Unsupervised Domain Adaptation for ToF Data Denoising with Adversarial Learning

Gianluca Agresti
University of Padova
agrestig@dei.unipd.it

Henrik Schaefer
Sony Europe B.V.
henrik.schaefer@sony.com

Piergiorgio Sartor
Sony Europe B.V.
sartor@sony.de

Pietro Zanuttigh
University of Padova
zanuttigh@dei.unipd.it

Additional Material

This additional material document contains two sections. The first describes the datasets acquired for the training of the deep network and for the evaluation of the paper results. The second presents some additional visual results which did not fit in the paper for space limitations.

1. Real World Datasets

Table 1 shows the 5 different datasets used for training and experimental evaluation in the paper. Note that the datasets S_1 and S_4 come from [1], while the datasets S_2,S_3 and S_5 have been acquired ad-hoc for this work. Here we focus on the new datasets introduced in the paper:

1. the unlabeled real dataset S_2 used for the proposed unsupervised domain adaptation;
2. the labeled real dataset S_3 used for validation;
3. the labeled real dataset S_5, i.e. the *box* dataset, used for experimental evaluation.

We will detail their characteristics and present some examples of the contained data. The datasets are available online at http://lttm.dei.unipd.it/paper_data/MPI_DA_CNN .

Dataset	Type	GT	# scenes	Used for
S_1	Synth	Yes	40	Supervised train
S_2	**Real**	**No**	**97**	**Adversarial train**
S_3	**Real**	**Yes**	**8**	**Validation**
S_4	Real	Yes	8	Testing
S_5	**Real**	**Yes**	**8**	**Testing**

Table 1: Datasets exploited in the paper. The datasets S_2, S_3 and S_5 have been acquired ad-hoc for this work.

The real datasets S_2, S_3 and S_5 have been acquired with a SoftKinectic ToF camera. Before starting the acquisitions, the ToF camera has been calibrated in order to remove the wiggling (cyclic) error from the depth and amplitude images. In the paper, we used ToF data acquired at 20, 50 and 60 MHz. The data used by the proposed method has been phase unwrapped by using the multi-frequency information, in order to have the maximum unambiguous rage equal to 15 m. The resolution of the ToF depth and amplitude images is 320×239 px. Note that all the 3 datasets contain structures originating MPI. In the following of this section we are going to explain the specific characteristics of each of the 3 datasets.

1.1. Unlabeled Real Dataset S_2

The unlabeled real dataset S_2 is composed by scenes captured in a office environment in uncontrolled light conditions (ambient light was present). The acquisitions frame static scenes containing tables, chairs, lockers and many other different objects that can be found in a office. The dataset contains 97 recorded scenes, and for each of them the calibrated depth and amplitude images have been stored. The depth values are in the range from 0.5 to 6 m. Figure 2 shows some examples of depth and amplitude images of sample scenes in S_2.

1.2. Real Validation Set S_3

The subjects of the recordings from the real dataset S_3 are static scenes containing puppets, small boxes, wooden corners and polystyrene cones and spheres. The recorded depth images are in the range between 0.5 and 2 m. The depth ground truth of the ToF acquisitions has been generated with an active stereo system registered with the ToF camera. First, the ToF camera captures the scene, then a standard light projector illuminates it with a series of phase shifted patterns while the stereo system is recording. In this way, we can uniquely label the scene points observed by each single row of the stereo cameras, helping the triangulation operation and obtaining an accurate depth estimation. We used high frequency sinusoidal patterns in order to reduce the distortion due to diffuse reflection, as also suggested by Gupta et al. in [2].

Finally, this ground truth depth map is projected on the ToF sensor.

The acquired ground truth is not ideal, but it is still far more precise than ToF acquisitions. As an example, we compared a captured V grove with a synthetic 3D model

of it. Fig. 1 contains the ToF amplitude image and the 3D synthetic model of the corner. After the registration of the 2 depth fields, the mean absolute error (MAE) of the captured ground truth corner is 0.9 mm, about 100 times more accurate than the ToF acquisition with MPI.

