# Supplementary Material for D2HNet: Joint Denoising and Deblurring with Hierarchical Network for Robust Night Image Restoration

Yuzhi Zhao[1], Yongzhe Xu[2], Qiong Yan[2], Dingdong Yang[2], Xuehui Wang[3], and Lai-Man Po[1]

[1] Department of Electrical Engineering, City University of Hong Kong, China
[2] SenseTime Research and Tetras.AI
[3] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
yzzhao2-c@my.cityu.edu.hk, wangxuehui@sjtu.edu.cn, eelmpo@cityu.edu.hk,
{xuyongzhe1,yanqiong,yangdingdong}@tetras.ai

## 1 More Results on Captured Real Images

We show more visual results of D2HNet and SOTA methods on real images in Figure 1, which are captured with *Xiaomi Mi Note 10* smartphone. The texture learning ability, denoising quality, and artifact removal performance of the proposed D2HNet are all better than SOTA methods. The more detailed analysis is in the captions.

## 2 More Results on Validation Set

We show more visual results of D2HNet and SOTA methods on the validation set of the collected D2-Dataset. The results on 1440p data and 2880p data are shown in Figure 2 and Figure 3, respectively. The D2HNet produces more distinguishable details and achieves better deblurring quality. It also achieves consistent and better performance on different image resolutions.

## 3 Burst-image Method Experiments

We compare D2HNet with a burst-image denoising method KPN [1]. The training set of KPN is also generated from the same video source of D2-Dataset and 4 successive short-exposure images are synthesized by a similar process used in D2-Dataset, then augmented with the same noise parameters as D2HNet. The results are shown in Figure 4, where D2HNet produces richer textures (e.g., flowers in ♯2) and has fewer visual artifacts (e.g., black car in ♯1 and dark road in ♯3) than KPN. Since KPN defines a fixed size of output convolutional kernels, it is not flexible to image resolutions larger than training images, i.e., it cannot address the domain gap issue. In addition, burst capturing with 4 shots takes more time than 2 shots due to hardware constraints. And more shots introduce more misalignment issues. Hence our D2HNet framework is more favorable.

## 4 More Results Related to Domain Gap

The domain gap in the task means differences between synthetic training images and real-world photos, e.g., blur area and resolution between them. To further demonstrate that D2HNet addresses the domain gap issue, we add an experiment setting that uses D2HNet architecture but does not perform downsampling for the input images of DeblurNet. The visual comparisons are shown in Figure 5. We observe that the pixel shifts of most highly blurry tuples are in the range of [40, 100], where some samples are shown in Figure 5 (b). Since D2HNet architecture without downsampling only sees a maximum pixel shift of approximately 100, while the pixel shifts of the input pairs shown in Figure 5 (a) are much larger than 100 (e.g., larger than 150 for the black T-shirt patch), it cannot handle such cases. Therefore, there are obvious artifacts in the results.

## 5 Illustration of Data Acquisition

We synthesize a D2-Dataset for training and benchmarking. There are three steps of the data synthesis pipeline, where the details are shown in Figure 6 (a). For the data synthesis pipeline for training the burst-image denoising method, the details are shown in Figure 6 (b). We also show some long- and short-exposure image pairs in Figure 6 (c).

## 6 Illustration of Data Processing Schemes

To further visualize the effectiveness of VarmapSelection and CutNoise schemes, we show 4 examples in Figure 7. The variance maps of VarmapSelection can well represent the regional blur degree; therefore, it helps select blurry patches at the training. It makes the D2HNet better generalize to blurry long-exposure inputs. The CutNoise makes a region of the short-exposure input image the same as ground truth; therefore, D2HNet learns to directly use the short-exposure input at this region. It makes D2HNet learn where to deblur and enhance long-exposure images in addition to how to deblur and enhance long-exposure images [3]. Also, it helps balance the usage of long- and short-exposure inputs.

## 7 More details of Noise Model

We use the physics-based noise model [2] to calibrate the *Xiaomi Mi Note 10* smartphone for training the D2HNet. The ISO range of this smartphone is [100, 12800]. At the training, we randomly select the long-exposure ISO from [1000, 4000] and the short-exposure ISO from [6400, 12800] uniformly. It ensures that the noises in the long-exposure input are slighter than in the short-exposure input. At the validation, we add noises to clean validation images from D2-Dataset as inputs. The same ISO ranges are used for validation images. At the testing, since the D2HNet is trained with the calibrated noise model, it can directly enhance the long- and short-exposure image pair captured by the smartphone. We show some samples in Figure 8 to illustrate the noise calibration results.


