X-Section: Cross-section Prediction for Enhanced RGB-D Fusion. Supplementary Materials.

Multi-Frame Fusion Results

This section expands further the multi-frame fusion experimental analysis. We obtain a ground truth approximation by computing a solid voxelisation of the scene. Cad models and objects poses are given by the dataset. Since it is not possible to obtain the ground truth of the environment (such as tables, walls and monitors), we mask the depth using segmentation information to reconstruct objects only.

We run the pipeline for the first 50 frames and fuse them. The resulting reconstruction is compared with the ground truth. Finally, we compute Intersection over Union, precision and recall for every frame and report them in Figure 2. The reconstructions are shown in Figure 3 and 4.

Overall, the results show that in all scenes we recover more information than depth only fusion. For most sequences the precision of the reconstruction using thickness is close to the one obtained by fusion of only depth. This means that our propose method recovers correctly the information and does not hallucinate details that are not present. However, there are two in which X-Section has a significantly lower precision than depth only fusion. Sequences *0049* and *0057* result particularly difficult to reconstruct.

As shown in the second row of Figure 3, among the objects in *0049* there is an upside-down bowl. In this case the network struggles to generate a consistent prediction outputting a different internal shape as expected. An detail of the reconstruction of the scene is reported in Figure 1. The same object is present also in sequence *0053* in which our method has really high precision and does not present any artefact. Since the network has been trained using random position of cameras and objects, it should be robust to changes in the point of view. Among the possible causes of this failure there is the occlusion between objects and the lack of context at training time.

The analysis above highlights one of the current drawbacks of the proposed approach. Isolating objects simplifies the task and yields effective generalisation to unseen views. However, this introduces a compromise in terms of robustness due to the lack of context. The impact remains limited, and tackling this issue is left as possible improvement.

Another problematic sequence is *0057*. The sources of error are twofold. One is a clear over estimation of the thickness of the objects (also present in *0049* as Figure 1 shows). The other is the presence of artefacts around the shape of the objects. These recur often in the dataset and they are typical of dense reconstruction systems. In the case of the proposed pipeline they are caused by drift in the pose and imprecise segmentation at the edges of object. The dataset used for evaluation has camera poses associated to each frame, but they are the result of global optimisation. This causes a pose drift that affects all the scenes. Both segmentation accuracy and pose drift are currently open research questions and the integration of more advanced methods addressing this issues will improve the performances of our method.

Figure 1: Reconstruction of Sequence *0049* using predicted thickness. Highlighted in blue the upside down bowl that causes the network to fail. On the right the corresponding object is segmented in the same colour.

In summary, the experiments show a measurable improvements in the estimation of the occupied space. The results also show that breaking down the problem produces good generalisation to both novel views and novel objects (see main body of the work for this) at the cost of loosing contextual information and of a compromise in terms of stability. However, we measured a marginal impact of the drawbacks on the validation sequences.

