

Zoom to Learn, Learn to Zoom

Supplementary Material

Xuaner Zhang
UC Berkeley

Qifeng Chen
HKUST

Ren Ng
UC Berkeley

Vladlen Koltun
Intel Labs

A. Training Patch Examples

SR-RAW is a diverse dataset that covers images of indoor and outdoor scenes under different illuminations. During training, we randomly crop 64×64 patches from a full-resolution Bayer mosaic as input. Target patches are 256×256 , correspondingly cropped from images taken with a long focal length. Example input-output training pairs are shown in Figure 3.

B. Visual Analysis of CoBi

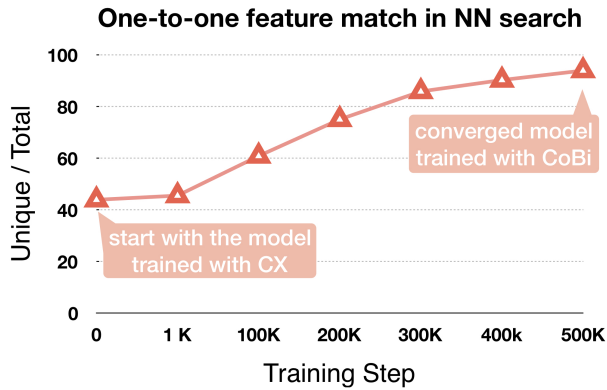


Figure 1: Number of unique feature matches increases as training progresses, indicating a more diverse nearest neighbor feature field and thus potentially closer feature matching to ground truth.

As described 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), we plot how the number of unique feature matches changes with training step in Figure 1. On the x-axis, training step 0 denotes the starting point of the model trained with only the contextual loss. The number of unique feature matches for our model trained with CoBi increases from 43.7% to 93.9% when training

converges. A high percentage of unique feature matches implies a more diverse nearest neighbor feature field and closer feature matching to ground truth.

C. Additional Qualitative 4X Results

Comparison with Baselines More results on 4X baseline comparisons are shown in Figure 4. 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 5 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 6 in the main paper.

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

D. Qualitative 8X Results

We use SR-RAW to train a 8X zoom model. We set patch size in Equation 3 to $n = 15$ and keep the training parameter and scheduling the same as described in the paper. Results are shown in Figure 7.

E. Generalization to Smartphone Sensor

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 2 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.

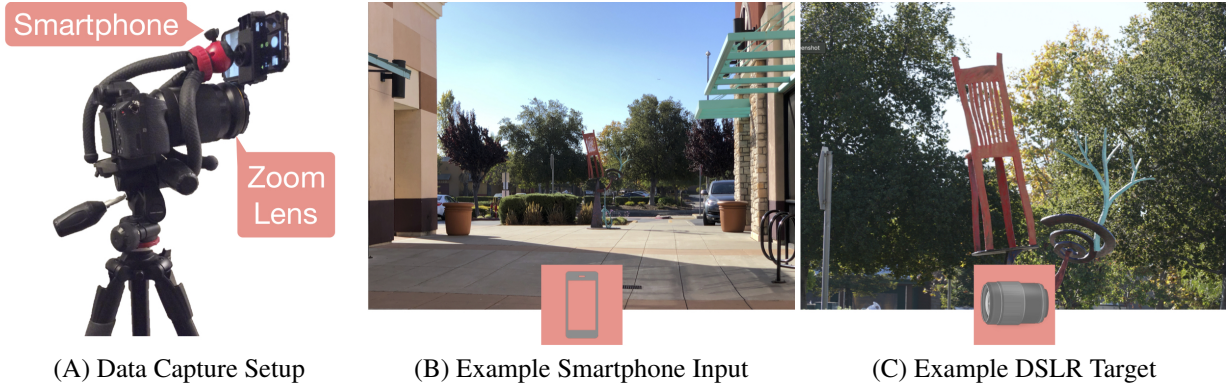


Figure 2: Smartphone-DSLR data capture and an example data pair.

