ImagePairs: Realistic Super Resolution Dataset via Beam Splitter Camera Rig

Hamid Reza Vaezi Joze Ilya Zharkov Karlton Powell Carl Ringler Luming Liang
Andy Roulston Moshe Lutz Vivek Pradeep
Microsoft
www.microsoft.com/en-us/research/project/imagepairs

Abstract

Super Resolution is the problem of recovering a high-resolution image from a single or multiple low-resolution images of the same scene. It is an ill-posed problem since high frequency visual details of the scene are completely lost in low-resolution images. To overcome this, many machine learning approaches have been proposed aiming at training a model to recover the lost details in the new scenes. Such approaches include the recent successful effort in utilizing deep learning techniques to solve super resolution problem. As proven, data itself plays a significant role in the machine learning process especially deep learning approaches which are data hungry. Therefore, to solve the problem, the process of gathering data and its formation could be equally as vital as the machine learning technique used. Herein, we are proposing a new data acquisition technique for gathering real image data set which could be used as an input for super resolution, noise cancellation and quality enhancement techniques. We use a beam-splitter to capture the same scene by a low resolution camera and a high resolution camera. Since we also release the raw images, this large-scale dataset could be used for other tasks such as ISP generation. Unlike current small-scale dataset used for these tasks, our proposed dataset includes 11,421 pairs of low-resolution high-resolution images of diverse scenes. To our knowledge this is the most complete dataset for super resolution, ISP and image quality enhancement. The benchmarking result shows how the new dataset can be successfully used to significantly improve the quality of real-world image super resolution.

1. Introduction

Super Resolution (SR) is the problem of recovering high-resolution (HR) image from a single or multiple low-resolution (LR) images of the same scene. In this paper we are focusing on single-image SR which uses a single LR image as input. It is an ill-posed problem as the high frequency visual details of the scene are lost in the LR image while the HR image is being recovered. Therefore, the SR techniques are proven to be restrictive for usage in the practical applications [2]. SR could be used for many different applications such as satellite and aerial imaging [58], medical image processing [75], infrared imaging [78], improvement of text, sign and license plate [3], and finger prints [12].

Figure 1. Super Resolution Process on an image from ImagePairs dataset using bicubic, SRGAN [39] and EDSR [42] methods.

Figure 1 shows an example of single-image SR process where the recovered HR image is 4 times larger that its LR input image. We show the result of different super resolution techniques in this figure. If the technique fails to recover adequate detail from the LR input, the output will be blurry without sharp edges.

The SR problem has been studied comprehensively in the past [59, 48] and many machine learning techniques has been proposed to solve this problem. Examples would include Bayesian [61], steering kernel regression [73], adaptive Wiener filter [26], neighbor embedding [22, 6], matching [57] and example-based [21] methods.

Deep learning techniques have been proven a success in many areas of computer vision. This involves application of deep learning techniques by the lead image restoration researcher to solve SR [27, 38, 35, 68, 45, 44, 24, 15]. Because of the nature of deep the learning networks, being a

multi-layered feature extraction cascade [14], more data is required in order to train these complex methods [51].

As proven, the input data itself plays a significant role in the machine learning processes [4, 79], especially deep learning approaches which are data hungry. Hence, the process of gathering data and its formation may be equally as vital to solving the machine learning problem as the technique used. The sole purpose of SR is not to upscale or to increase the number of pixels in an image, but to increase the quality of it as closely to an image with the target resolution as possible. An example would be capturing a photo using a cellphone with a $5MP$ front facing camera and a $20MP$ rear facing camera where a 2X SR technique applied to the front facing camera will make it $20MP$. This is an attempt to increase the number of pixel from $5MP$ to $20MP$ while expecting an increase in the the quality the output image similar to that of the high quality rear facing camera. An example is presented in Fig. 1, where the same scene was photographed with a $5MP$ camera and $20.1MP$ camera in the same lighting condition. The same part of the image was cropped to show the nature of the difference in the quality of the images (ground truth vs. bicubic). This shows that maintaining the $20MP$ quality of the SR technique output requires SR, noise cancellation, image sharpening and even color correction to some extend while the-state-of-the-art methods such as SRGAN [39] and EDSR [42] fail to do so as seen in Fig. 1. We believe that the main reason for failure of these methods is lack of realistic training data that we focus on this paper.

A more complex version of this task could be Image Signal Processing (ISP) pipeline with various stages including denoising [8, 74], demosaicing [40], gamma correction, white balancing [63, 64] and so on. ISP pipeline has to be tuned by camera experts for a relatively long time before it can be used in the commercial cameras. Domain knowledge such as optics, mechanics of the cameras, electronics and human perception of colors and contrast are necessary in this tuning process. Replacing this highly skilled and tedious tuning process with a deep neural network is a recent research direction in computational photography [52, 50, 41, 32]. Current datasets [69, 1] widely used for training SR models increase the number of pixel without taking the quality of the image into consideration. The new data acquisition technique proposed herein may be used for SR, noise cancellation and quality enhancement techniques. A dataset of 11,421 pairs of LR-HR images is presented which was used to solve the SR problem. We use a beam-splitter to capture the same scene by two cameras: LR and HR. The proposed device can capture the same scene by two cameras, there still have a different perspective due to the different focal lenses, but we solve it by local alignment technique. Since we also release the raw images, this large-scale dataset could be used for other tasks such as ISP generation. To our knowledge, this is the most complete dataset for SR, ISP and image quality enhancement with far more real LR-HR images compared to existing dataset for SR and ISP task. This dataset is more than $10\times$ larger than current SR dataset while it includes real LR-HR pairs and more than $2\times$ larger than current ISP dataset (except [32] which is a concurrent work) while it includes diverse scenes. The benchmark result shows how the new dataset can be successfully used to significantly improve the quality of real-world image super resolution.

