Camera Lens Super-Resolution

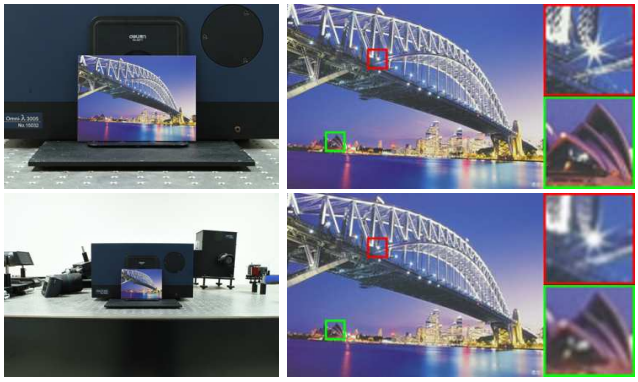
Chang Chen Zhiwei Xiong* Xinmei Tian Zheng-Jun Zha Feng Wu
University of Science and Technology of China

Abstract

Existing methods for single image super-resolution (SR) are typically evaluated with synthetic degradation models such as bicubic or Gaussian downsampling. In this paper, we investigate SR from the perspective of camera lenses, named as CameraSR, which aims to alleviate the intrinsic tradeoff between resolution (R) and field-of-view (V) in realistic imaging systems. Specifically, we view the R-V degradation as a latent model in the SR process and learn to reverse it with realistic low- and high-resolution image pairs. To obtain the paired images, we propose two novel data acquisition strategies for two representative imaging systems (i.e., DSLR and smartphone cameras), respectively. Based on the obtained City100 dataset, we quantitatively analyze the performance of commonly-used synthetic degradation models, and demonstrate the superiority of CameraSR as a practical solution to boost the performance of existing SR methods. Moreover, CameraSR can be readily generalized to different content and devices, which serves as an advanced digital zoom tool in realistic imaging systems.

1. Introduction

Single image super-resolution (SR) is a typical inverse problem in computer vision. Generally, SR methods assume bicubic or Gaussian downsampling as the degradation model [33]. Based on this assumption, continuous progress has been achieved to restore a better high-resolution (HR) image from its low-resolution (LR) version, in terms of reconstruction accuracy [9, 13, 15, 17, 23, 25, 27, 31, 32, 35, 36] or perceptual quality [2, 3, 5, 12, 16, 22, 28]. However, these synthetic degradation models may deviate from the ones in realistic imaging systems, which results in a significant deterioration on the SR performance [20]. To better simulate the challenging real-world conditions, additional factors including noise, motion blur, and compression artifacts are integrated to characterize the LR images in either a synthetic [26] or a data-driven [4] manner. These modified degradation models promote the SR performance of learning-based methods when the LR images indeed have corresponding degradations.



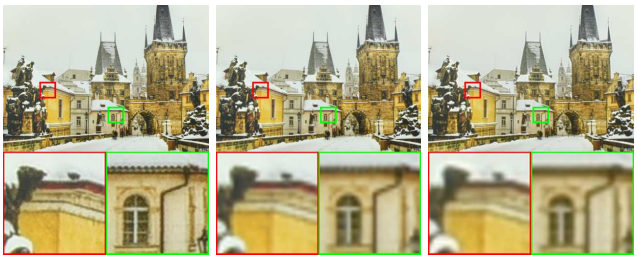
(a) (b)

Figure 1. (a) Resolution-FoV (R-V) degradation. Zooming out the optical lens in a DSLR camera, the FoV is enlarged at the cost of resolution loss. (b) Aligned realistic LR-HR image pair after rectification. LR image is displayed after interpolation for a side-by-side comparison. (Bicubic interpolation is used throughout this paper unless noted otherwise.)

In this paper, we investigate SR from the perspective of camera lenses, named as CameraSR, which aims to alleviate the intrinsic tradeoff between resolution (R) and field-of-view (FoV, V for short hereafter) in realistic imaging systems. An instance of the R-V tradoff is shown in Fig. 1(a). When zooming out the optical lens in a DSLR camera, the obtained image has a larger FoV but loses details on subjects; when zooming in the lens, the details of subjects show up at the cost of a reduced FoV. This R-V tradeoff also applies to cameras with fixed focal lenses such as those on smartphones when the shooting distance changes. Inspired by learning-based single image SR, we view the *R-V degradation* (i.e., resolution loss due to enlarged FoV) as a latent model in the SR process and learn to reverse it with a number of LR-HR image pairs. Specifically, we define a subject captured at a long focal length or a short distance as the HR ground truth, and the same one captured at a short focal length or a long distance as its paired LR observation.

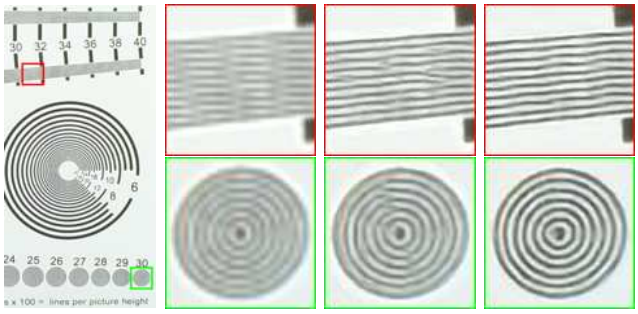
To obtain such paired images, we first use a DSLR camera mounted on a tripod with a zoom lens. To avoid the out-of-focus blur, we adopt a small aperture size and capture

*Correspondence should be addressed to zwxiong@ustc.edu.cn



(a) HR Ground Truth PSNR | (b) Bicubic Downsampling 28.56 dB | (c) R-V Degradation 27.22 dB

Figure 2. Visual comparison between the LR image with bicubic downsampling and the realistic LR image with R-V degradation (both are displayed after interpolation). The latter loses more information than the former in visual compared with the HR ground truth, which is also quantitatively verified by PSNR.