The ground truth acquisition is time consuming and requires a properly calibrated system. This makes quite impractical to build big and various enough datasets as the ones required to properly train complex deep networks. This is one of the main motivations for the unsupervised domain adaptation approach proposed in the paper. Figure 3 shows the depth and amplitude images captured at 60 MHz of all the 8 scenes contained in S_3.

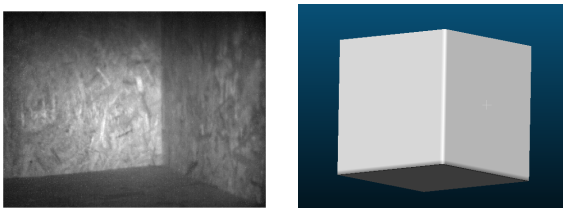

Figure 1: On the left, the ToF amplitude image of the corner used for the evaluation of the ground truth accuracy. The ideal 3D model is on the right.

1.3. Real Test Set S_5

The subjects of the recordings from the real dataset S_5 are static scenes containing boxes of various shapes and dimensions. We decided to create this *box* dataset since ToF sensors can be used in logistics and manufacturing for inspection, handling and dimensioning of box-shaped objects and we would like to evaluate which are the performance of our approach in this scenario. The dataset also contains the ground truth depth maps related to the ToF acquisitions, that have been acquired with the approach described in the previous section. Figure 4 shows the depth and amplitude images captured at 60 MHz of all the 8 scenes contained in S_5.

2. Qualitative Analysis

We show the qualitative results of our method on all the scenes of datasets S_4 and S_5 in Figure 5 and 6 respectively. The figures contain the ground truth, the input data with the corresponding error maps, the output of the proposed approach with its error maps and finally the error maps relative to the method of [1] (that is the best among the competitors according to the evaluation in Section 8 of the paper). The depth and error maps have been visualized only in the pixels where the depth ground truth is available (the points for which no ground truth is available are labeled with a dark blue color). Note that [1] exploits a denoiser CNN similar to the generator of our method. The key difference is the unsupervised domain adaption:

- in [1] the training process is carried on labeled synthetic data only;
- in our method instead, we are using an unsupervised domain adaptation in which we use the same synthetic dataset of [1] and the new unlabeled real dataset S_2.

As it is possible to note by looking at the error images (Figures 5 and 6), the depth over-estimation (red regions) due to MPI is strongly reduced in the output of the proposed method if compared with the raw input data and the error maps of [1]. There is still some MPI corruption on the floor of the scenes, but the improvement is clear when compared with the other approaches. Furthermore, the proposed method is also more accurate than [1] on the depth edges, e.g. on the border of the sphere and of the head and on the details of the deer respectively on the third, fifth and sixth row of Figure 5. For additional comments and the quantitative analysis of the performances, please look at Section 8 of the paper.

References

[1] G. Agresti and P. Zanuttigh. Deep learning for multi-path error removal in tof sensors. In *Geometry Meets Deep Learning ECCV Workshop*, 2018. 1, 2, 6, 7

[2] Mohit Gupta and Shree K Nayar. Micro phase shifting. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 813–820. IEEE, 2012. 1