We compute alignment between SP-Low and DSLR-Low to geometrically align the smartphone and the DSLR, and then apply field of view matching and geometric alignment between DSLR-Low and DSLR-High. These enable us to get aligned training data: input sensor data from the smartphone and target RGB image from the DSLR.

Figure 8 with more qualitative results extends Figure 7 in the main paper.

References

- [1] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor. 2018 PIRM challenge on perceptual image super-resolution. *arXiv*, 1809.07517, 2018. 1
- [2] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand. Deep joint demosaicking and denoising. *ACM Trans. on Graphics (TOG)*, 2016. 1
- [3] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 1
- [4] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017. 1
- [5] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang. ESRGAN: Enhanced super-resolution generative adversarial networks. *arXiv*, 1809.00219, 2018. 1

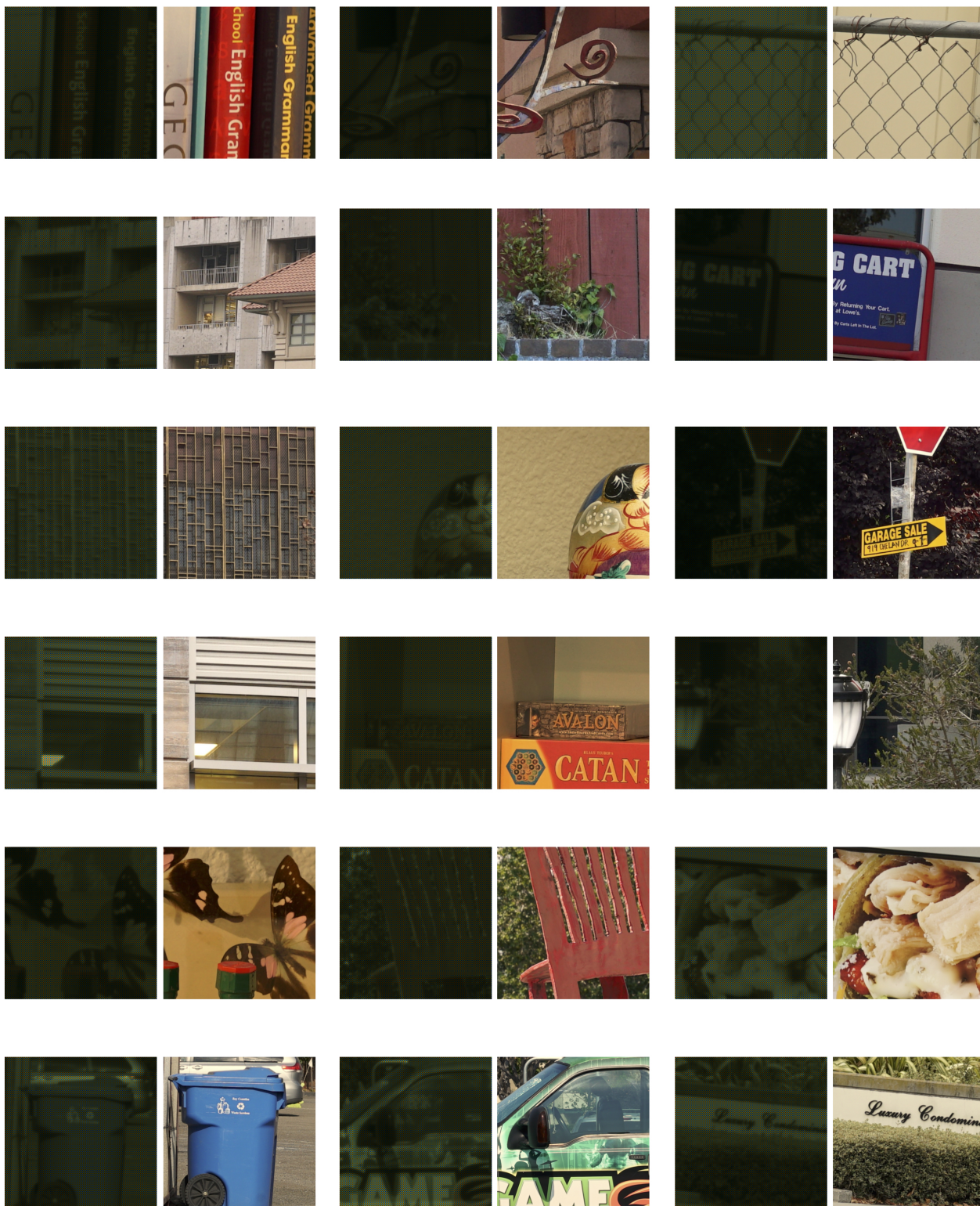


Figure 3: Example training pairs of Bayer mosaics and ground truth RGB images. Bayer mosaics are 64×64 patches randomly cropped from full-resolution sensor data, and RGB images are corresponding 256×256 patches cropped from RGB images.



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.): 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.