(a) (b) (c) (d)

Figure 3. An example to show the performance deterioration due to improper degradation modeling (bicubic downsampling here). (a) An image captured by a DSLR camera. (b) Interpolated result. (c) SR result using VDSR [13] trained under bicubic downsampling. (d) SR result using VDSR trained under R-V degradation.

100 city scenes printed on postcards as the subjects which can be well focused at different focal lengths. In practice, however, several issues due to the mechanical zoom prohibit the direct use of the captured raw data, including spatial misalignment, intensity variation, and color mismatching. After addressing these issues through an elaborate data rectification pipeline, we build a dataset consisting of 100 aligned image pairs named "City100". An example is shown in Fig. 1(b). Following the same pipeline, we then obtain a variant of City100 by using a smartphone camera mounted on a translation stage with a fixed focal lens. The City100 dataset, together with its smartphone version, characterizes the R-V degradation in two representative realistic imaging systems.

Based on City100, we conduct a quantitative analysis on the commonly-used synthetic degradation models, in terms of both LR observations and SR results. Take the bicubic downsampling as an example, due to the underestimation of R-V degradation (as shown in Fig. 2), it results in a significant deterioration on the SR performance (as shown in Fig. 3). This analysis validates the importance of degradation modeling for the resolution enhancement in realistic imaging systems. Observing the disadvantage of synthetic degradation models, we propose CameraSR as a practical solution to boost the performance of existing SR methods, by learning the R-V degradation from City100. Comprehensive experiments demonstrate that CameraSR achieves a significant improvement of SR results compared with those using synthetic degradation models.

More importantly, we demonstrate that CamereSR has a favorable capability of generalization in terms of both content and device. Specifically, an SR network trained on City100 can be readily generalized to other scene content, as well as to other type of devices belonging to the same category of imaging systems (e.g., from Nikon to other DSLRs and from iPhone to other smartphones). By effectively alleviating the R-V tradeoff or even breaking the physical zoom ratio of an optical lens in realistic imaging systems, CameraSR could find a wide application in practice as an advanced digital zoom tool.

Contributions of this paper are summarized as follows:

- A new perspective, i.e., R-V degradation of camera lenses, for SR modeling in realistic imaging systems.
- Two novel strategies for acquiring LR-HR image pairs as in City100 to characterize the R-V degradation under DSLR and smartphone cameras, respectively.
- Quantitative analysis on the commonly-used synthetic degradation models using realistic data.
- An effective solution, i.e., CameraSR, to promote existing learning-based SR methods in realistic imaging systems.

2. Related Work

Recent years have seen a remarkable improvement in single image SR. To promote the reconstruction accuracy, increasingly more learning-based methods adopt the convolutional neural network (CNN) following the seminal work of SRCNN [6]. For instance, Kim *et al*. proposed VDSR [13] which deepens the network for accuracy with the residual learning. Lai *et al*. proposed LapSRN [15] which improves the SR results at large scale factors with the Laplacian pyramid structure. Furthermore, various mechanisms have been integrated into the network design to advance the SR performance, such as sparsity [30], contiguous memory [36], deep supervision [27], recursion [14, 25], back-projection [9], information distillation [10], and attention [35]. Different from the above methods, Ledig *et al*. proposed SRGAN [16] which is optimized for perceptual quality instead of reconstruction accuracy. Along this line, Sajjadi *et al*. proposed EnhanceNet [22] which promotes the quality of texture synthesis with a perceptual loss. Wang *et*

al. proposed SFTGAN [28] which integrates a spatial feature transform layer into GAN [8] to further enhance the SR performance. However, most existing learning-based methods adopt a synthetic degradation model (e.g., bicubic or Gaussian downsampling) when formulating the SR problem, which hinders their performance in realistic imaging systems with much more complicated degradation.

There are a few works that involve realistic degradation modeling for single image SR. For instance, Timofte *et al*. introduced more degradation operators into the bicubic-downsampled LR images, including motion blur and Poisson noise [26]. Bulat *et al*. defined the LR face images with the low-quality assumptions (e.g., noise, blur, and compression artifacts) and trained a GAN [8] to learn the degradation process [4]. On the other hand, as a self-similarity based method, Michaeli and Irani adaptively estimated the degradation model relying on the inherent recurrence of the input image [20]. Shocher *et al*. further optimized an image-specific CNN with examples solely extracted from the input image [23].

Different from the above approaches, our proposed CameraSR models the R-V degradation from the perspective of camera lenses. The estimation of R-V degradation neither relies on the low-quality assumptions nor the inherent recurrence of LR images. Instead, it is characterized by the samples captured with realistic imaging systems. Such a degradation modeling is inspired by the prior work for realistic image denoising [21], where a subject captured at a high ISO value is defined noisy and the same one captured at a low ISO value is defined clean. We extend this definition to the SR scenario, which addresses the key challenge of obtaining realistic LR-HR image pairs. Note that the focus of this paper is not the network design. For the comparison purpose, we adopt VDSR [13] and SRGAN [16] as two representative embodiments to demonstrate the effectiveness and generalizability of CameraSR, which can be replaced with any CNN-based methods.