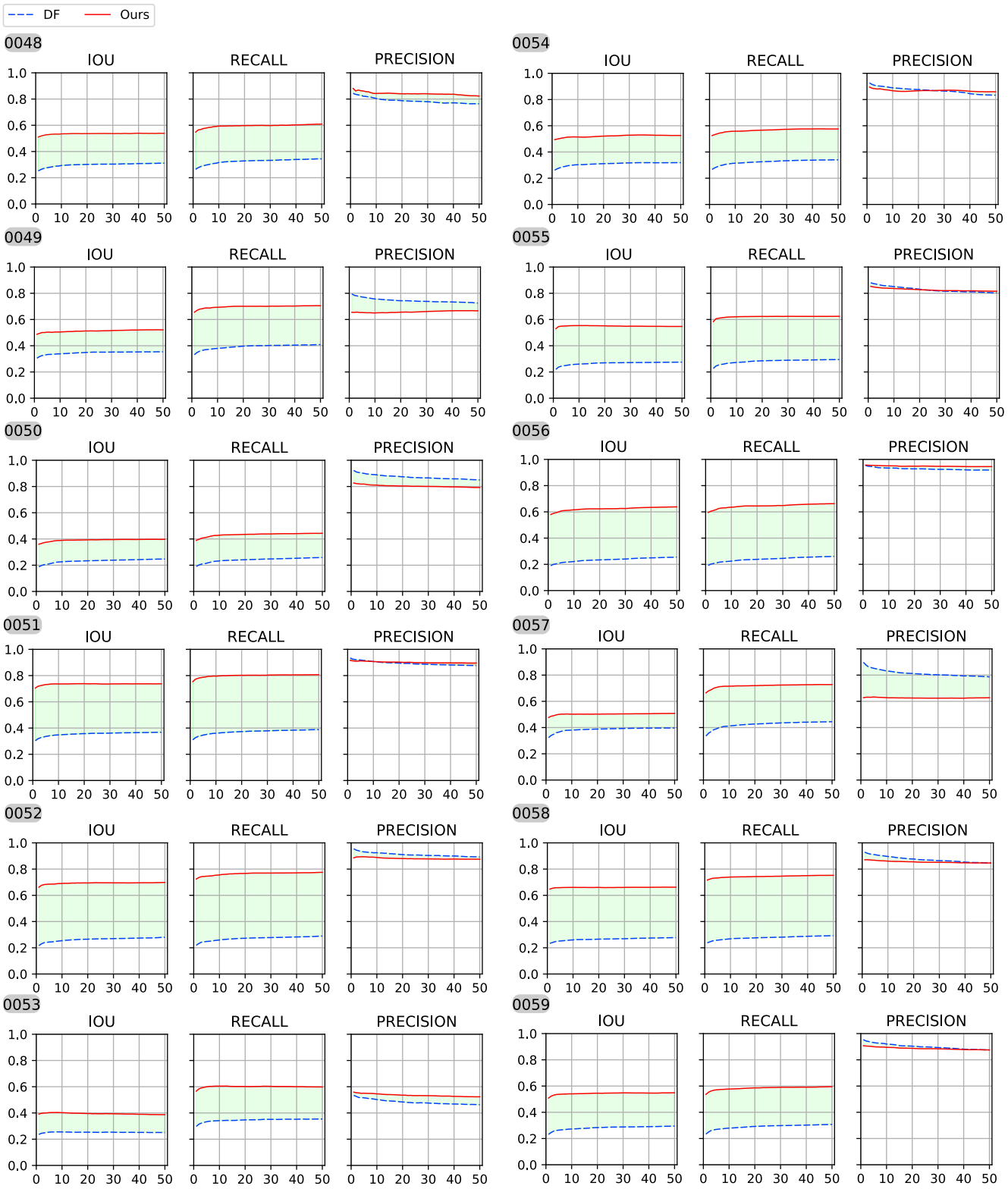

Figure 2: Plots of the metrics for each one of the validation sequence. The plots present tmetrics evaluated after the fusion of a single frame for the first 50 frames of the sequences. The solid red line is our approach while the blue dashed plot is obtained by fusing only depth information via TSDF averaging (DF).