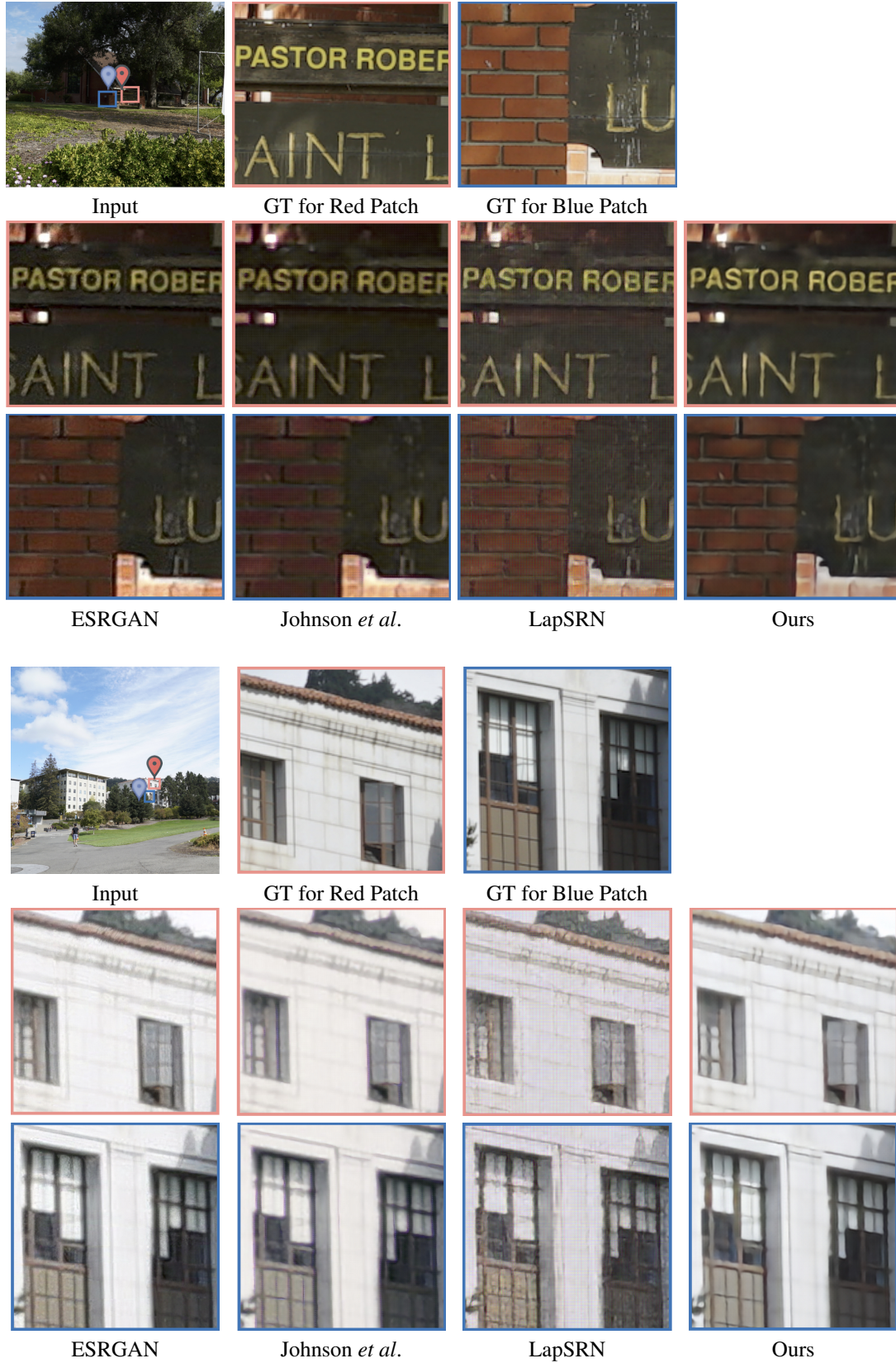


Figure 4 (Cont.): Our 4X zoom results show better perceptual performance in super-resolving distant objects against baseline methods that are trained under a synthetic setting and applied to processed RGB images.