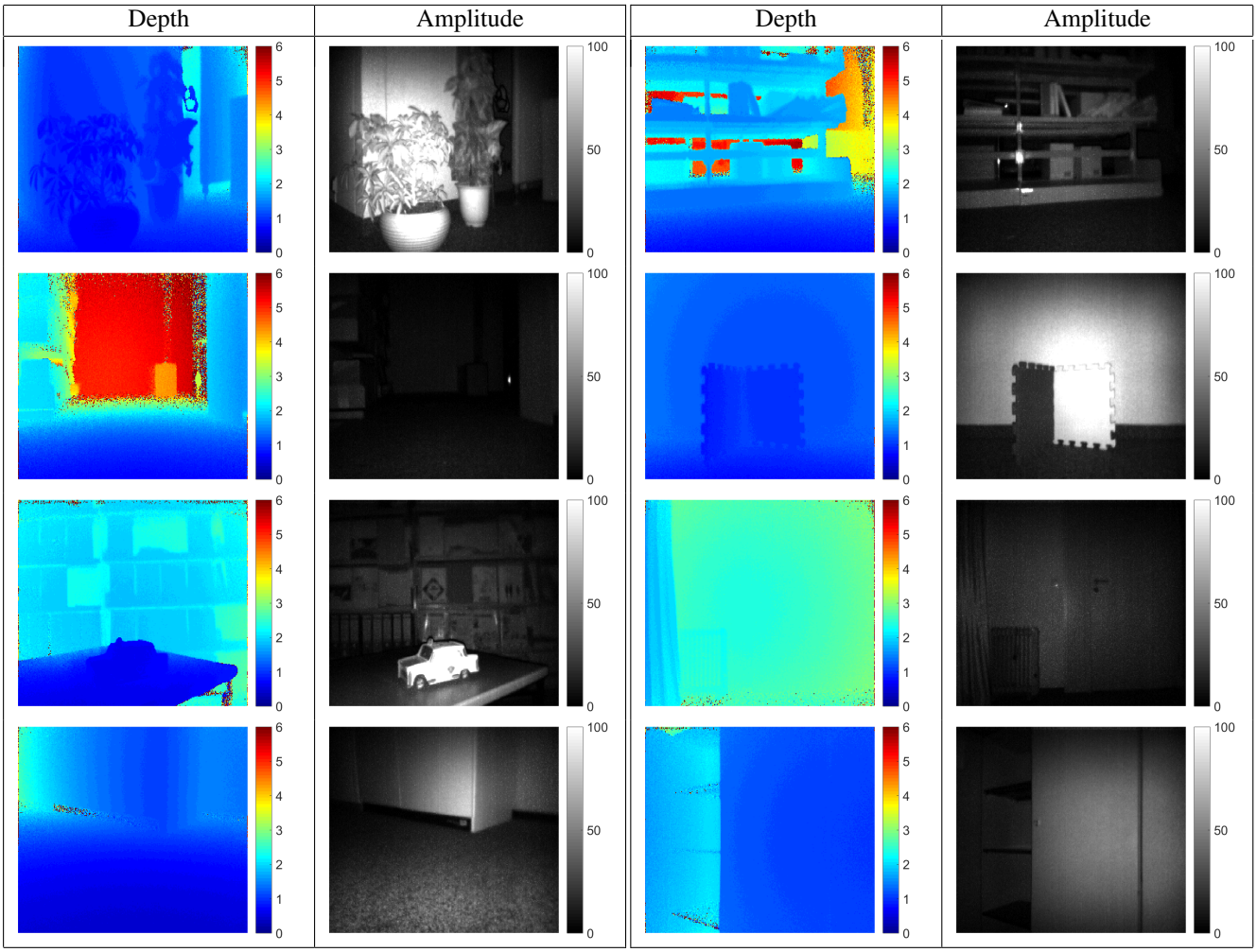

Figure 2: Representation of some of the ToF recordings contained in the S_2 dataset. Here we show the depth and amplitude images captured at 60 MHz. The depth values are measured in meters.