3. Problem Formulation

Consider again taking photos using a DSLR camera with an optical zoom lens. Zooming out the lens derives a larger FoV at the cost of resolution loss on the subject. Denote this R-V degradation as $D_{RV}(\cdot)$, our goal is to obtain a function $S(\cdot)$ that reverses $D_{RV}(\cdot)$ for realistic image SR. This problem can be formulated as

$$\hat{X} = S(D_{RV}(X)), \tag{1}$$

where X denotes the original image and $\hat{X}$ denotes the super-resolved one. Compared with previous SR formulations, the only difference lies in the modeling for the degradation process. For instance, the bicubic downsampling $D_{Bic}(\cdot)$ formulates the SR problem as $\hat{X} = S(D_{Bic}(X))$ and the Gaussian downsampling as $\hat{X} = S(D_{Gau}(X))$. For the more complicated degradation model imposed in [26], it is $\hat{X} = S(D_{Blur}(D_{Bic}(X)) + v)$, where $D_{Blur}(\cdot)$ denotes a blurring operator and v denotes a certain kind of noise.

Unlike the synthetic degradation models as mentioned above, it is difficult to derive an analytic expression for $D_{RV}(\cdot)$. Inspired by learning-based SR, we view the R-V degradation as a latent model $\hat{D}_{RV}(\cdot)$ in the SR process and directly learn the parametric SR function $S_\Theta(\cdot)$ with N pairs of realistic LR ($\mathbf{Y} = \{Y_1, Y_2, ..., Y_N\}$) and HR ($\mathbf{X} = \{X_1, X_2, ..., X_N\}$) samples, which can be represented as

$$\hat{X} = S_\Theta(\hat{D}_{RV}(X)), \tag{2}$$

where $\hat{D}_{RV}(\cdot)$ is subject to $\mathbf{Y} = \hat{D}_{RV}(\mathbf{X})$. With the increase of the number of samples N, we have $\hat{D}_{RV}(\cdot) \rightarrow D_{RV}(\cdot)$. Then, $S_\Theta(\cdot)$ can be optimized with a loss function $\mathcal{L}(\cdot)$ as

$$\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(X_i - S_\Theta(Y_i)), \tag{3}$$

where Θ denotes a set of trainable parameters and n denotes the size of mini-batch when optimizing Θ with the stochastic gradient descent algorithm.

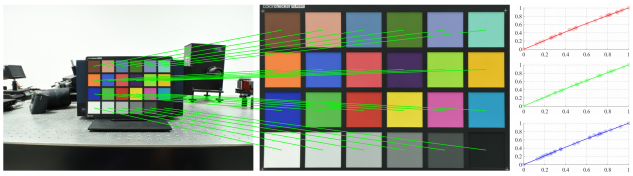
This is the main idea of CameraSR, which will be detailed in Sec. 5.2. While the problem formulation is quite intuitive, the key challenge is, how to obtain the LR-HR image pairs in realistic imaging systems?

4. Data Acquisition

4.1. DSLR imaging system

To capture the realistic LR-HR image pairs, we use a Nikon D5500 camera mounted on a tripod with a Zoom-Nikkor lens, whose focal length ranges from 18mm to 55mm. We define an image captured at 55mm focal length as the HR ground truth and the one captured at 18mm focal length as the LR observation. To alleviate the influence of noise, the ISO value is set to the lowest level. The other settings such as white balance and aperture size are fixed for each capture. In practice, however, we observe several issues for prohibiting the direct use of the captured raw data, including spatial misalignment, intensity variation, and color mismatching. It is probably due to the fact that the change of focal length is a mechanical process which cannot be ideally controlled. It thus results in slight dithering of the camera body as well as the exposure configuration. To address these issues, we elaborate a data rectification pipeline.

First, we model the spatial misalignment as a global 2D translation inspired by [11]. Specifically, we compute and match SIFT key-points [18] between the HR images and the interpolated LR ones. Then, the matched coordinates are used to estimate a homography using RANSAC [7].



(a) (b) (c)

Figure 4. Color calibration. The mean values obtained from each color blocks are adopted to fit three polynomial curves (c) for color calibration, from the LR observation (a) to its HR ground truth (b).

Figure 5. Acquisition strategy for the smartphone version of City 100. Translating the smartphone away from a subject (from A to B), the effective resolution is decreased due to the enlarged FoV (R-V degradation).

Having the translation parameters, we shift the LR images through interpolation to obtain the aligned results. Note that the interpolation will introduce some smoothing effects, but not critical for the already interpolated LR images which contain relatively fewer high frequencies. We avoid shifting the HR images since they contain a lot of desired details. Second, we model the intensity variation as a bias in the DC component of an image and estimate it by averaging the pixel intensities in the whole image. Then, we use the estimated bias to compensate this variation. Third, we model the color mismatching as a parametric non-linear mapping and fit it with polynomial parameters for calibration by leveraging a color checkerboard, as shown in Fig. 4. Specifically, we collect and average pixel values in each block from the color checkerboard to obtain paired samples from the LR observation to its HR ground truth. Then, we fit three polynomial curves for R, G, and B channels using the collected samples, respectively. Finally, we map pixels in LR observations using the obtained polynomial curves.

After the above data rectification, we build a City100 dataset using the DSLR camera, in which 100 city scenes printed on high-quality postcards are adopted as the subjects. The plane shape of postcards guarantees that the whole image can be well focused under a small aperture size at both long and short focal lengths, which avoids the out-of-focus blur. The resolution of final HR images in City100 is 1218×870, which is 2.9 times of the LR ones. Images from City100 have diverse colors and contents, which facilitate leaning-based SR. An overview of the City100 dataset is shown in the supplementary document.

