN [3]	29.6092 dB	0.9182
FlexISP* [15]+MDSR [12]	29.1237 dB	0.9192
SEM* [10]+SRCNN [3]	29.4978 dB	0.9348
SEM* [10]+MDSR [12]	29.3729 dB	0.9382
DemosaicNet* [14]+SRCNN [3]	30.1313 dB	0.9374
DemosaicNet* [14]+MDSR[12]	30.1177 dB	0.9291
Ours	**31.4093 dB**	**0.9476**

Qualitative Results

To further validate the quality of our results, we show qualitative comparisons in Fig. 6. Note that although MDSR [12] is a superior super-resolution algorithm to SRCNN [3], sometimes it performs worse than SRCNN, as it emphasis more strongly the artifacts produced by the demosaicing algorithms.

The combination of FlexISP [15] and SEM [10] produces some disturbing artifacts such as zippering around the edge and false color artifacts. These are particularly visible in the man's clothes (in the first column of Fig. 6) and the text (in the last column of Fig. 6).

Both DemosaicNet [14] and our network can produce demosaiced images without these artifacts, but our network is able to recover more realistic details. This is demonstrated in the second and the fourth column of Fig. 6. Our network is able to produce higher quality color images without the visually disturbing artifacts introduced by the other methods.

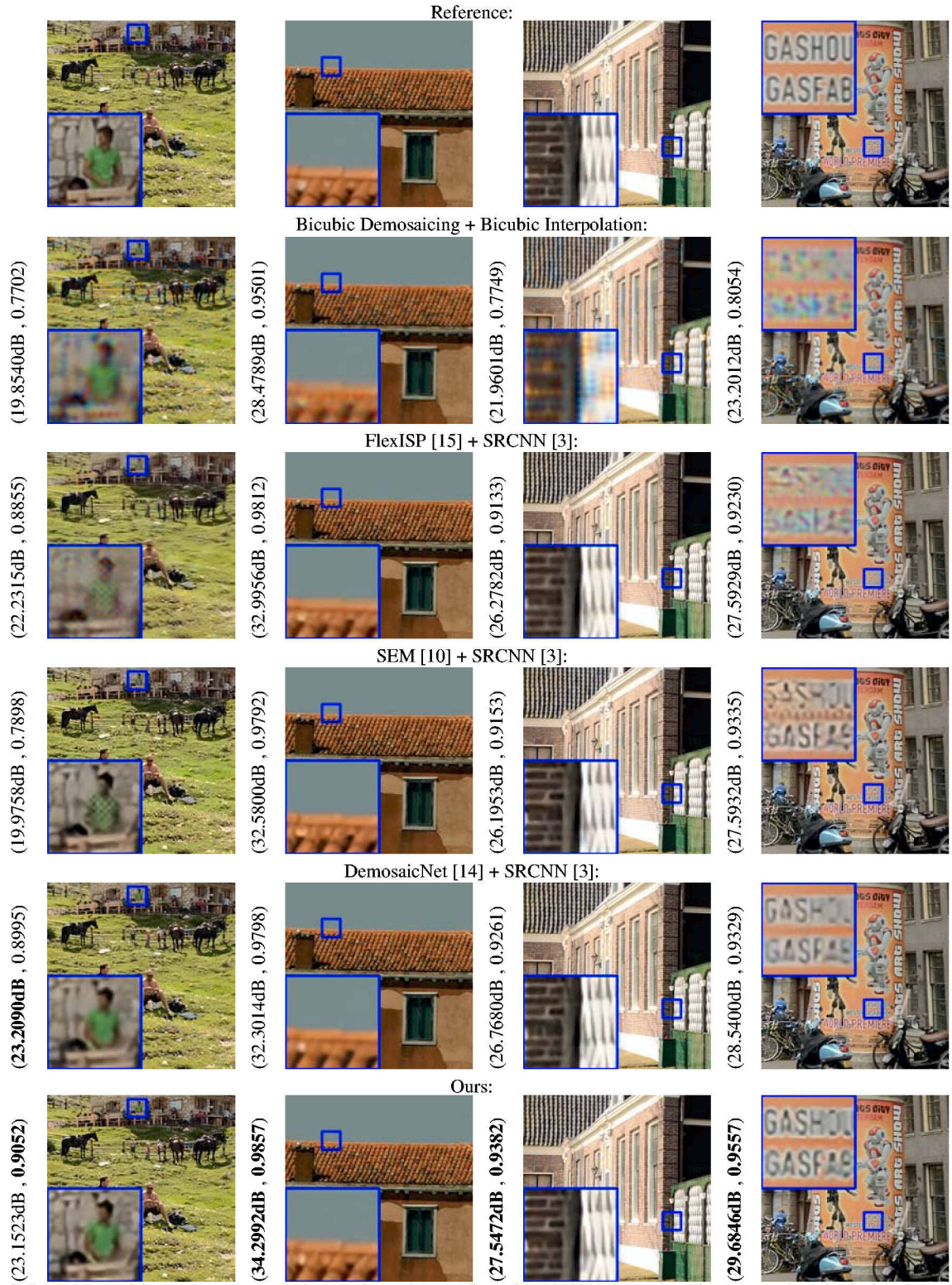


Figure 6: Joint demosaicing and super-resolution results on images from the RAISE [2] dataset. The two numbers in the brackets are the PSNR and SSIM scores, respectively.