2. Related Works

In recent years, the core of image SR methods has shifted towards machine learning, mainly the machine learning techniques and the datasets. Herein, a brief description is given on the single image SR methods and learning-based ISP methods as well as their common datasets. There are also multiple image SR methods [61, 7, 19, 15] which are not the main focus of this paper. More comprehensive SR methods descriptions may be found in [59, 48].

2.1. SR methods

Interpolation based: The early SR methods are known as interpolation based methods where new pixels are estimated by interpolating given pixels. This is the easiest way to updates the image resolution. Examples include Nearest Neighbor interpolation, Bilinear interpolation and Bicubic interpolation which uses 1, 4 and 16 neighbor pixels respectively to compute the value of new pixels. These methods are wildly in use in image resizing.

Patch based: More recent SR methods rely on machine learning techniques to learn the relation between patches of HR image and patches of LR images. These methods are referred to as patch-based methods in some literature [69, 66] and Exemplar-based in other [21, 23]. Unlike the first class of methods, these methods need training data in order to train their models. These training data are usually pairs or corresponding LR and HR images. The training dataset is further discussed in subsection 2.3. Depending on the source of a training patch, the corresponding method for patch based SR may be categorized into two main categories: external or internal.

External methods the external method uses a variety of learning algorithms to learn the LR-HR mapping from a large database of LR-HR image pairs. These include nearest neighbor [21], kernel ridge regression [36], sparse coding [70] and convolutional neural networks [15].

Internal Methods the internal method on the other hand assumes that patches of a natural image recurs within and across scales of the same image [5]. Therefore, it makes an attempt to search for a HR patch within a LR image with different scales. Glasner et al. [23] united the classical and example-based SR by exploiting the patch recurrence

Data Set	Size	Main purpose	HR Resolution	LR generation	Raw
Set5 [6], Set14 [77], Urban100 [28]	5/14/100	SR	512×512	down-sample HR	No
The Berkeley segmentation [46]	200	Segmentation	481×321	down-sample HR	No
DIV2K [1]	1000	SR	2048×1080	down-sample HR	No
See-In-the-Dark (SID) [11]	5094	Low-Light	4240×2832	-	Yes
Samsung S7 [52]	110	ISP	4032×3024	-	Yes
RealSR [10]	595	SR	3500×700	Real	Yes
Zurich RAW to RGB [32]	48043	ISP	448×448	Real	Yes
ImagePairs (Proposed)	11421	SR	3504×2332	Real	Yes

Table 1. Compression between proposed dataset to current datasets used for its task.

within and across image scales. Freedman and Fattal [20] gained computational speed-up by showing that self-similar patches can often be found in limited spatial neighborhoods. Yang et al. [69] refined this notion further to seek self-similar patches in extremely localized neighborhoods, and performed first-order regression. Michaeli and Irani [47] used self-similarity to jointly recover the blur kernel and the HR image. Singh et al. [28] used the self-similarity principle for super-resolving noisy images.

With the success of convolution neural networks, many internal patch-based SR methods were proposed which outperform the prior methods. As an example, SRGAN [39] used a generative adversarial network (GAN) [25] for this task that trained by perceptual loss function consisting of an adversarial loss and a content loss. The residual dense networks (RDN) [76] exploited the hierarchical features from all the convolutional layers. EDSR [42] did a performance improvement by removing unnecessary modules in conventional residual networks. WDSR [71] introduced a linear low-rank convolution in order to further widen activation without computational overhead.

2.2. ISP Methods

Image Signal Processing (ISP) pipeline is a method used to convert an raw image into a digital form in order to get an enhanced image. This consists of various stages including denoising [8, 74], demosaicing [40], gamma correction, white balancing [62, 16] and so on. Currently, this pipeline has to be tuned by camera experts for a long period of time for each new camera. Replacing the expert-tuned ISP with a fully automatic method has been done with few recent methods approach by training an end-to-end deep neural network [52, 50, 41]. Schwartz et al. [52] released a data set, named Samsung S7 data set, contains RAW and RGB image pairs with both short and medium exposures. They design a network that first processes the image locally then globally. Ratnasingam [50] replicates the steps of a full ISP with a group of sub networks and achieves the-state-of-the-art result by training and testing on a set of synthetic images. Liang et al. [41] used 4 sequential u-nets in order to solve this problem. They claimed that the same network can be used for en-lighting extreme low light images.

2.3. SR and ISP Datasets

SR dataset includes pairs of HR and LR images. Most existing datasets generate an LR image from the corresponding HR image by sub-sampling the image using various settings. Here the HR images are also called ground truth as the final goal of SR methods is to retrieve them from LR images. Therefore, SR dataset includes sets of HR images or ground truths and settings to generate LR image from HR images. Here is a list of common SR datasets:

1. The Berkeley segmentation dataset [46] is one of the first datasets used for single image SR [56, 20, 23]. It includes 200 professional photographic style images of 481×321 pixels with a diverse content.
2. Yang et. al. [69] proposed a benchmark for single image SR which includes The Berkeley segmentation dataset as well as a second set containing 29 undistorted high-quality images from the LIVE1 dataset [54] , ranging from 720×480 to 768×512 pixels. Huang et. al. [27] added 100 urban high resolution images from flicker100 with a variety of real-world structures to this benchmark, in order to focus more on man made object.
3. DIV2K dataset [1] has introduced a new challenge for single image SR. This database include 1000 images of diverse contents with train/test/validation split as 90/10/10.
4. RealSR dataset [10] captured images of the same scene using fixed DSLR cameras with different focal lengths. The focal length changes can capture finer details of the scene. This way, HR and LR image pairs on different scales can be collected with a registration algorithm. This dataset includes 595 LR/HR pairs of indoor and outdoor scenes.