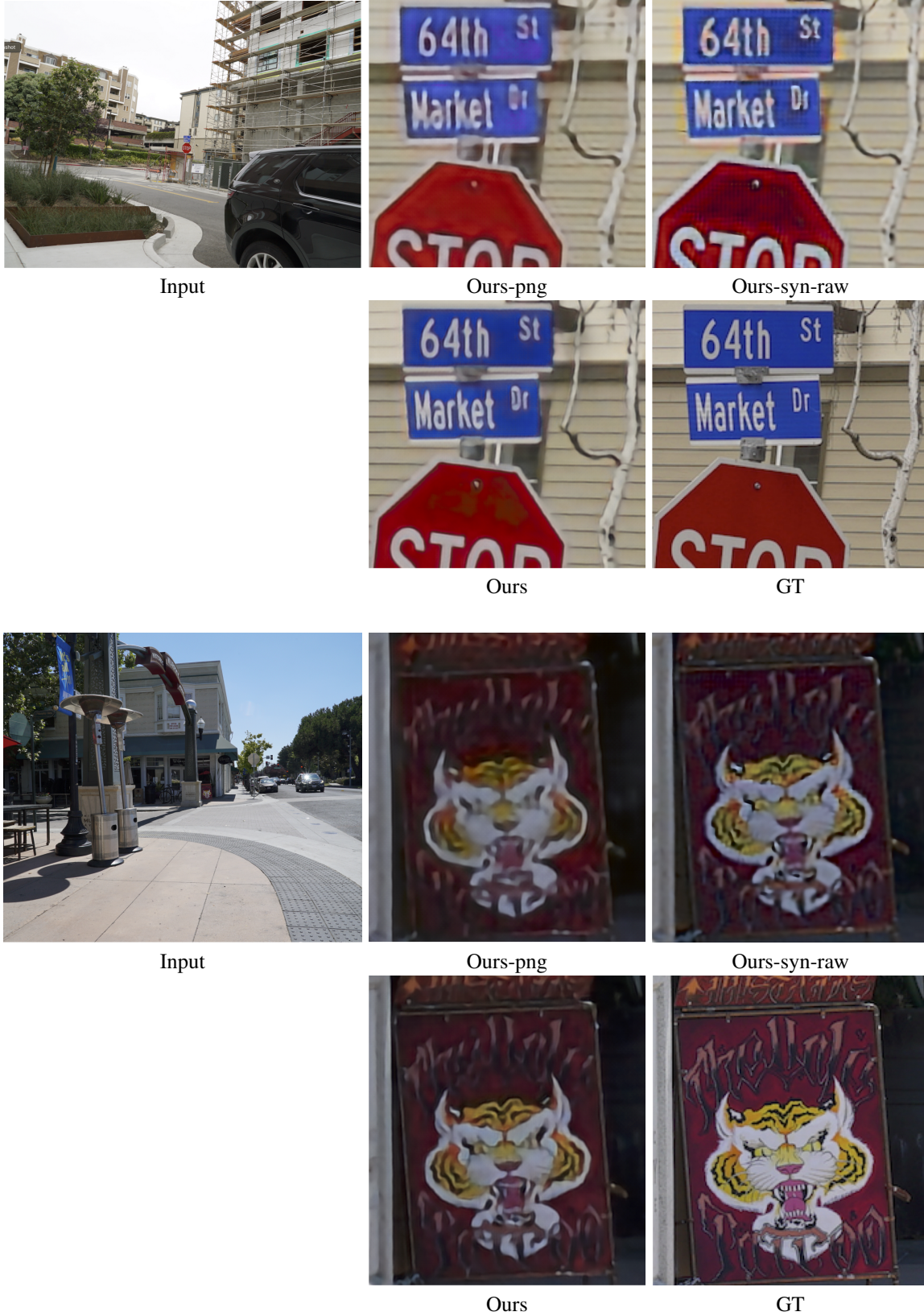


Figure 5: Results on 4X zoom show that “Ours-png”, our model trained on processed RGB images is limited in super-resolving details, and “Ours-syn-raw”, our model trained on synthetic sensor data often suffers from color and demosaic artifacts. Our model trained with real sensor data generates perceptually better results.



Input



Ours-png



Ours-syn-raw



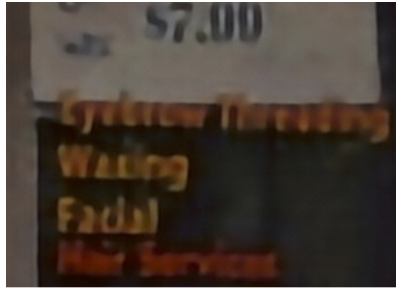
Ours



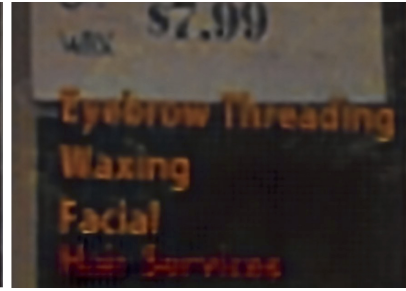
GT



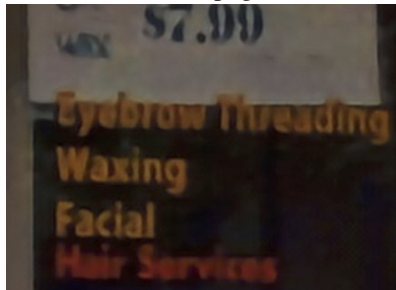
Input



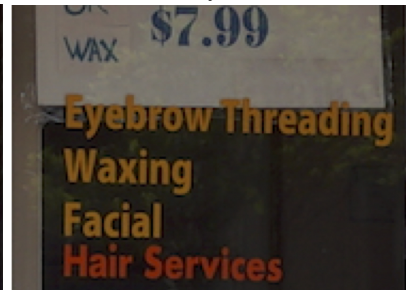
Ours-png



Ours-syn-raw



Ours



GT

Figure 5 (Cont.): Results on 4X zoom show that “Ours-png”, our model trained on processed RGB images is limited in super-resolving details, and “Ours-syn-raw”, our model trained on synthetic sensor data often suffers from color and demosaic artifacts. Our model trained with real sensor data generates perceptually better results.

