AdaDepth: Unsupervised Content Congruent Adaptation for Depth Estimation

Supplementary material

Introduction

In the supplementary material, we present additional qualitative results of the proposed *AdaDepth* approach for both NYUD and KITTI datasets. We also present architectural details of the two discriminators D_F and D_Y as introduced in the main paper.

Discriminator Architecture

In Table 1 and Table 2, we present architecture details of the two discriminator, D_Y and D_F respectively. For discriminator network D_Y, we follow Patch-GAN's [2] convolutional architecture with an input receptive field of size close to 80 × 80.

Layer	**Type**	**Filter Size**	**Filter Num**	**Stride**	**Output Size**
Input	-	-	-	-	$128 \times 160 \times 1$
c1	Conv	3×3	64	1	$128 \times 160 \times 64$
c2	Conv*	3×3	64	2	$64 \times 80 \times 64$
c3	Conv*	3×3	128	1	$64 \times 80 \times 128$
c4	Conv*	3×3	128	2	$32 \times 40 \times 128$
c5	Conv*	3×3	256	1	$32 \times 40 \times 256$
c6	Conv*	3×3	256	2	$16 \times 20 \times 256$
c7	Conv*	3×3	512	1	$16 \times 20 \times 512$
c8	Conv*	3×3	1024	2	$8 \times 10 \times 1024$
c9	Conv*	1×1	1024	1	$8 \times 10 \times 1024$
c10	Conv	1×1	1	1	$8 \times 10 \times 1$

Table 1: Network architecture of discriminator, D_Y applied on the single channel depth map. Padding is kept as "SAME" for all convolution layers. Conv* denotes standard convolutional layers followed by a batch-normalization with leaky-ReLU non-linearity.

Layer	**Type**	**Filter Size**	**Filter Num**	**Stride**	**Output Size**
Input	-	-	-	-	$8 \times 10 \times 2048$
layer-0	Conv	3×3	256	1	$8 \times 10 \times 256$
layer-1	Conv*	3×3	512	2	$4 \times 5 \times 512$
layer-1a	Conv*	1×1	128	1	$4 \times 5 \times 128$
layer-2	Conv*	3×3	512	2	$2 \times 3 \times 512$
layer-2a	Conv*	1×1	128	1	$2 \times 3 \times 128$
layer-3	Conv*	3×3	1024	2	$1 \times 2 \times 1024$
layer-3a	Conv*	1×1	256	1	$1 \times 2 \times 256$
layer-fc1	FC*	-	1024	-	1024
layer-fc2	FC	-	1	-	1

Table 2: Network architecture of discriminator, D_F applied on Res-5c activation map(input layer). Padding is kept as ”SAME” for all convolution layers. Conv* denotes standard convolutional layers followed by a batch-normalization with leaky-ReLU non-linearity.

Additional Qualitative Results

Figure 1 shows further qualitative results of our unsupervised depth adaptation approach *AdaDepth-U* along with the semi-supervised variant *AdaDepth-S* for NYUD dataset. In contrast to the existing fully-supervised approaches (Laina *et al.* [3] and Eigen *et al.* [1]), our semi-supervised approach achieves better performance by predicting depth values close to the ground-truth depth map, as can be seen in of Figure 1 and Figure 2.

For better visualization, Figure 2 shows the error-map of depth prediction with respect to the corresponding ground-truth. This clearly demonstrates the superiority of the proposed approach over the previous state-of-the-arts methods. The results of Eigen *et al.* [1] and Laina *et al.* [3] exhibit more error on major background regions as compared to our results. For an unbiased comparison, we also share some results based on our best, median and worst metric scores for images in Figure 3.

The qualitative results for KITTI dataset are shown in Figure 4. The test images are from the commonly used 697 test images from Eigen split [1]. As the ground-truth LIDAR data is very sparse, we interpolate the depth map for better visualization. For a fair comparison, we only show predictions for the cropped region following [1]. Compared to Zhou *et al.*, predictions from *AdaDepth-S* are sharper, preserve more local information and are closer to the ground-truth depth map.