4.2. Smartphone imaging system

Different from the zoom lenses in professional DSLR cameras, commodity smartphone cameras are generally equipped with prime lenses whose focal length cannot change. In this sense, the realistic degradation modeling is even more meaningful to smartphones, where CameraSR can serve as a powerful digital zoom tool. However, limited by the fixed focal lens, LR-HR image pairs for smartphone cameras cannot be captured with the same strategy as for DSLR cameras. Alternatively, we develop another strategy for obtaining the smartphone version of City100, as shown in Fig. 5. An iPhone X mounted on a translation stage is used for data acquisition, and the position of iPhone relative to the translation stage can be precisely adjusted. We define an image captured at a short distance as the HR ground truth, and the one captured at a long distance as the LR observation. To avoid the "intelligent" exposure configuration by the smartphone itself, we use the ProCam[1] software to manually control the settings such as ISO, white balance, exposure time and so on. The data rectification pipeline for smartphone is similar to that for DSLR as detailed in Sec. 4.1. In addition, considering that smartphone images have notably heavier noise than DSLR images due to the much smaller sensor size, we repeat the capture of each scene 20 times and average the resulting images to alleviate the influence of noise. The resolution of final HR images is 2.4 times of the LR ones.

It is worth mentioning that, the City100 dataset and its smartphone version are obtained by two representative realistic imaging systems, i.e., DSLR and smartphone. Although two specific devices, i.e., Nikon D5500 and iPhone X are used here, the trained CameraSR network has a favorable capability of generalization and can be readily applied to different devices belonging to the same category of imaging systems (as detailed in Sec. 6.2).

5. Analysis on Degradation Models

In this section, our goal is to quantitatively analyze the performance of commonly used synthetic degradation models $D_{Bic}(\cdot)$ and $D_{Gau}(\cdot)$, in comparison with the realistic R-V degradation $D_{RV}(\cdot)$ based on the paired samples from our developed City100 dataset. Since $D_{RV}(\cdot)$ has not an analytic expression, it is difficult to conduct direct comparisons between them. Thus, we turn to the corresponding LR observations and SR results for quantitative comparisons.

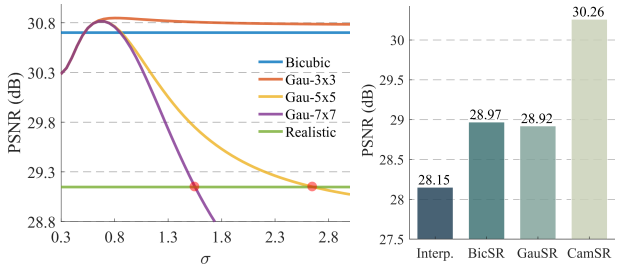
5.1. LR observation

Given an HR image X from City100, the LR observations are obtained by $D_{Bic}(X)$, $D_{Gau}(X)$, and $D_{RV}(X)$ (i.e., the paired Y from City100), respectively. As demonstrated in Fig. 2, $D_{Bic}(\cdot)$ underestimates the degradation

[1] https://www.procamapp.com

Test image	Interpolated LR	BicubicSR	GaussianSR	CameraSR
	PSNR / SSIM / Ma's / VGG	PSNR / SSIM / Ma's / VGG	PSNR / SSIM / Ma's / VGG	PSNR / SSIM / Ma's / VGG
St. Petersb.	28.74 / 0.8630 / 3.58 / 0.8543	29.69 / 0.8874 / 5.05 / 0.7756	29.61 / 0.8934 / 6.16 / 0.7019	31.00 / 0.9116 / 6.58 / 0.4791
Dubai	30.21 / 0.8443 / 3.37 / 0.5650	30.91 / 0.8599 / 4.73 / 0.4193	30.71 / 0.8603 / 5.86 / 0.3856	31.94 / 0.8788 / 6.74 / 0.3390
Venice	26.52 / 0.7317 / 3.58 / 0.9654	27.25 / 0.7686 / 4.43 / 0.8254	27.21 / 0.7813 / 5.93 / 0.7798	28.19 / 0.8062 / 6.71 / 0.6167
Rome	30.65 / 0.8654 / 3.60 / 0.3825	31.45 / 0.8806 / 4.77 / 0.3625	30.99 / 0.8768 / 6.17 / 0.3525	33.04 / 0.9039 / 6.68 / 0.2891
New York	24.62 / 0.7520 / 3.83 / 1.1808	25.55 / 0.7921 / 4.85 / 1.1528	26.06 / 0.8113 / 5.85 / 1.1345	27.14 / 0.8416 / 6.76 / 0.8381
Average	28.15 / 0.8113 / 3.59 / 0.7896	28.97 / 0.8377 / 4.77 / 0.7071	28.92 / 0.8446 / 5.99 / 0.6709	30.26 / 0.8684 / 6.69 / 0.5124

Table 1. Quantitative results of SR on the five test images from City100 (as shown in Fig. 7). PSNR and SSIM [29] (the higher, the better) are adopted for the evaluation of reconstruction accuracy (VDSR [13] network). Ma's metric [19] (the higher, the better) and the VGG metric (the lower, the better) are adopted for the evaluation of perceptual quality (SRGAN [16] network). We denote the Euclidean distance between SR results and ground truth in the feature space of a trained VGG-19 [24] network as the VGG metric ($\times 10e^4$) [34].