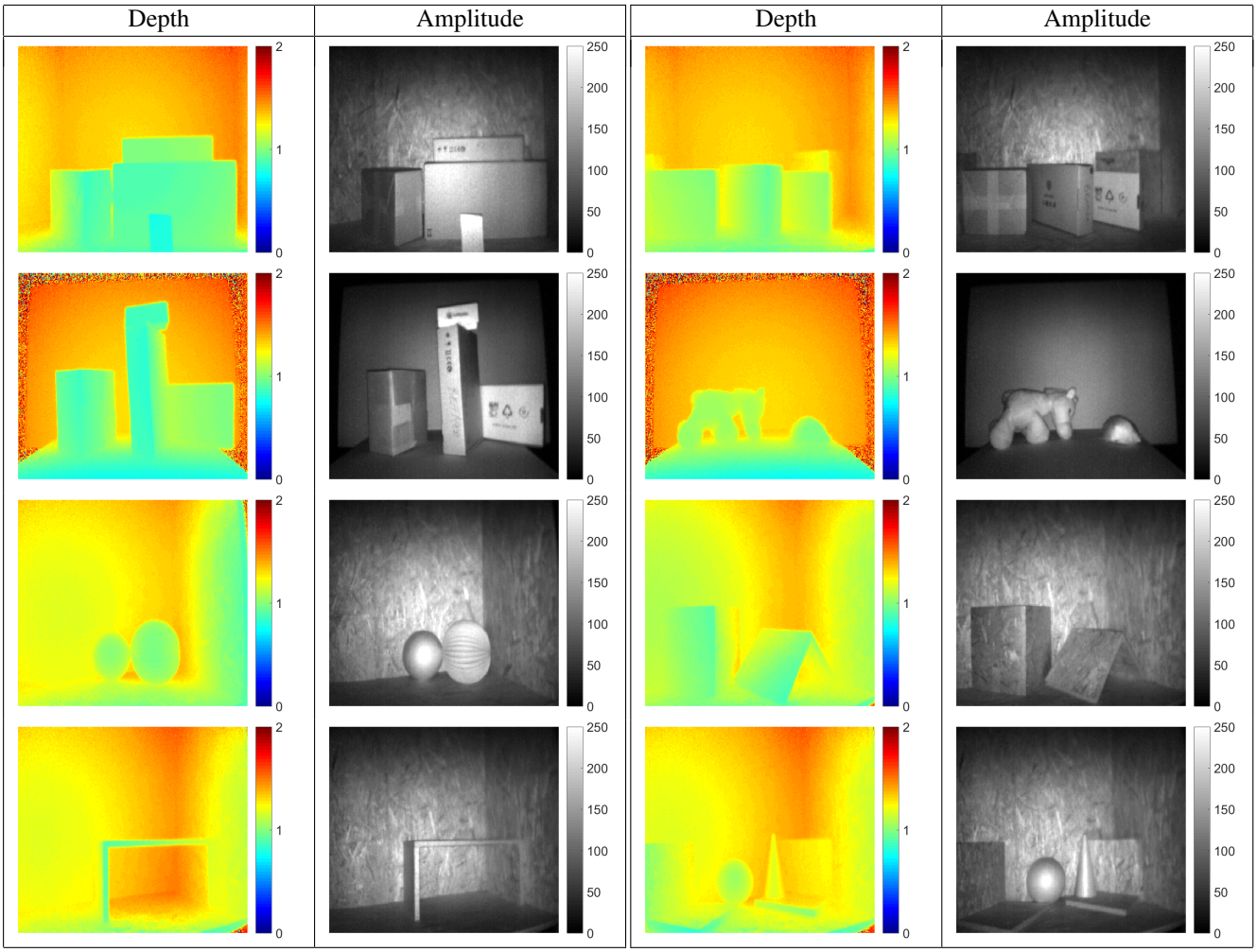

Figure 3: Representation of the ToF recordings contained in the S_3 dataset. Here we show the depth and amplitude images captured at 60 MHz. The depth values are measured in meters.