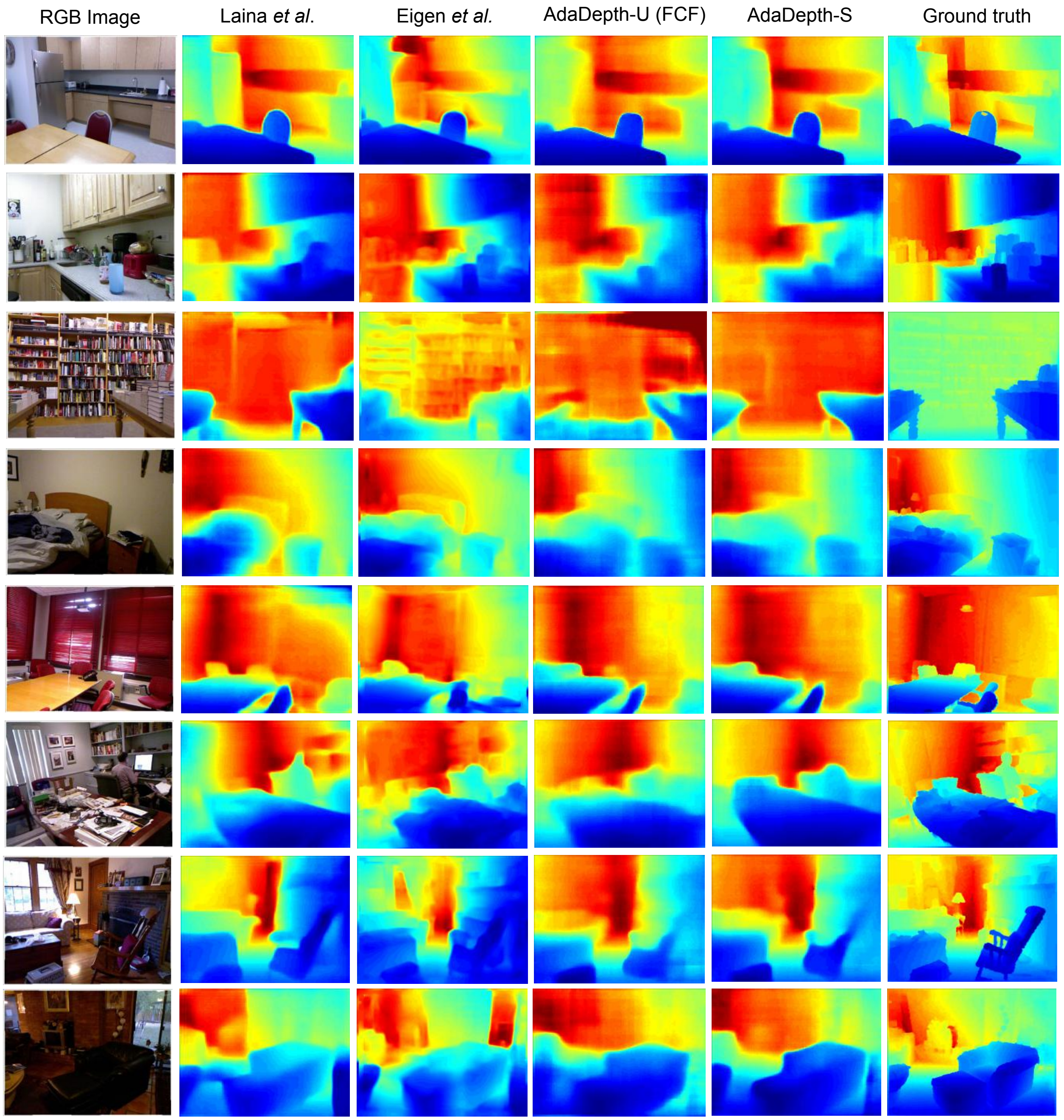

Figure 1: Qualitative comparison of *AdaDepth-U* and *AdaDepth-S* against fully-supervised Laina *et al.* [3] and Eigen *et al.* [1]. The results of Eigen *et al.* [1] exhibit low-level spatial details but fail to regress to ground-truth depth values (the upper part of sofa in 8^{th} row and the part of table towards the wall in 6^{th} row). Laina *et al.* on the other hand regress to better depth values generally but suffer with lack of precision on the edges (bed in row 4^{th}). Row-3(prediction on new scenes) and row-8 (low-contrast background) shows some of our failure cases.