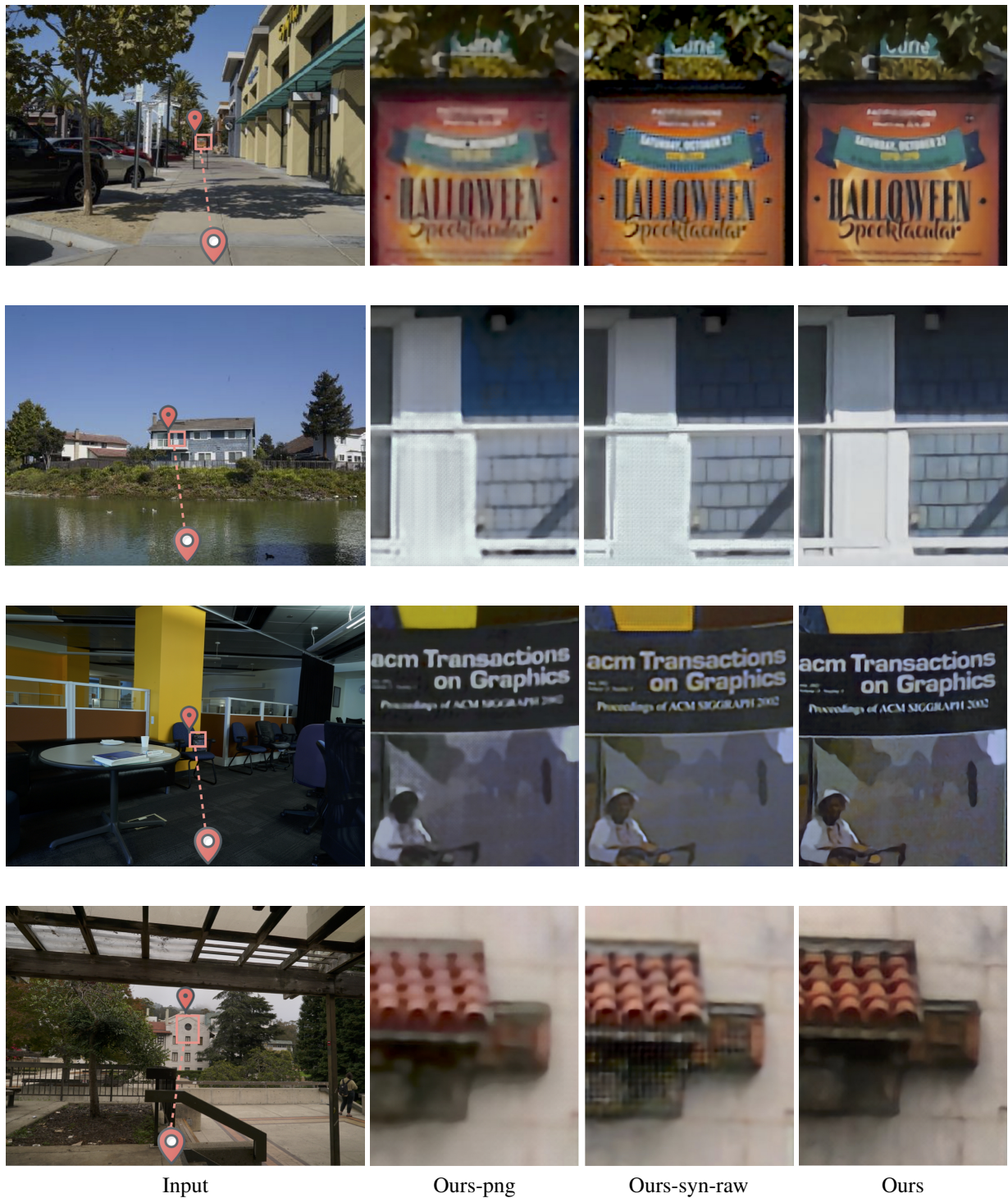


Figure 6: More inference results on 4X zoom compared against “Ours-png” and “Ours-syn-raw”.