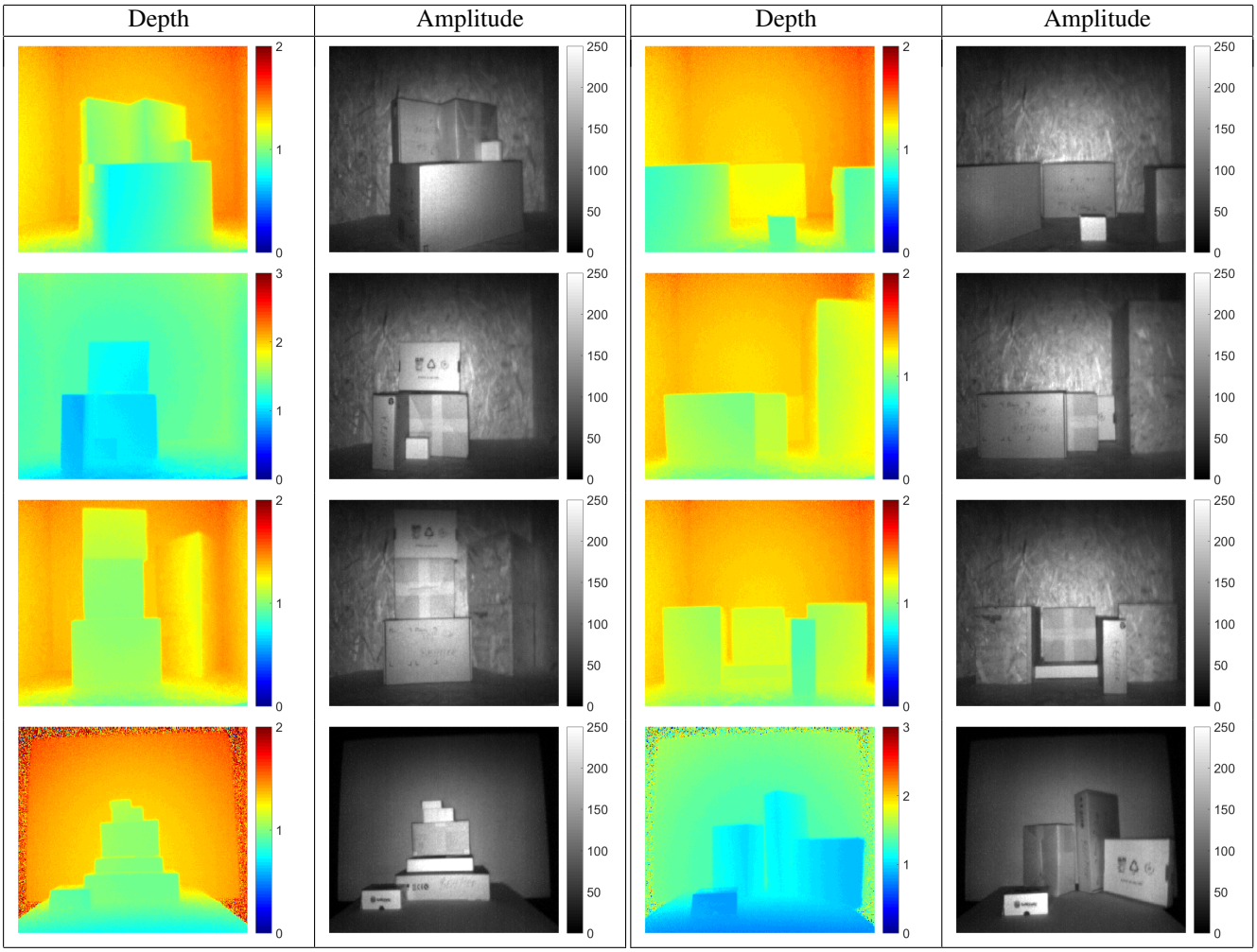

Figure 4: Representation of the ToF recordings for the S_5 datasets. Here we show the depth and amplitude images captured at 60 MHz. The depth values are measured in meters.