**Fig. 1.** Visual comparisons of the proposed D2HNet with other methods on real photos. From ♯1 and ♯2, there are no visual artifacts of D2HNet, while there are obvious artifacts for other methods. From ♯3, there are very obvious remaining noises in the dark sky of other methods, while the D2HNet output is much cleaner. From ♯4 to ♯6, we observe that D2HNet can well recover textures and remove artifacts simultaneously when there are a lot of details in the input images (especially in the long-exposure inputs). For instance, the Chinese characters in ♯4 of D2HNet are cleaner and clearer than other methods; the letters and numbers "B 4020X" in ♯5 of D2HNet are more distinguishable than other methods; D2HNet better learns the textures from ♯6 inputs.

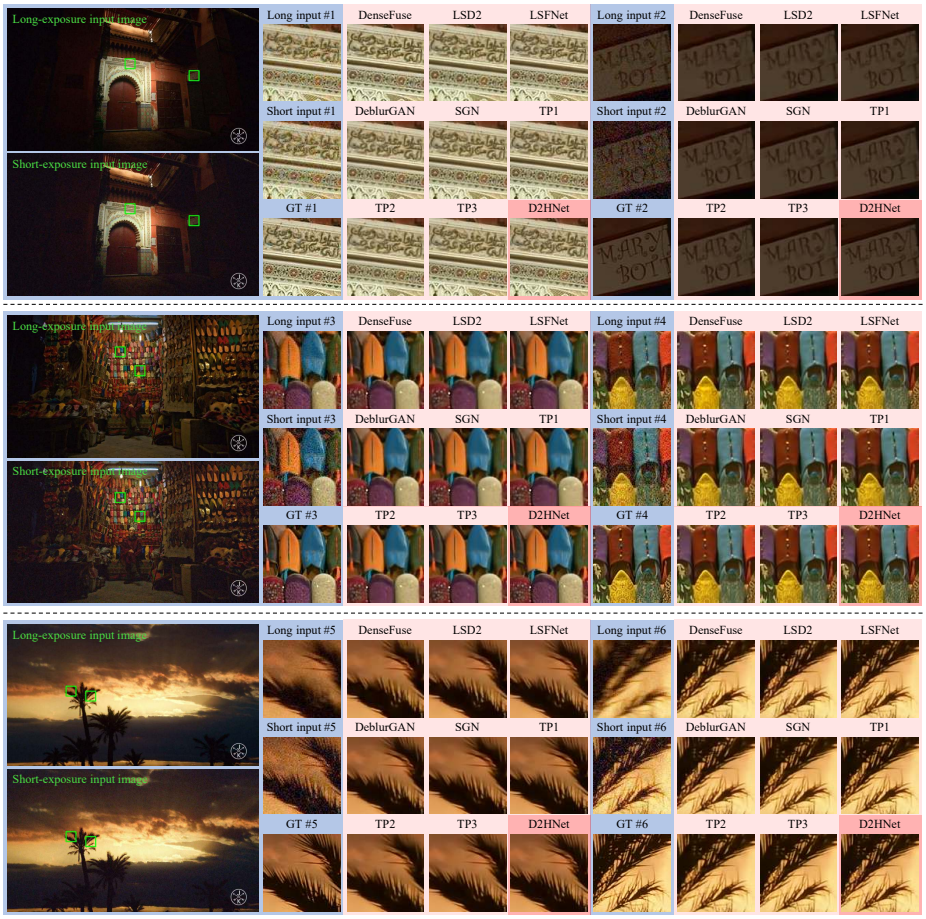



**Fig. 2.** Visual comparisons of the proposed D2HNet with other methods on D2-Dataset 1440p validation set. Note that, we also show the ground truth since the experiments are performed on validation set. From ♯1, ♯3, and ♯4, textures of the painting and shoes in D2HNet are clearer than other methods (please compare the details of different results based on ground truth, i.e., GT ♯1, GT ♯3, and GT ♯4). From ♯2, the letters of D2HNet are more distinguishable than other methods, e.g., the edges and clarity. From ♯5 and ♯6, the edges of D2HNet results are better than other methods.