There are also few standard benchmark datasets, Set5 [6], Set14 [77], and Urban100 [28] commonly used for performance comparison. These datasets include 5 ,14 and 100 images, respectively. Apart from RealSR [10], all other datasets do not include LR images so the LR image should generate synthetically from corresponding HR image. There are several ways to generate LR test images

from HR images (the ground truth) [53, 60, 55] such that the generated LR test images may be numerically different. One common way to achieve this is to generate a LR image in a Gaussian blur kernel to down-sample the HR image using a noise term [33, 36, 69]. The parameter for this task will be s as scale factor, α for Gaussian kernel and ϵ for noise factor. There are other datasets dedicated to image enhancements such as MIT5K [9] and DPED [29]. MIT5K [9] includes 5,000 photographs taken with SLR cameras, each image retouched by professionals to achieve visually pleasing renditions. DPED [29, 30] consists of photos taken synchronously in the wild by three smartphones and one DSLR camera. The smartphone images were aligned with DSLR images to extract 100×100 patches for CNN training including 139K, 160K and 162K pairs for each settings. This dataset was used in a challenge on image enhancement [30] as well as a challenge on RAW to RGB Mapping [31].

There are not many publicly available ISP dataset which requires raw image as well as generated image from that. Here we describe two datasets that were used for ISP.

1. **See-In-the-Dark (SID)**: proposed by Chen et al. [11], is a Raw-RGB dataset captured in extreme low-light where each short-exposure raw image is paired with its long-exposure RGB counterpart for training and testing [72]. Images in this dataset were captured using two cameras: Sony $\alpha 7SII$ and Fujifilm X-T2, each subset contains about 2500 images, with about 20% as test set. The raw format of Sony subset is the traditional 4-channel Bayer pattern that of Fuji subset is XTrans format with 9 channels. Beside raw and RGB data, their exposure times are provided alongside.
2. **Samsung S7**: captured by Schwartz et al. [52], contains 110 different RAW-RGB pairs, with train/test/validation split of 90/10/10. Different to the SID dataset, this one does not provide related camera properties such as the exposure time associated with the image pairs. The raw format here is also the traditional 4-channel Bayer pattern.
3. **Zurich RAW to RGB**: is a concurrent work proposed by Ignatov et al. [32] where 20 thousands photos were collected using a smartphone capturing RAW photos and a professional high-end DSLR camera. The captured RAW–RGB image (RAW from smartphone and RGB from DSLR) pairs are aligned using the same procedure as in [29]. The patches of size 448×448 were extracted from pairs, resulting in 48,043 RAW-RGB image pairs as a dataset for training ISP.

Current SR methods as well as learning based ISP methods are mainly focused on their learning process as mentioned before. Different machine learning techniques have been applied to these problems and recent efforts have involved training different deep neural network models.

Camera	Low-resolution	High-resolution
Image sensor format	1/4”	1/2.4”
pixel size	$1.4\mu m$	$1.12\mu m$
Resolution	5MP	20.1MP
FOV (H,V)	$64°, 50.3°$	$68.2°, 50.9°$
Lens focal length	$2.9mm$	$4.418mm$
Focus	fixed-focus	auto-focus

Table 2. Camera Specifications.

Comparing to datasets for popular computer vision tasks such as image classification [13], detection [18, 43], segmentation [43], video classification [34] and sign language recognition [65], there is an obvious lack of large realistic dataset for SR and ISP tasks despite of the potential to produce significant result by neural network techniques. Table 1 shows all these datasets currently used for SR and ISP tasks and their specification compared to our proposed dataset ImagePairs. Our proposed dataset is not only at least 10 times larger than other SR datasets and 2 times from other ISP datasets (except [32] which is a concurrent work), but also has real LR-HR images and includes raw images which could be used for other tasks.

3. Data Acquisition

3.1. Hardware Design

The high resolution camera used had a $20.1MP$, 1/2.4” format CMOS image sensor supporting $5344(H) \times 3752(V)$ frame capture, $1.12\mu m$ pixel size, and lens focal length of $f = 4.418mm$ (F/1.94), providing a $68.2° \times 50.9°$ field of view (FOV). The camera also featured bidirectional auto-focus (open loop VCM) and 2-axis optical image stabilization (closed loop VCM) capability.

The lower resolution fixed-focus camera used had a similar FOV with approximately half the angular pixel resolution. it also featured a $5MP$, 1/4” format CMOS image sensor supporting 2588×1944 frame capture, $1.4\mu m$ pixel

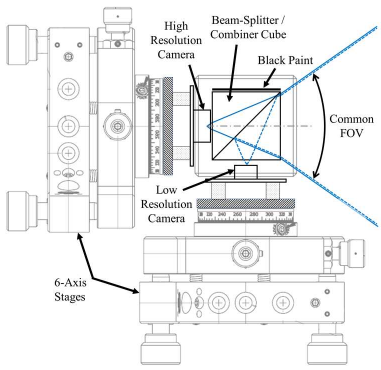

Figure 2. Opto-mechanical layout of dual camera combiner, showing high resolution camera (transmission path) and low resolution camera (reflective path) optically aligned at nodal points.

size, and lens focal length $f = 2.9mm$ (F/2.4), providing a $64°(H)x50.3°(V)$ FOV. Table 3.1 shows the specifications for these cameras.