(a) (b)

Figure 6. Analysis on the synthetic degradation models. (a) Investigation on the LR observations from City100. The PSNR is calculated between an interpolated LR image and its HR ground truth. (b) Investigation on the SR results from test set (as shown in Fig. 7). VDSR [13] is adopted as a representative network for BicubicSR, GaussianSR, and CameraSR. Although the Gaussian downsampling matches the degradation level of the realistic LR observation at the red points, the reconstruction accuracy of GaussianSR still has a gap compared with CameraSR. It reveals the disadvantage of synthetic degradation models.

level of $D_{RV}(\cdot)$, which results in a significant deterioration on the SR performance as shown in Fig. 3.

Besides $D_{Bic}(\cdot)$, we further investigate $D_{Gau}(\cdot)$. In practice, the Gaussian downsampling first blurs X with a Gaussian filter and then conducts pixel decimation at designated scale factors. To match the scale factor of samples from City100, we adapt $D_{Gau}(\cdot)$ for $\times 2.9$ downsampling by first interpolating an image X 3/2.9 times followed by a $\times 3$ decimation. In contrast to the bicubic downsampling, the Gaussian downsampling is more flexible as its kernel size $k \times k$ and standard deviation σ can be manually controlled. Here, we consider an ideal condition when the degradation level of $D_{Gau}(X)$ matches $D_{RV}(X)$ in terms of the LR observation. To this end, we traverse k and σ as shown in Fig. 6(a). After interpolating $D_{Gau}(X)$ and $D_{RV}(X)$ to the same resolution as X, we calculate the mean PSNR between them on City100 and find two matched parameters at the red points (with $k_1 = 5$, $\sigma_1 = 2.65$ and $k_2 = 7$, $\sigma_2 = 1.55$), which are adopted as the representatives of $D_{Gau}(\cdot)$.

Figure 7. Thumbnails of the five test images from City100.

5.2. SR result

Obtained the LR observations, we then evaluate the performance of different degradation models on the SR results, by comparing $S(D_{Bic}(X))$, $S(D_{Gau}(X))$, and $S(D_{RV}(X))$ to the ground truth X. We name the corresponding SR processes as BicubicSR, GaussianSR, and CameraSR for short, respectively. To train an SR network, we split City100 into two parts: 5 selected pairs for test (as shown in Fig. 7) and the other 95 pairs for training. Among the training set, 5 images are used for validation. For the baseline network, we adopt two representative CNN architectures considering the perception-distortion tradeoff reported in [1]. For reconstruction accuracy, we adopt the VDSR network [13] with a mean square loss

$$\mathcal{L}_{MSE} = ||S_\Theta(D(x)) - x||_2^2, \tag{4}$$

where x denotes an image patch cropped from X on City100, $D(\cdot)$ denotes the a certain degradation model, and $S_\Theta(\cdot)$ denotes the parametric SR network.

For perceptual quality, we adopt the SRGAN network [16] with a combined loss

$$\mathcal{L}_{Comb} = \mathcal{L}_{MSE} + \mathcal{L}_{VGG} + 10e^{-3}\mathcal{L}_{Gen}, \tag{5}$$

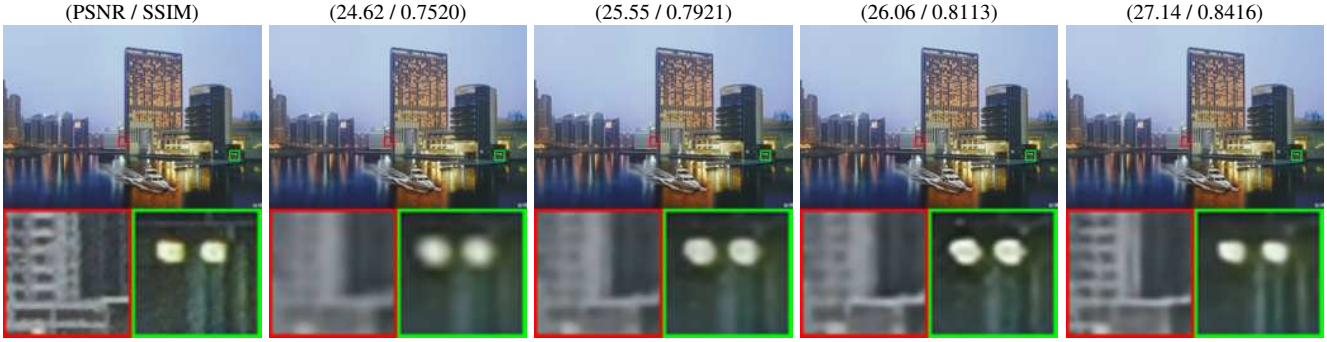
where the VGG loss $\mathcal{L}_{VGG}$ represents the pixel-wise distance in the feature space $\phi(\cdot)$ of a VGG-19 network [24]

$$\mathcal{L}_{VGG} = ||\phi(S_\Theta(D(x))) - \phi(x)||_2^2, \tag{6}$$

and the generative loss $\mathcal{L}_{Gen}$ is defined based on the probability of a discriminator $\mathcal{D}_{\Theta'}(\cdot)$ as