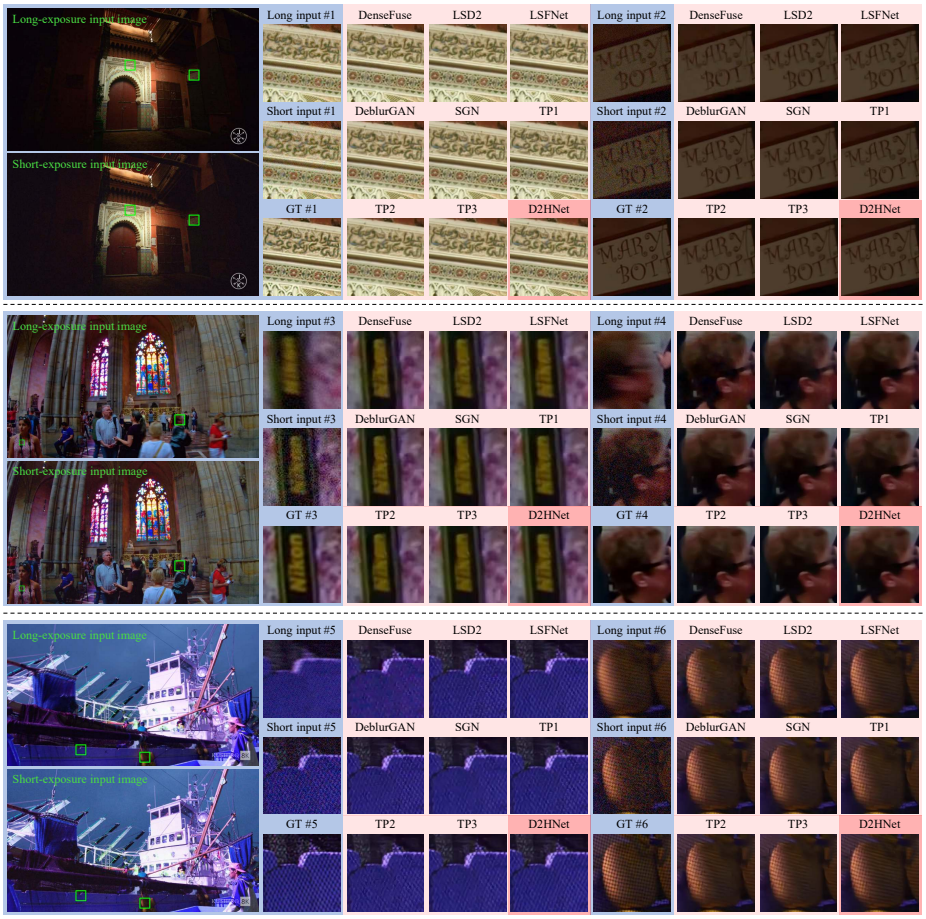



**Fig. 3.** Visual comparisons of the proposed D2HNet with other methods on D2-Dataset 2880p validation set. Note that, we also show the ground truth since the experiments are performed on validation set. The input examples ♯1 and ♯2 are at the same relative positions to Figure 2, but with different image resolutions. The textures and edges of D2HNet results are better than other methods. The proposed D2HNet performs well on both 1440p and 2880p validation images, which demonstrates that D2HNet has the ability to address the domain gap issue.

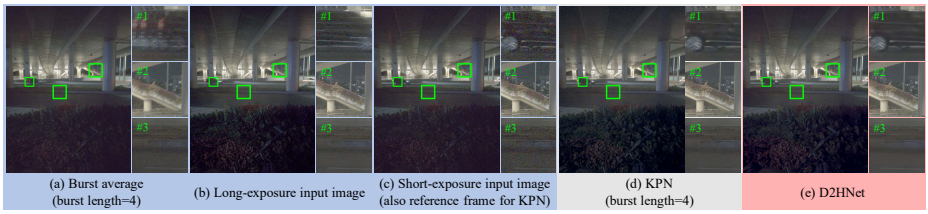



**Fig. 4.** Visual comparisons of the proposed D2HNet with KPN. The input long-short pairs and 4-frame images are captured by the same smartphone and in the same scene.