In order to simultaneously capture frames on both cameras with a common perspective, the FOVs of both cameras are combined using a Thorlabs BS013 50/50 non-polarizing beam-splitter cube. They are then aligned such that pointing angle of the optical axes are at far distance and entrance pupils at each camera (nodes)are at near distance. The high resolution camera, placed behind the combiner cube in the transmission optical path, is mounted on a Thorlabs K6XS 6-axis stage so that the x and y position of the entrance pupil is centered with the cube and the z position in close proximity. The tip and tilt of camera image center field pointing angle is aligned with a target at distance while rotation about camera optical axis is aligned by matching pixel row(s) with a horizontal line target. Fig. 2 illustrates the opto-mechanical layout of the dual camera combiner. The low resolution camera is placed behind the combiner cube in the lateral $90°$ folded optical path and also mounted on a 6-axis stage. It is then aligned in x, y and z such that entrance pupil optically overlaps that of the high resolution camera. The tip/tilt pointing angle as well as camera rotation about optical axis may be adjusted so as to achieve similar scene capture. In order to refine the overlap toward pixel accuracy, a live capture tool displays the absolute difference of camera frame image content between cameras such that center pointing and rotation leveling may be adjusted with high sensitivity. Any spatial and angular offsets may be substantially nulled by mechanically locking the camera in position. The unused combiner optical path is painted with carbon black to limit image contrast loss due to scatter. The opto-mechanical layout of dual camera combiner is illustrated at figure 4.

The proposed device can capture the same scene by two different cameras. The two cameras have a difference in perspective due to the different focal lenses which was solved by a local alignment technique described in section 4. Furthermore, the two camera sensors get half of the light because of 50/50 split with poorer image quality mainly on low-resolution camera.

Figure 3. The data acquisition device install on a tripod while the trolley is used for outdoor manoeuvre.

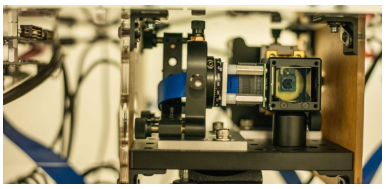

Figure 4. Two camera setup

3.2. Software Design

A data capturing software was developed to connect to both cameras, allowing them to synchronize with each other. The software may capture photo from both cameras at the same time as well as adjusting camera parameters such as gain, exposure and lens position for the HR camera. The raw data was stored for each camera, allowing later use of the arbitrary ISP. For each camera, all the meta data was stored on a file including the image category selected by the photographer. Figures 3 shows the data acquisition device installed on a tripod while the trolley is used for outdoor maneuvering.

4. ImagePairs Dataset

The dataset was called ImagePairs as it includs pairs of images of the exact scene using two different cameras. Images are either LR or HR where the HR image is twice as big in each dimensions as the corresponding LR image; all LR images are 1752×1166 pixels and HR images are 3504×2332 pixels. Unlike other real world datasets, we do not use zooming levels or scaling factor to increase the number of pairs so each pair corresponds to a separate scene. This means that we captures 11,421 distinct scenes with the device which generates 11,421 image pairs.

For each image pair, the meta data such as gain, exposure, lens position and scene categories were stored. Each image pair was assigned to a category which may later be used for training purposes. These categories include Document, Board, Office, Face, Car, Tree, Sky, Object, Night and Outdoor. The pairs are later divided in two sets of train and test, each including 8591 and 2830 image pairs, respectively. The two cameras have a difference in perspective due to the different focal lenses. Therefore, in order to generate pairs corresponding to each other in pixel level, the following steps were applied: (1) ISP (2) image undistortion (3) pair alignment (4) margin cropping. Figure 6 illustrates diverse samples from proposed dataset after the final alignments. In order to show the accuracy of pixel-by-pixel

Figure 5. Undistorted image of HR at right and LR at left.

alignment, each sample image is divided by half horizontally to show LR at left and HR at right in Fig. 6.

ISP : The original images were stored in the raw format. The first step was to convert the raw data to color images, using a full-stack powerful ISP for both LR and HR. Since we have access to the raw data, the ISP can be replaced with a different one or a simple linear one to ignore the non-linearity in the pipeline.

Image Undistortion : CMOS cameras introduce a lot of distortion to images. Two major distortions are radial distortion and tangential distortion. In radial distortion, straight lines will appear curved while in the tangential distortion the lens is not aligned perfectly parallel to the imaging plane. To overcome these distortions in both LR and HR images, we calibrated both cameras by capturing several checkerboard images. These images were later used to solve a simple model for radial and tangential distortions [49]. Figure 4 shows the un-distorted image for both LR and HR images.

Alignment : We use two steps in order to align the LR and HR images. First we try to globally match two images using image registration technique specifically homography transformation. Although now HR and LR image are globally aligned but they may not be aligned pixel by pixel due to some geometry constrains. So as the second step, we use a 10 by 10 grid for LR image and do a local search to find the best match on HR image for that grid. Lastly, we use matching position for grids on HR image to warp the LR image so that the LR and HR are globally and locally matched to each other.