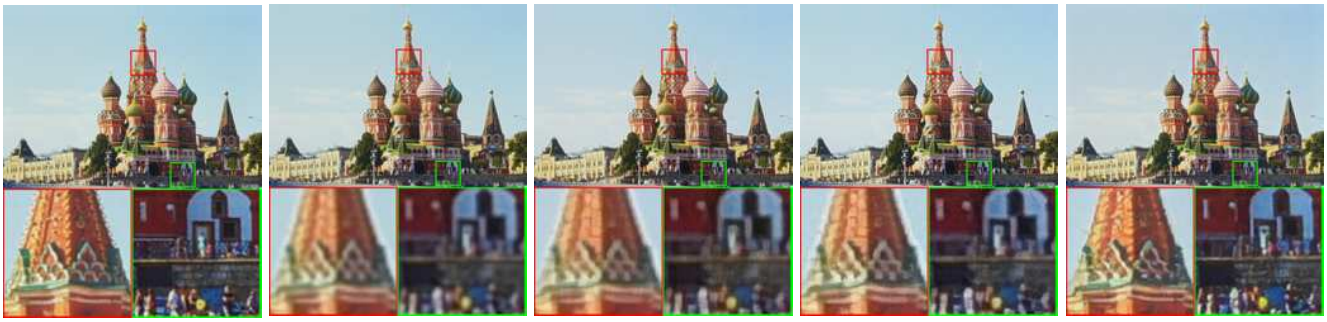
$$\mathcal{L}_{Gen} = -log\mathcal{D}_{\Theta'}(S_\Theta(D(x))), \tag{7}$$

Ground Truth HR (PSNR / SSIM) | Interpolated LR (24.62 / 0.7520) | BicubicSR (25.55 / 0.7921) | GaussianSR (26.06 / 0.8113) | CameraSR (27.14 / 0.8416)



Ground Truth HR (PSNR / SSIM) | Interpolated LR (28.15 / 0.8113) | BicubicSR (28.97 / 0.8377) | GaussianSR (28.92 / 0.8446) | CameraSR (30.26 / 0.8684)

Figure 8. Visual comparison of SR results under different degradation models in terms of reconstruction accuracy (VDSR [13] network). PSNR and SSIM [29] (the higher, the better) are adopted for evaluation metrics.



Ground Truth HR (VGG / Ma's metric) | Interpolated LR (0.8543 / 3.58) | BicubicSR (0.7756 / 5.05) | GaussianSR (0.7019 / 6.16) | CameraSR (0.4791 / 6.58)

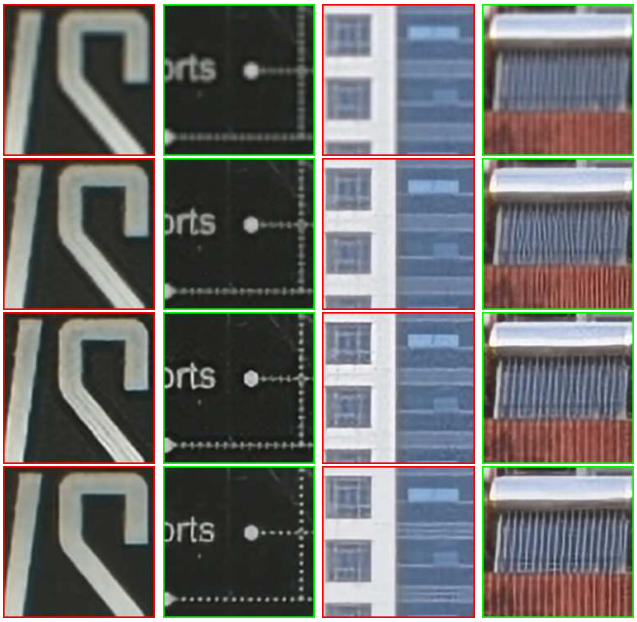
Ground Truth HR (VGG / Ma's metric) | Interpolated LR (0.9654 / 3.58) | BicubicSR (0.8254 / 4.43) | GaussianSR (0.7798 / 5.93) | CameraSR (0.6167 / 6.71)

Figure 9. Visual comparison of SR results under different degradation models in terms of perceptual quality (SRGAN [16] network). The VGG metric [24] (the lower, the better) and the Ma's metric [19] (the higher, the better) are adopted for evaluation.

where $\mathcal{D}_{\Theta'}(\cdot)$ denotes the probability that a reconstructed image $S_{\Theta}(D(X))$ is a natural one. The generative component $S_{\Theta}(\cdot)$ and the discriminator $\mathcal{D}_{\Theta'}(\cdot)$ are trained in an adversarial manner [8].

