Supplementary Material: Deep Stereo using Adaptive Thin Volume Representation with Uncertainty Awareness

Shuo Cheng[1*] Zexiang Xu[1*] Shilin Zhu[1]
Zhuwen Li[2] Li Erran Li[3,4] Ravi Ramamoorthi[1] Hao Su[1]
[1]University of California, San Diego [2]Nuro Inc. [3]Scale AI [4]Columbia University

Overview

In this Supplementary Material, we evaluate the uncertainty estimation with additional experiments, show the sub-networks of our network architecture in detail, and demonstrate our final point cloud reconstruction results of the DTU testing set and the Tanks and Temple dataset.

1. Additional experiments of uncertainty estimation.

In this section, we discuss additional experiments and analysis about our uncertainty estimation evaluated on the DTU validate set.

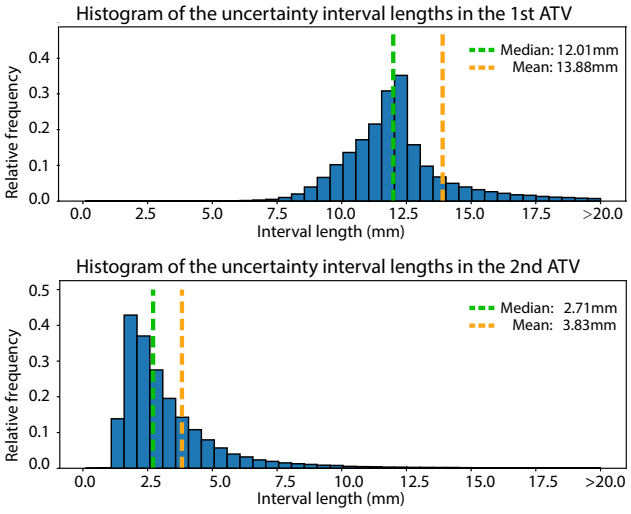

Figure 1: Histograms of the uncertainty interval lengths. We create bins for every 0.5mm to compute the histograms of the lengths of the uncertainty intervals in the two ATVs. We mark the median and the mean values of the lengths in the histograms.

We have shown the average lengths of the uncertainty intervals and the corresponding average sampling distances between the depth planes of the ATVs in Tab. 3 of the main paper. We now show the histograms of the uncertainty interval length in Fig. 1 to better illustrate the distributions of the interval length. We also mark the average lengths and the median lengths in the histograms. Note that, the distributions of the two ATVs are unimodal, in which most lengths distribute around the peaks; however, the average interval lengths differ much from the modes in the histograms, because of small portions of the intervals that have very large uncertainty. This means that using the average interval lengths – as what we do for Tab. 3 in the main paper – to dis-

[*] Equal contribution.

Layer	Stride	Kernel	Channel	Input
conv_unit0_0	1x1	3x3	3->8	rgb
conv_unit0_1	1x1	3x3	8->8	conv_unit0_0
conv_unit1_0	2x2	5x5	8->16	conv_unit0_1
conv_unit1_1	1x1	3x3	16->16	conv_unit1_0
conv_unit1_2	1x1	3x3	16->16	conv_unit1_1
conv_unit2_0	2x2	5x5	16->32	conv_unit1_2
conv_unit2_1	1x1	3x3	32->32	conv_unit2_0
conv_unit2_2	1x1	3x3	32->32	conv_unit2_1
conv_out1	1x1	1x1	32->32	conv_unit2_2
deconv_unit1_0	2x2	3x3	32->16	conv_unit2_2
concat1	-	-	-	deconv_unit1_0, conv_unit1_2
conv_unit3_0	1x1	3x3	32->16	concat1
conv_out2	1x1	1x1	16->16	conv_unit3_0
deconv_unit2_0	2x2	3x3	16->8	conv_unit3_0
concat2	-	-	-	deconv_unit2_0, conv_unit0_1
conv_unit4_0	1x1	3x3	16->8	concat2
conv_out3	1x1	1x1	8->8	conv_unit4_0

Table 1: The U-Net architecture of our multi-scale feature extractor. We show the detailed convolutional units of our multi-scale feature extractor; each convolutional unit is composed by a 2D convolution layer, a BN (batch normalization) layer and a ReLU layer. The colored cells (conv_out1, conv_out2, conv_out3) apply only a single 2D convolution layer to provide multi-scale features for cost volume construction.