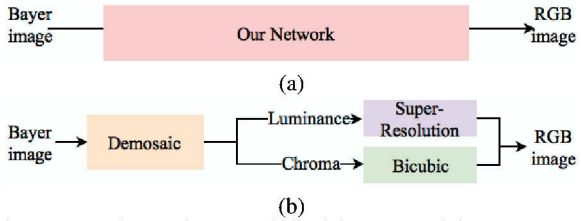


Figure 7: (a) is our framework for joint demosaicing and super-resolution, our network can perform the whole process in an end-to-end manner. (b) shows a typical pipeline to combine the demosaic algorithms and super-resolution algorithms, which we use for comparing with other algorithms. Unlike most super-resolution algorithms that output only the luminance channel, we directly generate full color output.

Running Time

We also test the running time of our method and the algorithms we compared to on 10 256×256 input images using a Nvidia TITAN X. As FlexISP and SEM reply on iterative optimization, they take more than 100,000 ms on average. While DemosaicNet takes on average 650 ms for demosaicing alone, our method has an average of 619 ms for the joint operation of demosaicing and super-resolution.

Discussion

In this paper, we propose a CNN-based framework for single image joint demosaicing and super-resolution, which is capable of directly recovering high-quality color super-resolution images from Bayer mosaics. Our proposed method outperforms all the tested combinations of the state-of-the-art demosaicing algorithms and the state-of-the-art super-resolution algorithms in both quantitative measurements of PSNR and SSIM as well as visually. Our approach does not produce disturbing color artifacts. Although these demosaicing artifacts (such as zippering artifacts in the first column of Fig. 6) may not appear in the real-world scenarios as some noise and aberrations are eliminated by the lens blur, our approach still provides the sharpest and the most realistic result compared to the other methods even when ignoring the artifacts. Our approach can be extended to videos, and can potentially be integrated into the camera imaging pipeline.

References

[1] Bryce E. Bayer, Color imaging array. US Patent 3,971,065. (1976).

[2] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, Giulia Boato, RAISE - a raw images dataset for digital image forensics. MMSys, pg. 219-224. (2015).

[3] Chao Dong, Chen Change Loy, Kaiming He, Xiaoou Tang, Learning a Deep Convolutional Network for Image Super-Resolution. ECCV, pg. 184-199. (2014).

[4] Karen Egiazarian, Vladimir Katkovnik, Single image super-resolution via BM3D sparse coding. EUSIPCO, (2015).

[5] Hong Chang, Dit-Yan Yeung, Yimin Xiong, Super-Resolution of Text Images through Neighbor Embedding. CVPR, (2012).

[6] Gilad Freedman, Raanan Fattal, Image and video upscaling from local self-examples. ACM Transactions on Graphics, 30, 12, (2011).

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep Residual Learning for Image Recognition. CVPR, (2016).

[8] Jian Sun, Zongben Xu, Heung-Yeung Shum, Image super-resolution using gradient profile prior. CVPR, (2008).

[9] Diederik P. Kingma, Jimmy Ba, Adam: A Method for Stochastic Optimization. ICLR, (2014).

[10] Teresa Klatzer, Kerstin Hammernik, Patrick Knöbelreiter, Thomas Pock, Learning joint demosaicing and denoising based on sequential energy minimization. ICCP, (2016).

[11] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, ImageNet Clarssification with Deep Convolutional Neural Networks. NIPS, (2012).

[12] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, Enhanced Deep Residual Networks for Single Image Super-Resolution. CVPR Workshops, (2017).

[13] Xiao-Jiao Mao, Chunhua Shen, Yu-Bin Yang, Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections. NIPS, pg. 2810-2818, (2016).

[14] Michael Gharbi, Gaurav Chaurasia, Sylvain Paris, Durand Frédo, Deep joint demosaicking and denoising. TOG, 39, 191 (2016).

[15] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajak, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, Jan Kautz, Kari Pulli, FlexISP - a flexible camera image processing framework. TOG, 33, 231 (2014).

[16] Samuel Schulter, Christian Leistner, Horst Bischof, Fast and accurate image upscaling with super-resolution forests. CVPR, (2015).

[17] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, Zehan Wang. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. CVPR, (2016).

[18] Radu Timofte, Vincent De Smet, Luc Van Gool, A+ : Adjusted Anchored Neighborhood Regression for Fast Super-Resolution. ACCV, (2014).

[19] Xiaolin Wu, Ning Zhang, Primary-consistent soft-decision color demosaic for digital cameras. IEEE Trans on Image Processing, pg. 1263-1274, (2004).

[20] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, Thomas Huang, Coupled Dictionary Training for Image Super-Resolution. IEEE Trans on Image Processing, pg. 3467-3478, (2012).

Author Biography

Ruofan Zhou received her BEng in Computer Science and Technology from Tsinghua University in 2015. Since then, she is a Research Assistant in Computational Photography, pursuing a PhD degree in the Image and Visual Representation Laboratory, EPFL.

Radhakrishna Achanta has a PhD in Computer Science from EPFL Switzerland, an MSc from NUS Singapore, and a BEng from JEC India. He is currently a Senior Data Scientist at the Swiss Data Science Center. During the past sixteen years he has worked in academia and industry, including start-ups. He has served as a reviewer and area chair for international conferences. His main interests are Computer Vision, Image Processing, and Machine Learning.

Sabine Süsstrunk leads the Images and Visual Representation Lab (IVRL) at EPFL, Switzerland. Her research areas are in computational photography, color computer vision and color image processing, image quality, and computational aesthetics. She has published over 150 scientific papers, of which 7 have received best paper/demos awards, and holds 10 patents. She received the IS&T/SPIE 2013 Electronic Imaging Scientist of the Year Award and IS&Ts 2018 Raymond C. Bowman Award. She is a Fellow of IEEE and IS&T.