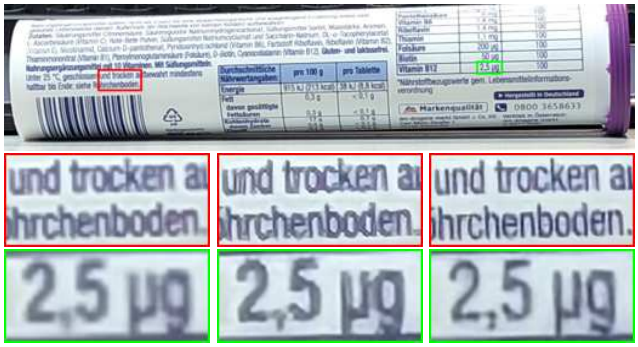
(a) Captured at 18mm focal length

(b) Captured at 55mm focal length

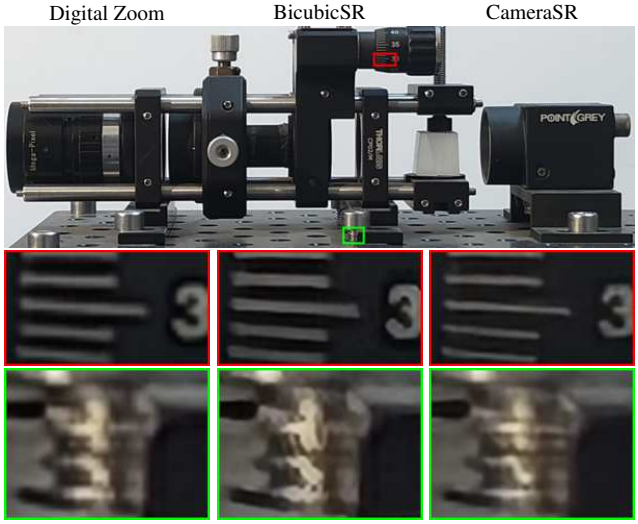


(c) From top to bottom: Interp. LR, BicubicSR, GaussianSR, CameraSR

Figure 10. Visual comparison of SR results on images captured by Nikon D5500. SR models are trained on City100 using the VDSR [13] network.



Digital Zoom BicubicSR CameraSR



Digital Zoom BicubicSR CameraSR

Figure 11. Visual comparison of SR results on images captured by iPhone X. SR models are trained on the smartphone version of City100 using the VDSR [13] network.

Then, we train two sets of SR networks for $D_{Bic}(\cdot)$, $D_{Gau}(\cdot)$, and $D_{RV}(\cdot)$ based on City100, respectively. All other hyper-parameters except the degradation model are kept the same to eliminate the influence of them. The quantitative results evaluated on PSNR are shown in Fig. 6(b), where both BicubicSR and GaussianSR have a notable performance gap (i.e., about 1.3 dB in average on the test set) compared with CameraSR. For GaussianSR, we evaluate two settings at the red points in Fig. 6(a) and report the better one. Detailed quantitative results are listed in Table 1. The corresponding visual comparisons are conducted in Figs. 8 and 9 for VDSR [13] and SRGAN [16] respectively, which again validates the significantly improved SR results achieved by CameraSR. More results for comparison can be found in the supplementary document.

6. Experiments

While the above analysis clearly demonstrates the importance of degradation modeling for the resolution enhancement of realistic imaging systems, it is not so surprising that CameraSR outperforms BicubicSR and GaussianSR since it directly learns the R-V degradation from City100. In this section, we show extensive SR results to demonstrate the generalizability of CameraSR (still trained on City100) to real-world scenes that are drastically different from City100 in content and even captured with different devices. Still, BicubicSR and GaussianSR are adopted for comparisons, in terms of reconstruction accuracy and perceptual quality.

6.1. Advanced digital zoom

Recall that our main goal is to alleviate the R-V tradeoff or even break the physical zoom ratio of an optical lens in realistic imaging systems, we now demonstrate that CameraSR achieves this goal. As shown in Fig. 10(a), given an image captured by a DSLR camera at the focal length of 18mm, CameraSR effectively super-resolves its details, which can be viewed as alleviating the R-V tradeoff of the camera lens (i.e., resolution and FoV are now obtained at the same time). Meanwhile, when the zoom lens of the same DSLR camera reaches its maximum magnification at the focal length of 55mm, CameraSR is capable of further enhancing the resolution of the captured image, as shown in

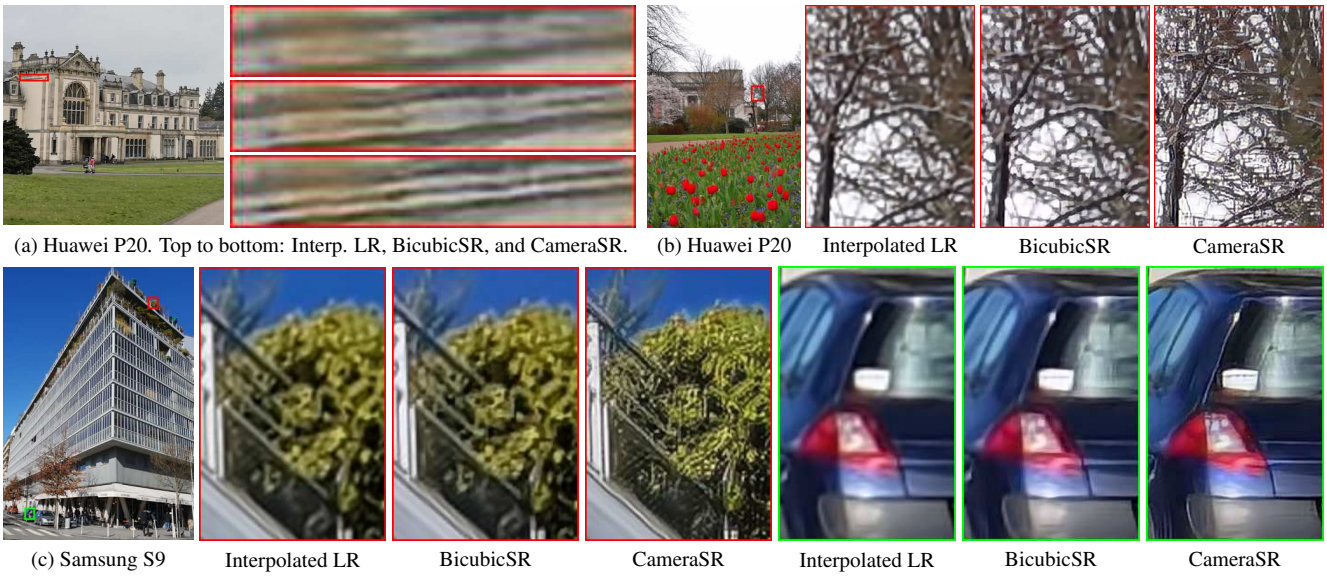


Figure 12. Visual comparison of SR results on images captured by Huawei P20 and Samsung S9 smartphone cameras. SR models are trained on the iPhone X version of City100 using the VDSR [13] network for (a) and SRGAN [16] network for (b) and (c), respectively.