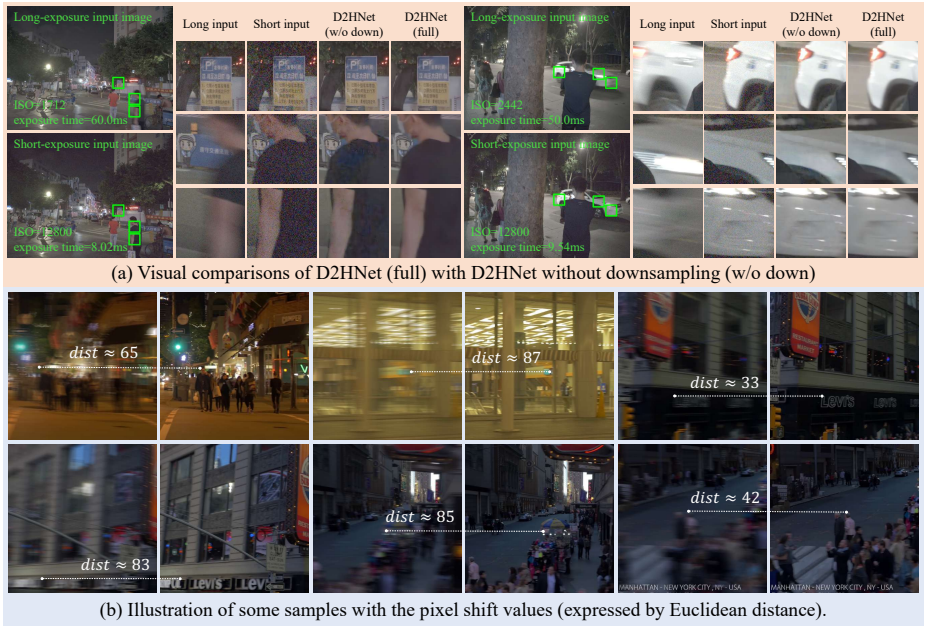
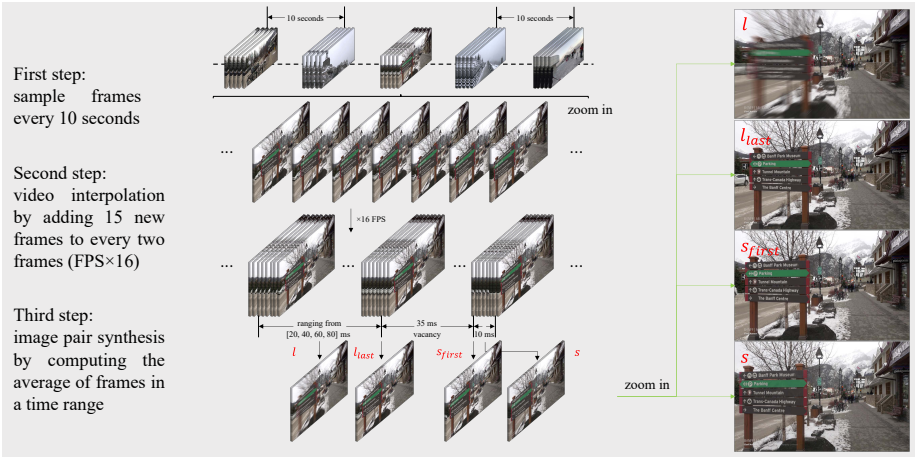
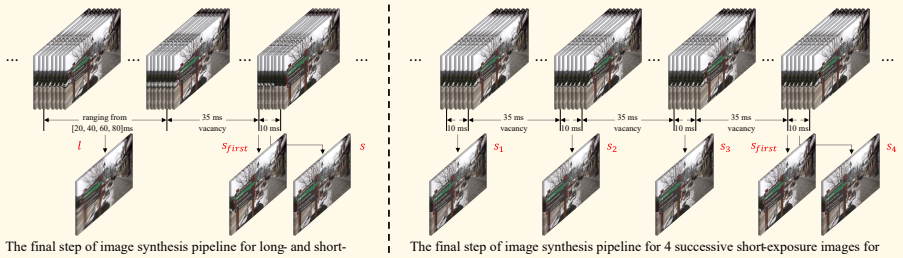



**Fig. 5.** (a) Visual comparisons of the proposed D2HNet with the same architecture but without downsampling. We select some patches from the highly blurry areas in the long-exposure input. There are obvious artifacts of the D2HNet (w/o down) results, while much fewer artifacts are in D2HNet (full) results. Since D2HNet (w/o down) does not consider the domain gap issue, it cannot handle real-world inputs with larger pixel shifts than training images. Therefore, it simply copies the pixels of the long-exposure input to the output, e.g., there are many blue pixels on the black T-shirt of D2HNet (w/o down) results; (b) Illustration of pixel shift values of some training long- and short-exposure image pairs. The image pairs are selected from 9453 highly blurry tuples, which are obtained by the VarmapSelection scheme.