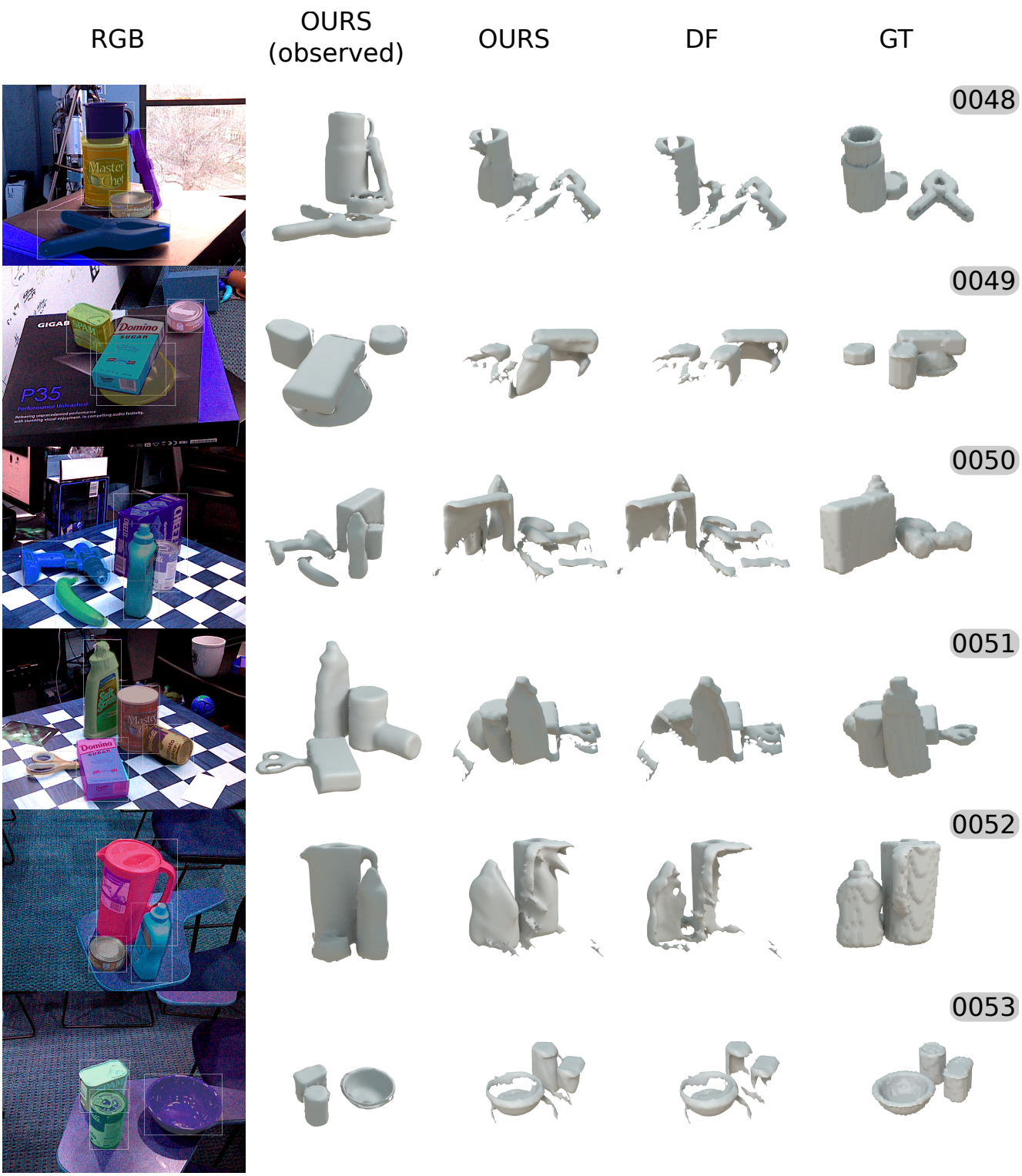

Figure 3: Reconstruction resulting from the fusion of the first 50 frames of the sequences. In this image are reported the first six sequences, *0048* to *0053*. From left, first RGB frame of the sequences, our reconstruction using thickness first showing the observed surfaces and then the estimated occupied space. The reconstruction labelled as DF is obtained by fusing only depth frames using traditional TSDF averaging. On the right, the voxelised scene.

