A. SR-RAW training patch examples

SR-RAW is a diverse dataset that covers indoor and outdoor scenes under different illuminations. During training, we randomly cropped $64 \times 64$ patches from a full-resolution Bayer mosaic as input for training. Example training patches are shown in Figure 1.

B. Visual Analysis of CoBi

As detailed in Section 4, we analyze the percentage of features that are matched uniquely (i.e., bijectively) in nearest feature search when applying directly contextual loss to training. The percentage of target features matched with a unique source feature is only 43.7%, much less than the ideal percentage of 100%. By using the proposed contextual bilateral loss (CoBi), in Figure 2, we plot the number of unique feature matches with training step. On the x-axis, training step 0 denotes the starting point of the model trained with only the contextual loss. On the y-axis, training step 0 denotes the starting point of the model trained with CoBi.

Figure 2: Number of unique feature matches increases as training progresses, indicating a more diverse nearest neighbor feature field and thus potentially closer feature matching to ground truth.

C. Additional Qualitative 4X Results

Comparison with Baselines More results on 4X baseline comparisons are shown in Figure 5. We compare against LapSRN [4] that demonstrates SR models with a different network architecture; a model by Johnson et al. [3] that adopts perceptual losses for SR, and finally ESRGAN [5], the winner of the most recent Perceptual SR Challenge PIRM [1]. This extends Figure 4 in the main paper.

Comparison with Our Model Variants More results of controlled experiments with our model are shown in Figure 5. We compare our model trained on real sensor data with “Ours-png” – our model trained on processed RGB images, and “Ours-syn-raw” – our model trained on synthetic sensor data. We adopt the standard sensor synthesis model described in [2] to generate synthetic Bayer mosaics from 8-bit RGB images. This extends Figure 5 in the main paper.

Additional inference results (without ground truth) are shown in Figure 6.

D. Qualitative 8X Results

This extends Figure 6 in the main paper.

E. Generalization to Smartphone Sensor

Data Capture To obtain ground truth zoomed images for a smartphone that has limited optical zoom power, we use a DSLR with a zoom lens to obtain ground truth high-resolution images. As shown in Figure 3 A, the smartphone is stably mounted on top of a DSLR. For calibration, we capture a pair of images using the smartphone and DSLR with the same focal length, call them SP-Low and DSLR-Low, and then use the DSLR to take a zoomed image, call it DSLR-High by adjusting the focal length to the desired ratio.

Data Pre-processing We align the pair of images taken by the smartphone and DSLR using SP-Low and DSLR-Low.

Additional Qualitative Results This extends Figure 7 in the main paper.
Figure 1: Example training patch pairs of Bayer mosaic (Left) and ground truth RGB image (Right). Bayer mosaics, after packing, are $64 \times 64$ randomly cropped from a full-resolution sensor data in SR-RAW, and RGB images are corresponding $256 \times 256$ ground truth patch.
Figure 3: Smartphone-DSLR data capture and example data pair.

References


Figure 4: Our 4X zoom results show better perceptual performance in super-resolving distant objects against baseline methods that are trained under a synthetic setting and applied to processed RGB images.
Figure 4 (Cont.):
Figure 4 (Cont.):
Figure 5: Our model trained on real sensor data produces clean and high quality zoomed images than “Ours-png”, our model trained on processed 8-bit RGB images and “Ours-syn-raw”, our model trained on synthetic sensor data.
Figure 5: .
Figure 5 (Cont.).
Figure 6: More inference results on 4X compared against “Ours-png” and “Ours-syn-raw”.
Figure 7: Our model fine-tuned on a much smaller dataset can adapt to a Bayer mosaic variant from an iPhone X sensor.