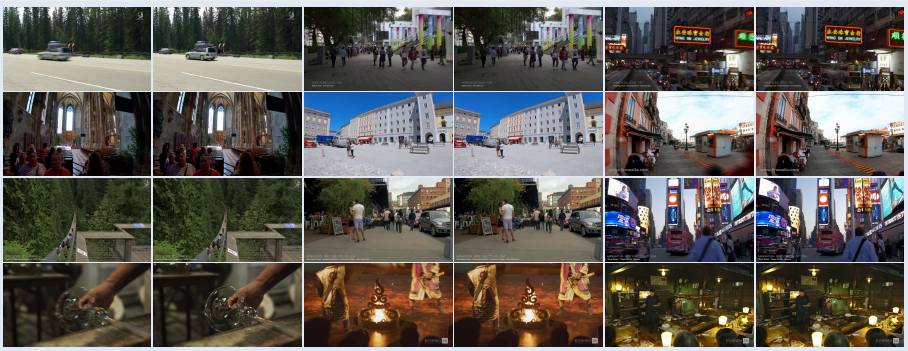

(a) Illustration of the data synthesis pipeline. An image tuple is shown as an example.




The final step of image synthesis pipeline for long- and short-exposure images for training D2HNet:
Long-exposure image: $l$; Short-exposure image: $s$; GT: $s_{first}$

The final step of image synthesis pipeline for 4 successive short-exposure images for training the burst-image denoising method:
4 successive short-exposure images: $s_1 - s_4$; Ground truth: $s_{first}$

(b) The comparisons of image synthesis pipelines for long-short fusion and burst-image denoising, which share the first and second steps but not share the final step.



(c) Illustration of some images in the D2-Dataset, where only long-short pairs are shown for simplicity.