Margin Crop : Although the images were aligned globally and locally, the borders are not as aligned as we expected, possibly due to differences in the camera specifications. Therefore, 10% of border from each image was removed, resulting in a change in the resolution of both LR and HR images; 1752×1166 pixels and 3504×2332 pixels respectively. For each image (LR or HR) we also stored meta data which is analogue gain, digital gain, exposure time, lens position and scene category. The scene category which is selected by the photographer includes Office, Document, Tree, Outside, Toy, Sky, Sign, Art, Building, Night, etc. Figure 7 illustrates the frequency of each categories for ImagePairs train/test sets.

At this point, the ImagePairs consists of a large dataset

Figure 6. Sample images from ImagePairs dataset. Each image divided by half horizontally to show LR (left) and HR (right).

of HR-LR images, allowing the easy application of patch-base algorithms. Random patches can pick from LR and the corresponding HR patches. Since the correspondence is pixel by pixel, there is no need to search for similar patches in different scales. Additionally, the ground truth (HR) has 4 times more pixels, is sharper and less noisy compared to the LR images, hence an increased image quality.

5. Experimental Results

5.1. Realistic Super Resolution

Before running a benchmark for state-of-the-art SR methods, we need to see their performance when trained on current SR datasets. As mentioned before, a real LR image usually has many other artifacts as it is captured with a weaker camera. We train a basic generative adveral network (GAN) model which includes 10 convolution layers for generator and a U-Net with 10 convolution/deconvolution for discriminator network with the proposed dataset. The sole reason of this experiment is to see if current state-of-the-art methods trained on synthetic images can outperform

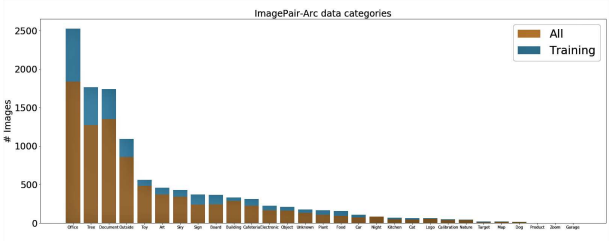

Figure 7. Frequency of ImagePairs train/test categories.

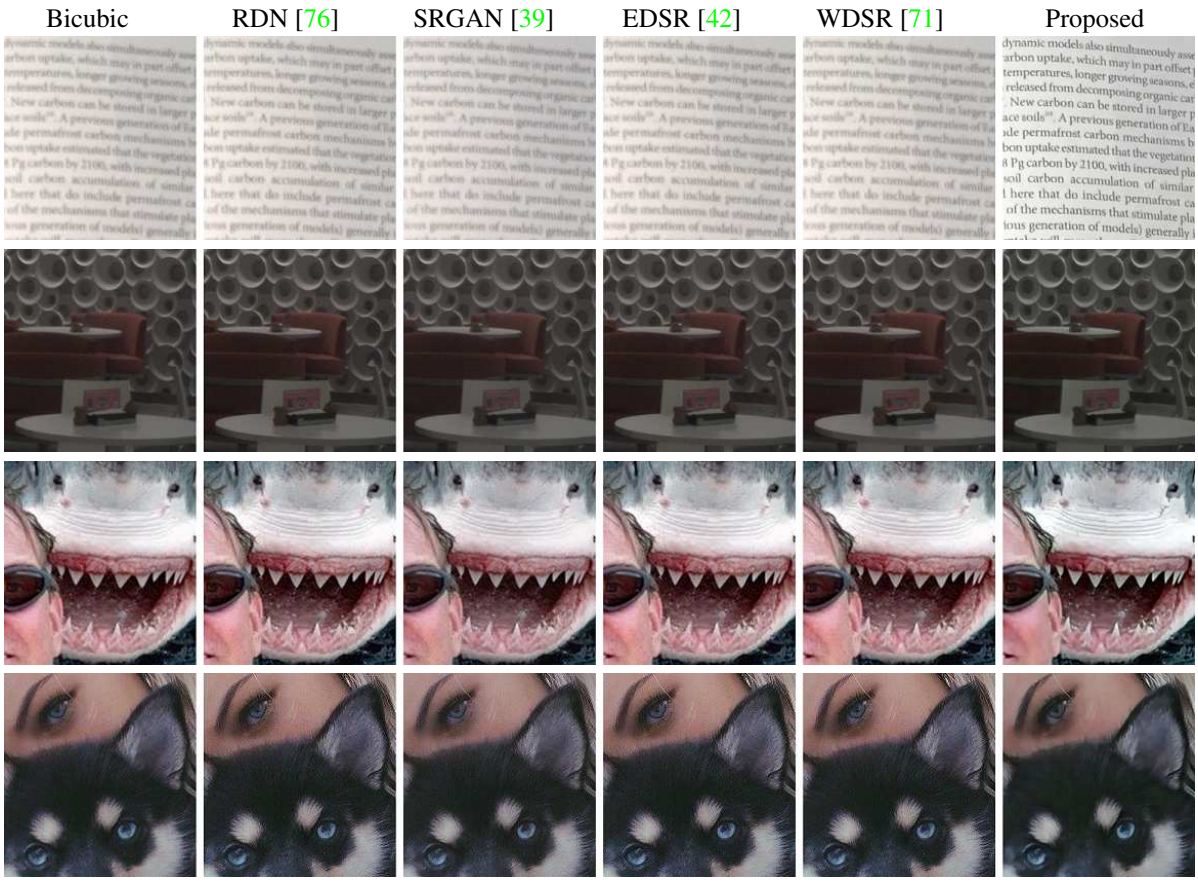

Figure 8. Qualitative compassion of the-state-of-the-art super resolution methods train on DIV2K [1] dataset set compare to simple network trained on ImagePairs dataset. The first two images are from ImagePairs test set and the next two are external images.

our simple method training on real images or not. Figure 8 shows the performance of this method compared to the super resolution methods: SRGAN [39], EDSR [42], WDSR [71] and RDN [76] trained on DIV2K dataset [1]. The first two images are from ImagePairs test set and the next two images are from real-world LR images from the internet. As expected, these methods only increase the pixel and do not effect image artifacts like noise and color temperature. Our method trained on ImagePairs dataset does well for images from the dataset and real-world LR images.