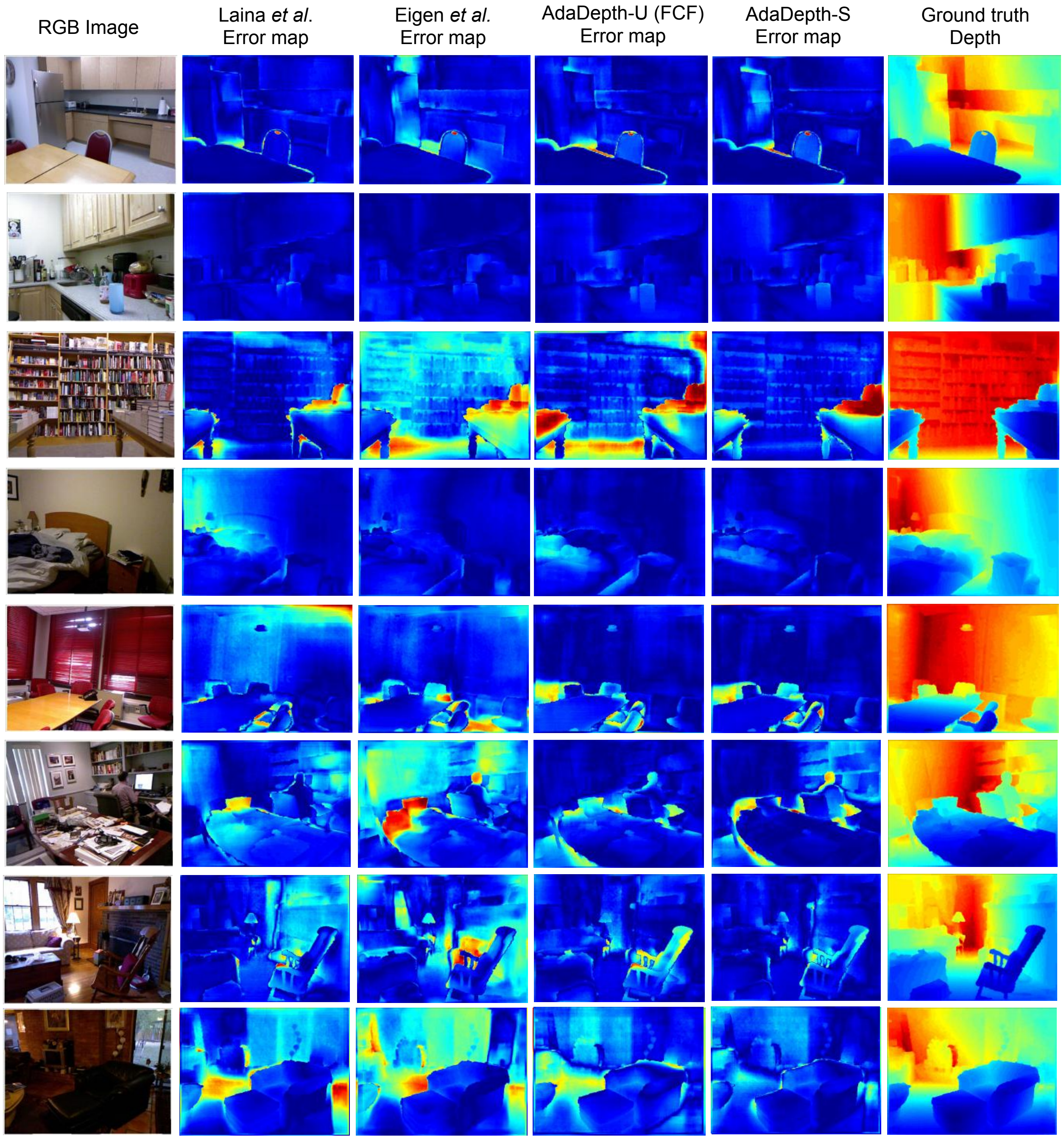

Figure 2: Qualitative comparison of *AdaDepth-U* and *AdaDepth-S* against Laina *et al.* [3] and Eigen *et al.* [1] from their corresponding error maps. The error-maps are computed as absolute difference between the prediction and corresponding ground-truth. For fair comparison the color scale on error map shows absolute error values from 0(blue) to 2(red) except for the last column, where it shows actual ground-truth depth.

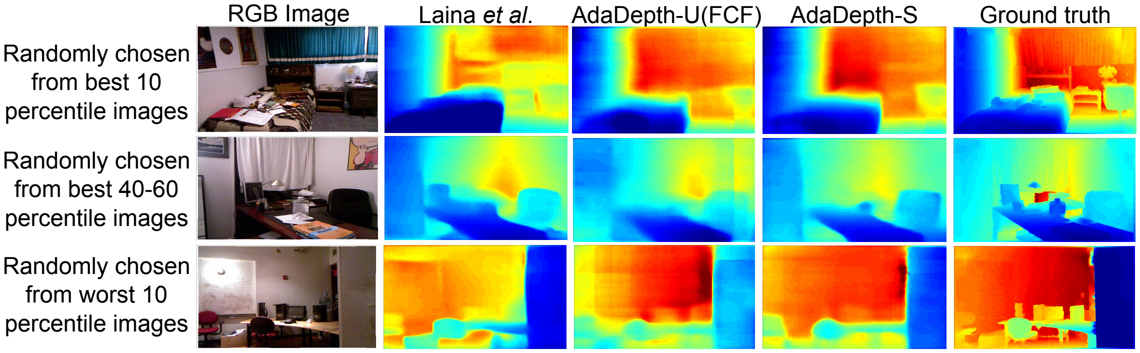
RGB Image
Laina et al.
AdaDepth-U(FCF)
AdaDepth-S
Ground truth
Randomly chosen from best 10 percentile images
Randomly chosen from best 40-60 percentile images
Randomly chosen from worst 10 percentile images

Figure 3: The percentile ranges are obtained by sorting results on NYUD test set as per the *rel* metric.