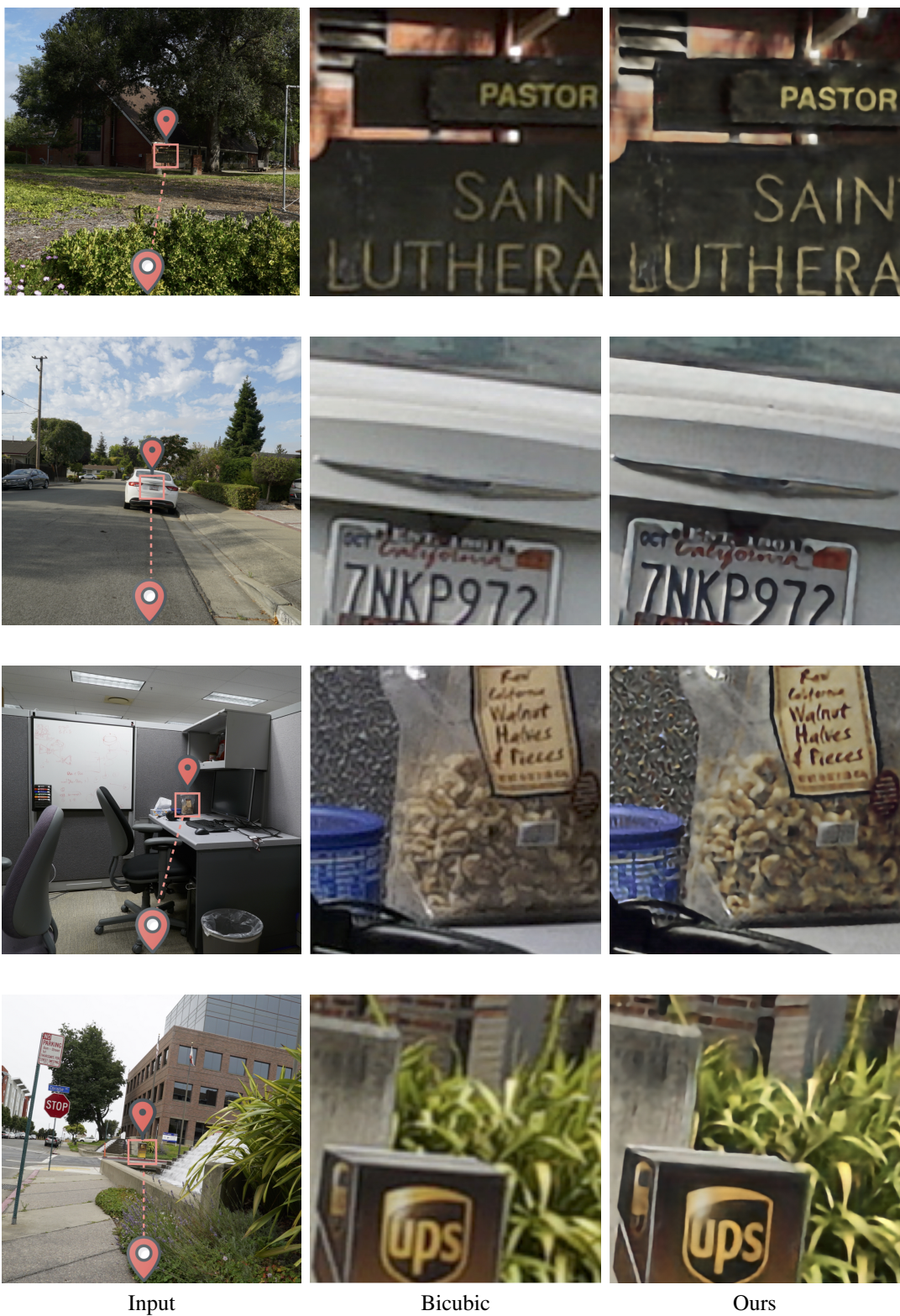


Figure 7: Results on our model trained for 8X zoom.



Figure 8: Results of 4X zoom show that our model fine-tuned on a much smaller dataset can adapt to a Bayer mosaic variant from an iPhone X sensor. Input is Bayer mosaic from iPhone X sensor data.