5.2. Super Resolution Benchmark

We trained three 2X super resolution methods on ImagePairs train set including SRGAN [39], EDSR [42] and WDSR [71] by using their model implementation by [37, 17]. All SR methods trained using LR-HR rgb images and we do not use raw images as input. We use same patch size of 128×128 for HR images and batch size equal to 16 for all training. All methods are trained for $150K$ iterations. For evaluation, we run trained models on centered quarter of cropped images of Imagepairs test set. Table 5.2 reports the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) [67] for trained model on ImagePairs as well as model trained on DIV2K dataset with similar parameters. As we discussed before, the PSNR and SSIM for methods trained on DIV2K is comparable with bicubic method. In some cases, they perform worst than bicubic since noise could boost with some SR methods. On the other hand, when we trained the same models with proposed ImagePairs dataset, all methods outperform their PNSR. SRGAN [39] and EDSR [42] is doing a good job in noise cancellation and outperform at least 2 db for PSNR and 0.6 on SSIM. On the other hand, SRGAN [39] which is not optimized for PSNR, mainly focuses on color correction and not much on noise cancellation. Figure 9 illustrates qualitative comparison of these methods trained on ImagePairs dataset. Needless to say, these models perform much better on nose cancellation, color correction and super resolution compared to similar models trained on DIV2K.

5.3. ISP Benchmark

For ISP task, we consider LR images and their corresponding raw images of ImagePairs train/test sets as the raw HR images are too large. We trained DeepISPnet [52], SIDnet [11] and GuidanceNet [41] on ImagePairs training set which contains raw and LR images. All networks read

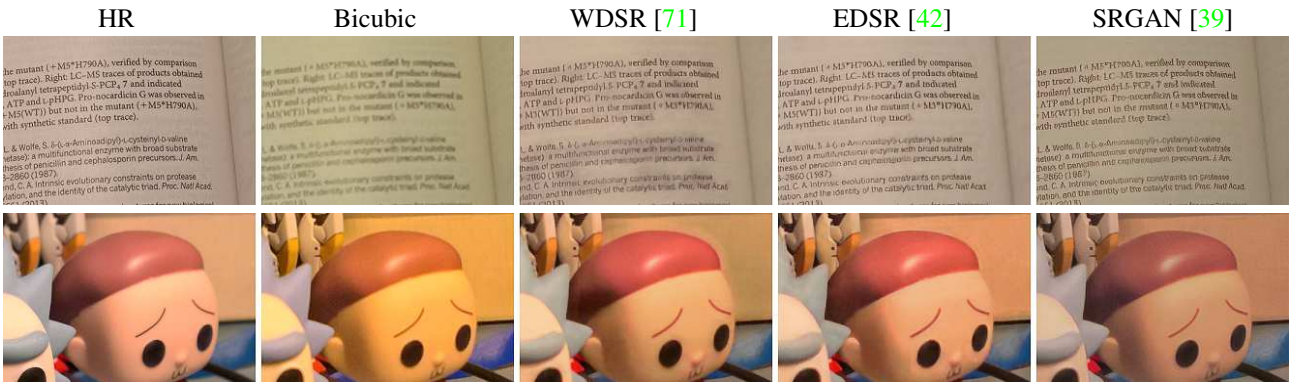

Figure 9. Qualitative comparison of the-state-of-the-art super resolution methods train on proposed dataset.

Figure 10. Qualitative comparisons of state-of-the-art ISP methods trained on ImagePairs dataset.

Model	Train data	PSNR (db)	SSIM
Bicubic	-	21.451	0.712
SRGAN [39]	DIV2K	21.906	0.699
WDSR [71]	DIV2K	21.299	0.697
EDSR [42]	DIV2K	21.298	0.697
SRGAN [39]	ImagePairs	22.161	0.673
WDSR [71]	ImagePairs	23.805	0.767
EDSR [42]	ImagePairs	23.845	0.764

Table 3. Comparisons of state-of-the-art single image super resolution algorithms on ImagePairs data set.

RAW images and associated 4 camera properties: analogue gain, digital gain, exposure time and lens position. Here, the exposure time is in microsecond; the lens position is the distance between the camera and the scene in centimeters. GuidanceNet [41] is designed to use camera properties in its bottleneck layers, but we modified DeepISPnet [52] and SIDnet [11]. For DeepISPnet [52], we tile and concatenate these features with the output of their local sub-network and then feed it to the global sub-network for estimating the quadratic transformation coefficients. For SIDnet [11], we tile and concatenate these features with the input image. Tables 5.3 reports the evaluation of these three models on ImagePairs test set in term of PSNR and SSIM. This shows

Model	PSNR (db)	SSIM
DeepISP [52]	20.30	0.89
SIDnet [11]	23.08	0.90
GuidanceNet [41]	29.22	0.96

Table 4. Comparisons of ISP algorithms on ImagePairs dataset.