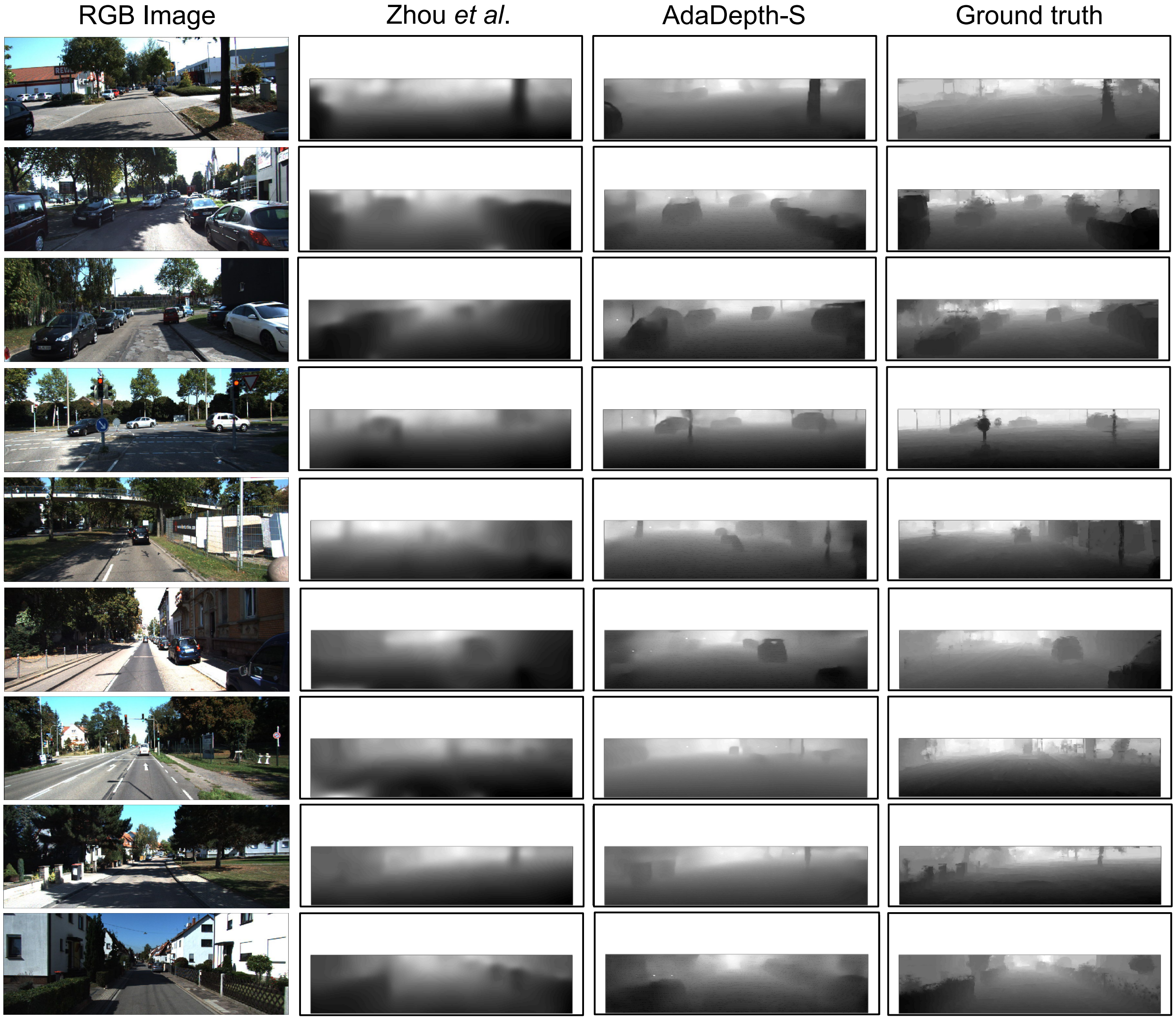

Figure 4: Qualitative comparison of *AdaDepth-S*, our semi-supervised adaptation approach against Zhou *et al.* [4]. Compared to Zhou *et al.*, predictions from *AdaDepth-S* are sharper, preserve more local information and are closer to the ground-truth depth map.

References

[1] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014.

[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[3] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, 2016.

[4] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.