Fig. 10(b). Similarly in Fig. 11, for a smartphone camera with a fixed focal lens, CameraSR serves as an advanced digital zoom tool, which significantly enhances the imaging quality compared with the built-in digital zoom function. The examples in Fig. 10(b) and Fig. 11 can be viewed as breaking the physical limit of zoom ratio.

6.2. Generalizability

Besides the significant improvement of SR performance, our proposed CameraSR also has a favorable generalization capability in terms of both content and device. For the content generalization, recall that the City100 dataset is captured under an indoor environment with a single category of subjects (i.e., postcard), yet the CameraSR model trained on City100 performs well in both indoor and outdoor environments with diverse subjects, as demonstrated in Figs. 10, 11, 12. For the device generalization, as shown in Fig. 12, the CameraSR model trained on the iPhone X version of City100 can be readily applied to different smartphones such as Huawei P20 and Samsung S9. More results for the generalization from Nikon to Canon DSLR cameras are shown in the supplementary document.

7. Conclusion and Discussion

In this paper, we investigate SR from the perspective of camera lenses, named as CameraSR, which models the R-V degradation in realistic imaging systems. With the proposed data acquisition strategies, we build a City100 dataset to characterize the R-V degradation in representative DSLR and smartphone cameras. Based on City100, we analyze the disadvantage of the commonly used synthetic degradation models and validate CameraSR as a practical solution to boost the performance of existing SR methods. Due to its favorable generalization capability, CameraSR could find a wide application as an advanced digital zoom tool in realistic imaging systems. Especially, besides the enhancement of natural images, we believe CameraSR has a great value for biomedical imaging with microscopes, where the resolution enhancement is essential for scientific observation.

Despite the promising preliminary results, there are still some real-world conditions that have not been considered in this paper. In terms of the LR observation, we consider a relatively ideal condition without noise. Yet the influence of noise is inevitable, especially in the smartphone imaging systems with small sensors. It is thus worth to jointly investigate the R-V degradation and noise to further promote the robustness of CameraSR. Besides single image SR discussed in this paper, the R-V degradation can be generalized to burst image SR, where a sequence of LR images are captured using the burst shooting mode to exploit the underlying information from the sub-pixel motion for a better HR reconstruction. Moreover, beyond the prior learned from external examples, the proposed CameraSR can be further extended for self-similarity based methods to utilize the inherent recurrence, by numerically estimating the R-V degradation kernel based on City100. The above extensions are considered as our future work.

Acknowledgement

We acknowledge funding from National Key R&D Program of China under Grant 2017YFA0700800, and Natural Science Foundation of China (NSFC) under Grants 61671419, 61425026, 61622211 and 61620106009.

References

[1] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, 2018. 5

[2] Adrian Bulat and Georgios Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *CVPR*, 2018. 1

[3] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. 2018 pirm challenge on perceptual image super-resolution. In *ECCV Workshop*, 2018. 1

[4] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *ECCV*, 2018. 1, 3

[5] Xin Deng. Enhancing image quality via style transfer for single image super-resolution. *IEEE Signal Processing Letters*, 25(4):571–575, 2018. 1

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 2

[7] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 3, 6

[9] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018. 1, 2

[10] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *CVPR*, 2018. 2

[11] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *ICCV*, 2017. 3

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 1

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 1, 2, 3, 5, 6, 7, 8

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 2016. 2

[15] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017. 1, 2

[16] Christian Ledig, Lucas Theis, Ferenc Huszr, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 1, 2, 3, 5, 6, 7, 8

[17] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR Workshop*, 2017. 1

[18] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 3

[19] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-rolution. *Computer Vision and Image Understanding*, 158:1–16, 2017. 5, 6

[20] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *CVPR*, 2013. 1, 3

[21] Tobias Pltz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *CVPR*, 2017. 3

[22] M. S. M. Sajjadi, B. Schlkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, 2017. 1, 2

[23] Assaf Shocher, Nadav Cohen, and Michal Irani. zero-shot super-resolution using deep internal learning. In *CVPR*, 2018. 1, 3

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5, 6

[25] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, 2017. 1, 2

[26] Radu Timofte, Shuhang Gu, Jiqing Wu, and Luc Van Gool. Ntire 2018 challenge on single image super-resolution: Methods and results. In *CVPR Workshop*, 2018. 1, 3

[27] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *ICCV*, 2017. 1, 2

[28] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 1, 3

[29] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5, 6

[30] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep networks for image super-resolution with sparse prior. In *ICCV*, 2015. 2

[31] Z. Xiong, X. Sun, and F. Wu. Robust web image/video super-resolution. *IEEE Transactions on Image Processing*, 19(8):2017–2028, 2010. 1

[32] Z. Xiong, D. Xu, X. Sun, and F. Wu. Example-based super-resolution with soft information and decision. *IEEE Transactions on Multimedia*, 15(6):1458–1465, 2013. 1

[33] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *CVPR*, 2018. 1

[34] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[35] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 1, 2

[36] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 1, 2