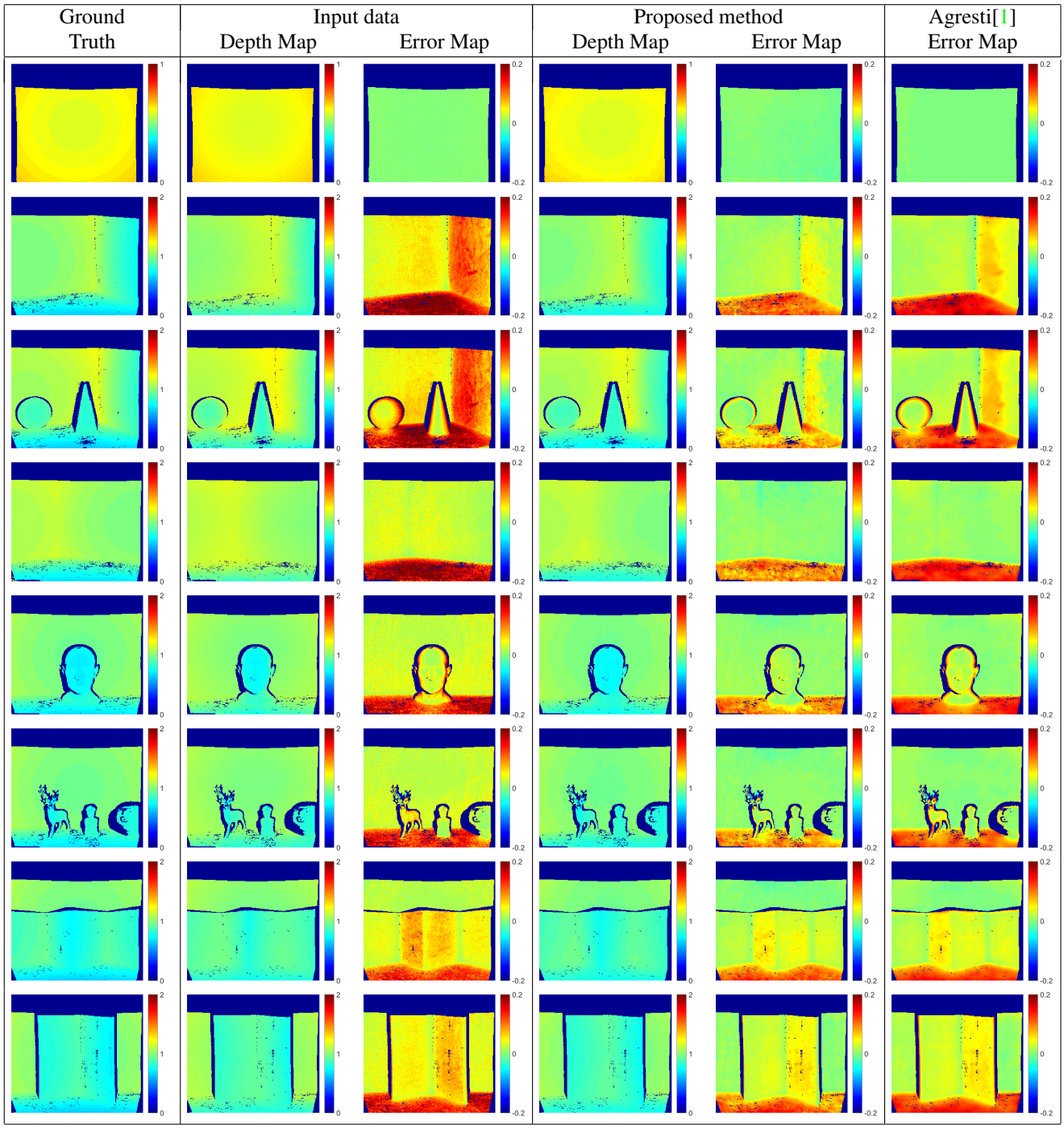

Figure 5: Comparison between the input depth at 60 MHz, the proposed method and the approach presented by Agresti et al. in [1]. The figure shows the computed depth and error maps for the scenes extracted from the dataset S_4. The values are measured in meters.