GuidanceNet [41] which properly used camera properties outperform others. Figure 10 illustrates examples for each of these models.

6. Conclusion

In this paper we proposed a new data acquisition technique which could be used as an input for SR, noise cancellation and quality enhancement techniques. We used a beam-splitter to capture the same scene by a low resolution camera and a high resolution camera. Unlike current small-scale datasets used for these tasks, our proposed dataset includes 11,421 pairs of low-resolution and high-resolution images of diverse scenes. Since we also release the raw images, this large-scale dataset could be used for other tasks such as ISP generation. We trained state-of-the art methods for SR and ISP tasks on this dataset and showed how the new dataset can be successfully used to improve the quality of real-world image super resolution significantly.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 2, 3, 7

[2] Simon Baker and Takeo Kanade. Limits on super-resolution and how to break them. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1167–1183, 2002. 1

[3] Jyotirmoy Banerjee and CV Jawahar. Super-resolution of text images using edge-directed tangent field. In *2008 The Eighth IAPR International Workshop on Document Analysis Systems*, pages 76–83. IEEE, 2008. 1

[4] Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting on association for computational linguistics*, pages 26–33. Association for Computational Linguistics, 2001. 2

[5] Michael F Barnsley. *Fractals everywhere*. Academic press, 2014. 2

[6] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. BMVA press, 2012. 1, 3

[7] Sean Borman and Robert L Stevenson. Super-resolution from image sequences-a review. In *Circuits and Systems, 1998. Proceedings. 1998 Midwest Symposium on*, pages 374–378. IEEE, 1998. 2

[8] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005. 2, 3

[9] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 4

[10] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3086–3095, 2019. 3

[11] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3, 4, 7, 8

[12] Jiali Cui, Yunhong Wang, JunZhou Huang, Tieniu Tan, and Zhenan Sun. An iris image synthesis method based on pca and super-resolution. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 471–474. IEEE, 2004. 1

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 4

[14] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014. 2

[15] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, pages 184–199. Springer, 2014. 1, 2

[16] Mark S Drew, Hamid Reza Vaezi Joze, and Graham D Finlayson. The zeta-image, illuminant estimation, and specularity manipulation. *Computer Vision and Image Understanding*, 127:1–13, 2014. 3

[17] Francesco Cardinale et al. Isr. https://github.com/idealo/image-super-resolution, 2018. 7

[18] M. "Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A." Zisserman. "the pascal visual object classes challenge: A retrospective". *"International Journal of Computer Vision"*, "111"("1"):98–136, "2015". 4

[19] Sina Farsiu, M Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and robust multiframe super resolution. *IEEE transactions on image processing*, 13(10):1327–1344, 2004. 2

[20] Gilad Freedman and Raanan Fattal. Image and video upscaling from local self-examples. *ACM Transactions on Graphics (TOG)*, 30(2):12, 2011. 3

[21] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002. 1, 2

[22] Xinbo Gao, Kaibing Zhang, Dacheng Tao, and Xuelong Li. Image super-resolution with sparse neighbor embedding. *IEEE Transactions on Image Processing*, 21(7):3194–3205, 2012. 1

[23] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 349–356. IEEE, 2009. 2, 3

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 1

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[26] Russell Hardie. A fast image super-resolution algorithm using an adaptive wiener filter. *IEEE Transactions on Image Processing*, 16(12):2953–2964, 2007. 1

[27] Jia-Bin Huang, Abhishek Singh, and Narendra 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. 1, 3

[28] Jia-Bin Huang, Abhishek Singh, and Narendra 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. 3

[29] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3277–3285, 2017. 4

[30] Andrey Ignatov and Radu Timofte. Ntire 2019 challenge on image enhancement: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 4

[31] Andrey Ignatov, Radu Timofte, Sung-Jea Ko, Seung-Wook Kim, Kwang-Hyun Uhm, Seo-Won Ji, Sung-Jin Cho, Jun-Pyo Hong, Kangfu Mei, Juncheng Li, et al. Aim 2019 challenge on raw to rgb mapping: Methods and results. In *IEEE International Conference on Computer Vision (ICCV) Workshops*, volume 5, page 7, 2019. 4

[32] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. *arXiv preprint arXiv:2002.05509*, 2020. 2, 3, 4

[33] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 53(3):231–239, 1991. 4

[34] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 4

[35] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral)*, June 2016. 1

[36] Kwang In Kim and Younghee Kwon. Single-image super-resolution using sparse regression and natural image prior. *IEEE transactions on pattern analysis and machine intelligence*, 32(6):1127–1133, 2010. 2, 4

[37] Martin Krasser. Single image super-resolution with edsr, wdsr and srgan. https://github.com/krasserm/super-resolution, 2018. 7

[38] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a adversarial network. *CoRR*, abs/1609.04802, 2016. 1

[39] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 2, 3, 7, 8

[40] Xin Li, Bahadir Gunturk, and Lei Zhang. Image demosaicing: A systematic survey. In *Visual Communications and Image Processing 2008*, volume 6822, page 68221J. International Society for Optics and Photonics, 2008. 2, 3

[41] Luming Liang, Ilya Zharkov, Faezeh Amjadi, Hamid Reza Vaezi Joze, and Vivek Pradeep. Guidance network with staged learning for computational photography. In *arXiv*, 2020. 2, 3, 7, 8