Layer	Stride	Kernel	Channel	Input
conv_unit0	1x1x1	3x3x3	8->8	cost volume
conv_unit1	2x2x2	3x3x3	8->16	conv_unit0
conv_unit2	1x1x1	3x3x3	16->16	conv_unit1
conv_unit3	2x2x2	3x3x3	16->32	conv_unit2
conv_unit4	1x1x1	3x3x3	32->32	conv_unit3
conv_unit5	2x2x2	3x3x3	32->64	conv_unit4
conv_unit6	1x1x1	3x3x3	64->64	conv_unit5
deconv_unit7	2x2x2	3x3x3	64->32	conv_unit6
deconv_unit8	2x2x2	3x3x3	32->16	conv_unit4 + deconv_unit7
deconv_unit9	2x2x2	3x3x3	16->8	conv_unit2 + deconv_unit8
conv_out	1x1x1	3x3x3	8->1	conv_unit0 + deconv_unit9

Table 2: The network architecture of the 3D U-Net. We show the 3D U-Net architecture that is used to process the cost volume and predict the depth probabilities at each stage. Similarly, each convolutional unit is composed by a 3D convolution layer, a BN (batch normalization) layer and a ReLU layer. The colored cell (conv_out) apply only a single 3D convolution layer. We apply soft-max on the final one-channel output over depth planes to compute the final depth probability maps.

cuss the depth-wise sampling is in fact underestimating the sampling efficiency we have achieved for most of the pixels, though our average lengths are good and correspond to a high sampling rate. Therefore, we additionally show the median values in the histograms, which are less sensitive to the large-value outliers and are more representative than the mean values for these distributions. As shown in Fig. 1, the median interval lengths of the two ATVs are 12.01mm and 2.71mm respectively, which are closer to the peaks of the histograms; these lengths correspond to depth-wise sampling distances of 0.38mm and 0.34mm, given our specified 32 and 8 depth planes. These are significantly higher sampling rates than previous works, such as MVSNet [3] – which uses 256 planes to obtain a sampling distance of 1.99mm – and RMVSNet [4] – which uses 512 planes to obtain a sampling distance of 0.99mm. Our ATV allows for highly efficient spatial partitioning, which achieves a high sampling rate with a small number of depth planes.

To illustrate how the per-pixel uncertainty estimates vary in a predicted depth map, we show the pixel-wise difference between the ground truth depth and the boundaries of the uncertainty intervals in Fig. 2. We can see that, while our estimated uncertainty intervals have small lengths (as shown in Fig. 1), the uncertainty estimation is very reliable, reflected by the fact that most intervals are covering the ground truth depth in both ATVs (the red and white colors in the right two columns of Fig. 2) . This verifies the high average covering ratios of 94.7% and 85.2% of the two ATVs, which we have shown in Tab. 3 of the main paper. We also observe more white colors in the 3rd-stage ATV than those in the 2nd-stage ATV, which reflects that the uncertainty is well reduced after a stage and the prediction becomes more precise. Note that, while our method may predict incorrect intervals (blue colors in the right two columns of Fig. 2) that fail to cover the ground truth for some pixels, those pixels are mostly around the shape boundaries, oblique surfaces and highly textureless regions, which are known to be challenging and still open problems for depth estimation. On the other hand, our method predicts large uncertainty for these challenging pixels, which is as we expect and reflects the inaccuracies in the predictions.

2. Network architecture.

We have shown the overview of our network in Fig. 2 of the main paper and discussed our network in Sec. 3 of the paper. Our network consists of a 2D U-Net for feature extraction and three 3D U-Nets with the same architecture for cost volume processing. We show the details of the 2D U-Net in Tab. 1, which is used for our multi-scale feature extractor (see Sec. 3.1 of the paper); we also show the details of our 3D U-Net in Tab. 2 which is used to process the cost volume at each stage (see Sec. 3.3 of the paper).