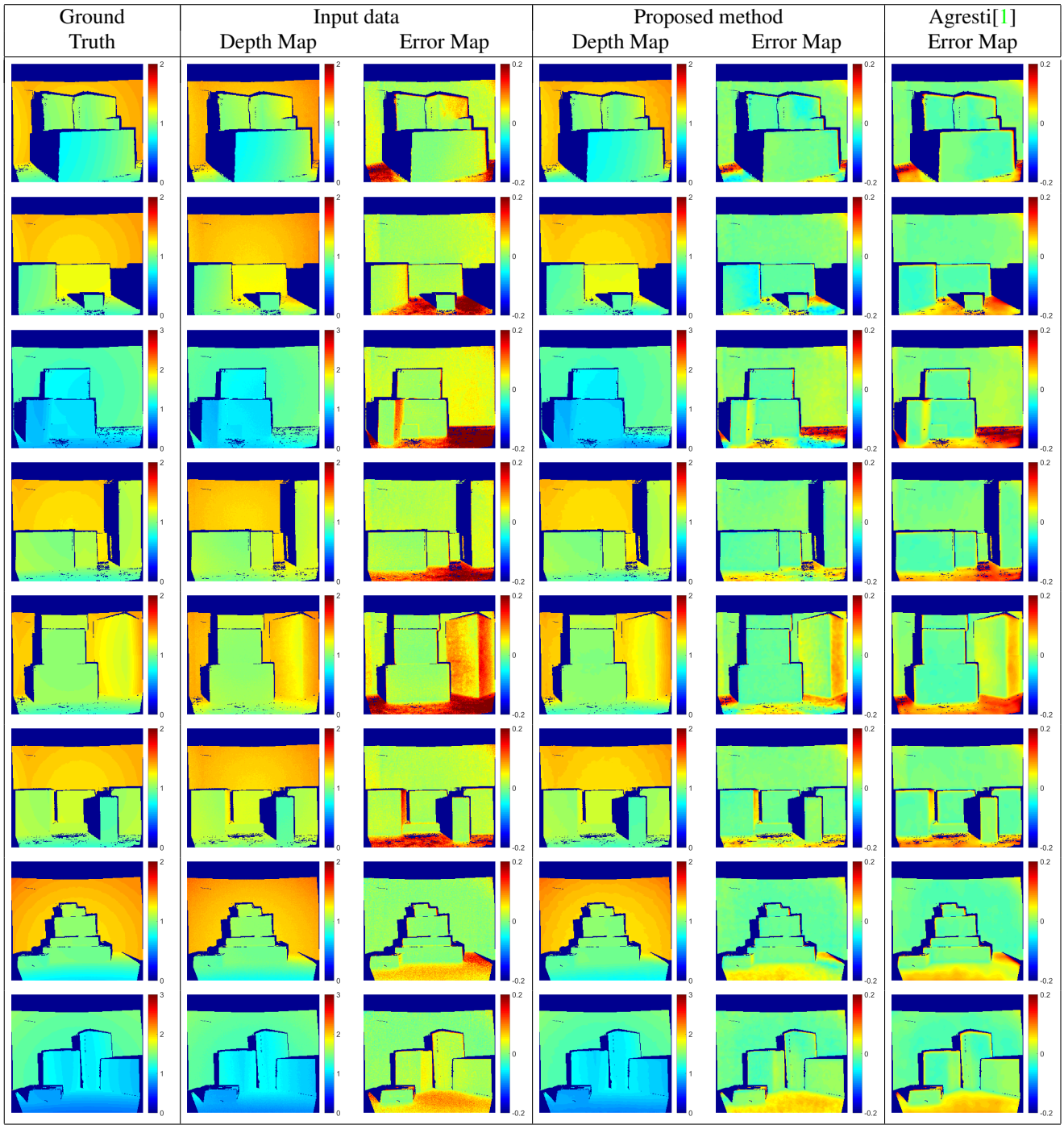

Figure 6: Comparison between the input depth at 60 MHz, the proposed method and the approach presented by Agresti et al. in [1]. The figure shows the computed depth and error maps for the scenes extracted from the dataset S_5. The values are measured in meters.