**Fig. 6.** Illustration of the image synthesis pipeline of D2-Dataset and some examples.

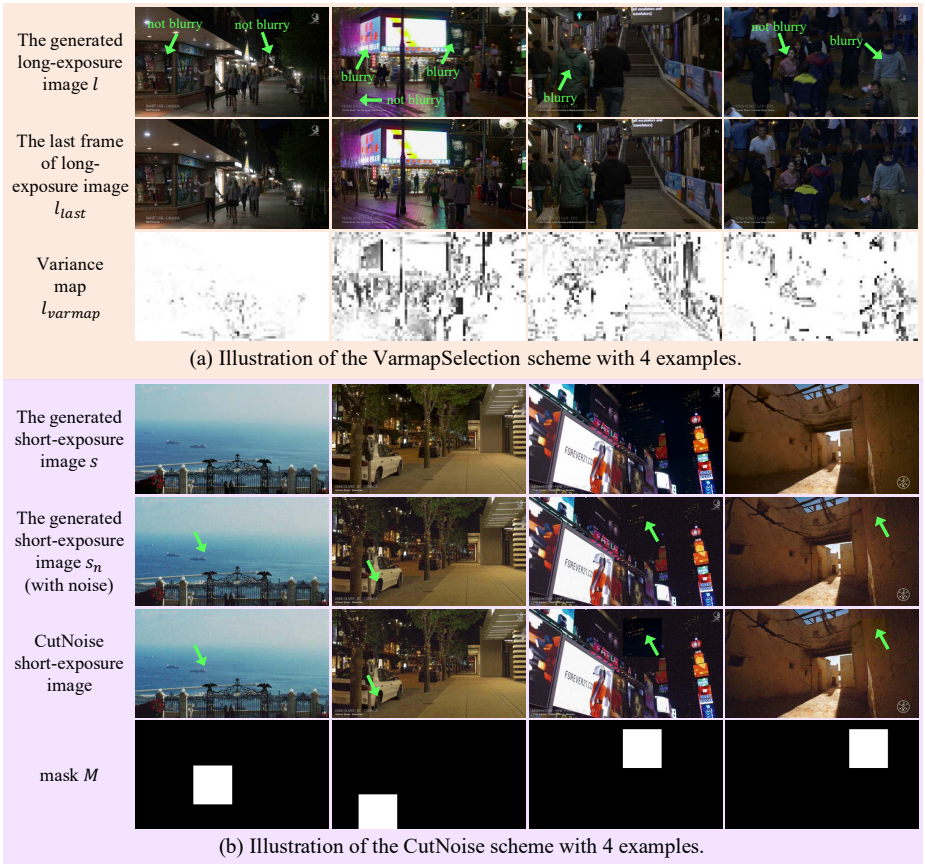



**Fig. 7.** Illustration of the VarmapSelection and CutNoise schemes. The variance maps can reflect the blurry regions or regions with large motions, e.g., the dark regions in $l_{varmap}$. The VarmapSelection is effective and robust to select blurrier training patches from the whole dataset. The CutNoise can be expressed as $s_n^{CutNoise} = M \odot s_{first} + (\mathbb{1} - M) \odot s_n$, where $\odot$ denotes matrix dot product and $\mathbb{1}$ is an all-1 matrix with the same dimension of the binary mask $M$.

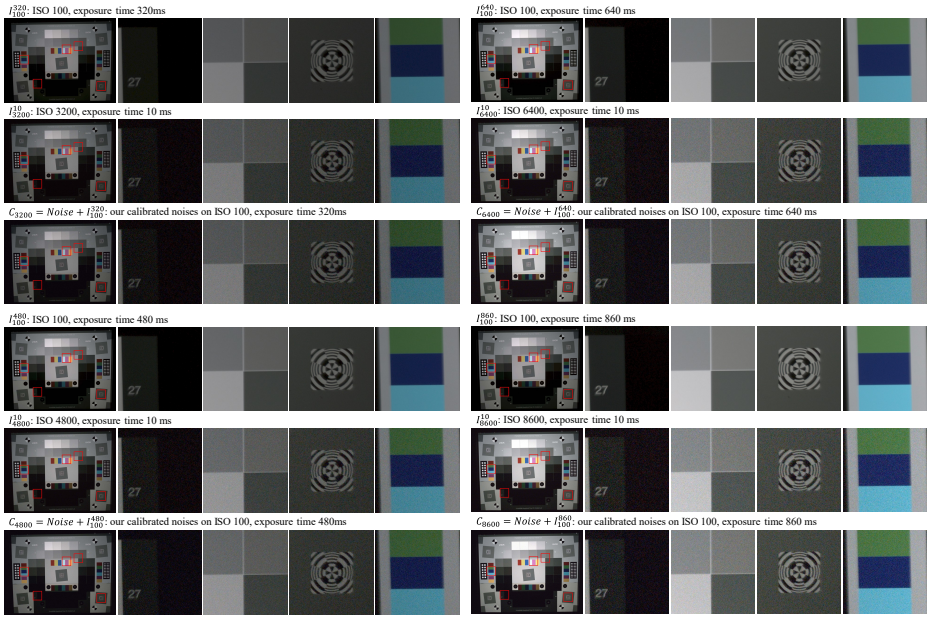



**Fig. 8.** Illustration of noise calibration results on ISO 3200, 4800, 6400, and 12800. There are 5 patches selected from the greyworld chart for readers to compare specific regions: dark region, checkerboard edges, round edges, color blocks. $I^*_{100}$ denote photos captured under ISO 100, which we assume there are almost no noises. $I^{10}_*$ are photos with real noises. $C_*$ are the addition of calibrated noises on clean images with specific ISO values, i.e., $I^*_{100}$. The overall brightness is generally equal for $I^*_{100}$, $I^{10}_*$, and $C_*$ since ISO$\times$exposure time is equal. Please compare the patterns of real and calibrated noises.

## References

1. Mildenhall, B., Barron, J.T., Chen, J., Sharlet, D., Ng, R., Carroll, R.: Burst denoising with kernel prediction networks. In: Proc. CVPR. pp. 2502–2510 (2018)
2. Wei, K., Fu, Y., Yang, J., Huang, H.: A physics-based noise formation model for extreme low-light raw denoising. In: Proc. CVPR. pp. 2758–2767 (2020)
3. Yoo, J., Ahn, N., Sohn, K.A.: Rethinking data augmentation for image super-resolution: A comprehensive analysis and a new strategy. In: Proc. CVPR. pp. 8375–8384 (2020)