3. Point cloud reconstruction.

We show our final point cloud reconstruction results of the DTU testing set [1] in Fig. 3 and Fig. 4, and the results of the Tanks and Temple dataset [2] in Fig. 5. Please refer to Tab. 1 and Tab. 2 in the main paper for quantitative results on these datasets.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 2

[2] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):78, 2017. 2

[3] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 2

[4] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. *arXiv preprint arXiv:1902.10556*, 2019. 2

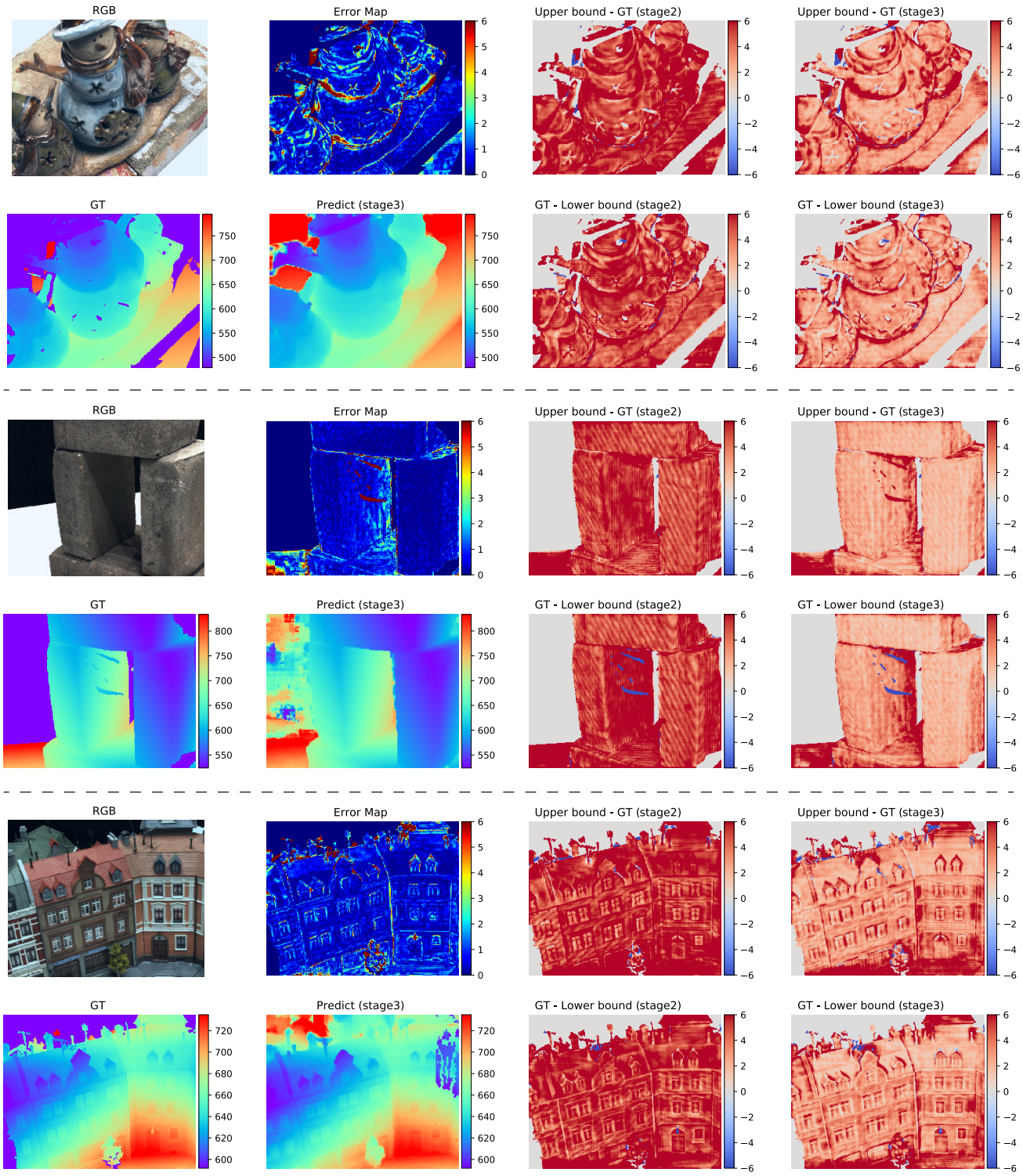

Figure 2: Uncertainty in depth predictions. We show three examples from the DTU validate set regarding the depth predictions and their pixel-wise uncertainty estimates. In each example, we show the reference RGB image, the ground truth depth, the depth prediction and a corresponding error map; we also illustrate the uncertainty intervals by showing the difference between the ground truth depth and the interval boundaries (lower bound and upper bound). Note that, in the right two columns, the white colors represent small intervals with low uncertainty, the red colors represent large intervals with large uncertainty, and the blue colors correspond to the intervals that fail to cover the ground truth.

Figure 3: Point cloud reconstruction on the DTU test set.

Figure 4: Point cloud reconstruction on the DTU test set.

Figure 5: Point cloud reconstruction on the Tanks and Temple dataset.