[42] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 1, 2, 3, 7, 8

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[44] Ding Liu, Zhaowen Wang, Nasser Nasrabadi, and Thomas Huang. Learning a mixture of deep networks for single image super-resolution. In *Asian Conference on Computer Vision*, 2016. 1

[45] Ding Liu, Zhaowen Wang, Bihan Wen, Jianchao Yang, Wei Han, and Thomas S Huang. Robust single image super-resolution via deep networks with sparse prior. *IEEE Transactions on Image Processing*, 25(7):3194–3207, 2016. 1

[46] David Martin, Charless Fowlkes, Doron Tal, and Jitendra 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. 3

[47] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–952, 2013. 3

[48] Kamal Nasrollahi and Thomas B Moeslund. Super-resolution: a comprehensive survey. *Machine vision and applications*, 25(6):1423–1468, 2014. 1, 2

[49] OpenCV. Camera calibration with opencv. http://docs.opencv.org/2.4/doc/tutorials/calib3d/camera_calibration/camera_calibration.html. 6

[50] Sivalogeswaran Ratnasingam. Deep camera: A fully convolutional neural network for image signal processing. In *International Conference on Computer Vision Workshops (ICCVW)*, August 2019. 2, 3

[51] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015. 2

[52] Eli Schwartz, Raja Giryes, and Alexander M. Bronstein. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE Transactions on Image Processing*, 28:912–923, 2019. 2, 3, 4, 7, 8

[53] Qi Shan, Zhaorong Li, Jiaya Jia, and Chi-Keung Tang. Fast image/video upsampling. *ACM Transactions on Graphics (TOG)*, 27(5):153, 2008. 4

[54] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing*, 15(11):3440–3451, 2006. 3

[55] Jian Sun, Zongben Xu, and Heung-Yeung 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. 4

[56] Jian Sun, Jiejie Zhu, and Marshall F Tappen. Context-constrained hallucination for image super-resolution. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 231–238. IEEE, 2010. 3

[57] Libin Sun and James Hays. Super-resolution from internet-scale scene matching. In *Proceedings of the IEEE Conf. on International Conference on Computational Photography (ICCP)*, 2012. 1

[58] Matt W Thornton, Peter M Atkinson, and DA 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. 1

[59] Jing Tian and Kai-Kuang Ma. A survey on super-resolution imaging. *Signal, Image and Video Processing*, 5(3):329–342, 2011. 1, 2

[60] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1920–1927, 2013. 4

[61] Michael E Tipping, Christopher M Bishop, et al. Bayesian image super-resolution. In *NIPS*, volume 15, pages 1279–1286, 2002. 1, 2

[62] Hamid Reza Vaezi Joze and Mark S Drew. Exemplar-based colour constancy. In *BMVC*, pages 1–12, 2012. 3

[63] Hamid Reza Vaezi Joze and Mark S Drew. Exemplar-based color constancy and multiple illumination. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):860–873, 2013. 2

[64] Hamid Reza Vaezi Joze, Mark S Drew, Graham D Finlayson, and Perla Aurora Troncoso Rey. The role of bright pixels in illumination estimation. In *Color and Imaging Conference*, volume 2012, pages 41–46. Society for Imaging Science and Technology, 2012. 2

[65] Hamid Reza Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. In *The British Machine Vision Conference (BMVC)*, September 2019. 4

[66] Qiang Wang, Xiaoou Tang, and Harry Shum. Patch based blind image super resolution. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 709–716. IEEE, 2005. 2

[67] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[68] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang. Deep networks for image super-resolution with sparse prior. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 370–378, 2015. 1

[69] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In *European Conference on Computer Vision*, pages 372–386. Springer, 2014. 2, 3, 4

[70] Jianchao Yang, Zhaowen Wang, Zhe Lin, Scott Cohen, and Thomas Huang. Coupled dictionary training for image super-resolution. *IEEE transactions on image processing*, 21(8):3467–3478, 2012. 2

[71] Jiahui Yu, Yuchen Fan, Jianchao Yang, Ning Xu, Zhaowen Wang, Xinchao Wang, and Thomas Huang. Wide activation for efficient and accurate image super-resolution. *arXiv preprint arXiv:1808.08718*, 2018. 3, 7, 8

[72] Syed Waqas Zamir, Aditya Arora, Salman Khan, Fahad Shahbaz Khan, and Ling Shao. Learning digital camera pipeline for extreme low-light imaging. Technical report, ArXiV, 2019. 4

[73] Kaibing Zhang, Xinbo Gao, Dacheng Tao, and Xuelong Li. Single image super-resolution with non-local means and steering kernel regression. *IEEE Transactions on Image Processing*, 21(11):4544–4556, 2012. 1

[74] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 2, 3

[75] S. Zhang, G. Liang, S. Pan, and L. Zheng. A fast medical image super resolution method based on deep learning network. *IEEE Access*, 7:12319–12327, 2019. 1

[76] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 3, 7

[77] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for neural networks for image processing. *arXiv preprint arXiv:1511.08861*, 2015. 3

[78] Yao Zhao, Qian Chen, Xiubao Sui, and Guohua Gu. A novel infrared image super-resolution method based on sparse representation. *Infrared Physics & Technology*, 71:506–513, 2015. 1

[79] Xiangxin Zhu, Carl Vondrick, Deva Ramanan, and Charless C Fowlkes. Do we need more training data or better models for object detection?. In *BMVC*, volume 3, page 5, 2012. 2