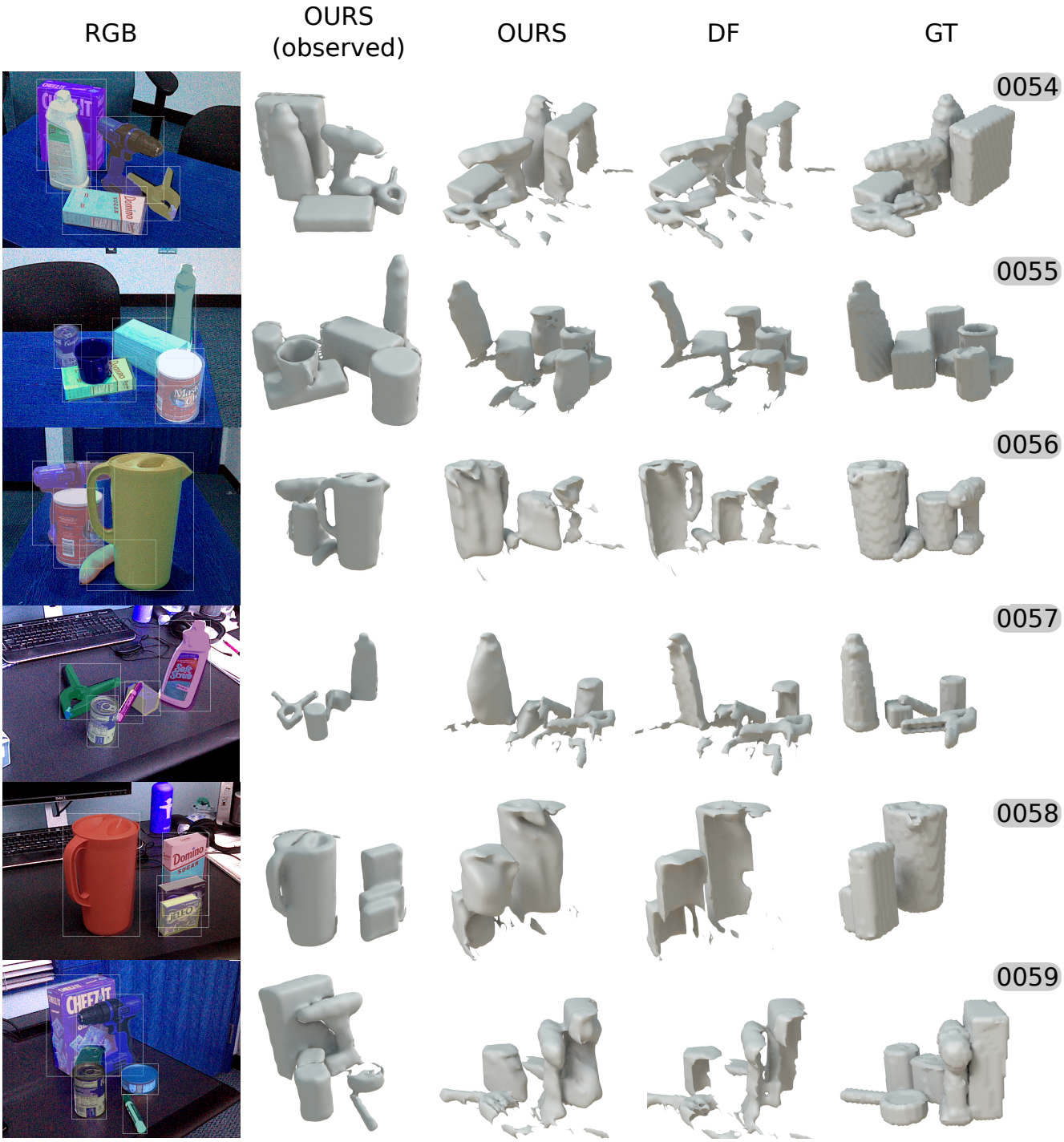

Figure 4: Reconstruction resulting from the fusion of the first 50 frames of the sequences. In this image are reported the second half of the sequences, *0054* to *0059*. From left, first RGB frame of the sequences, our reconstruction using thickness first showing the observed surfaces and then the estimated occupied space. The reconstruction labelled as DF is obtained by fusing only depth frames using traditional TSDF averaging. On the right, the voxelised scene.