# Easy3D: A Simple Yet Effective Method for 3D Interactive Segmentation

## Supplementary Material

This supplementary material provides:

- Additional details on the decoder architecture and attention operations (Sec. 7);
- Details on ScanNet++ and the creation of the Gaussian Splatting version of ScanNet40 (GS-ScanNet40) (Sec. 8);
- Additional implementation details of our method (Sec. 9);
- Analysis on efficiency (Sec. 10);
- Analysis on the impact of the number of clicks during training (Sec. 11).

## 7. Additional Decoder details

In our decoder, depicted in Fig. 6, we use a combination of attention-based [30] layers which are used to exchange information between the scene $S_E$, clicks $C_E$ and output $O_E$ embeddings. We use the symbol "$\rightarrow$" to describe the *direction of the exchange of information* that a specific layer is performing. More in detail, we refer to the definition of attention operation in [30] which relies on *queries*, *keys* and *values*. In the operation in Fig. 7, which is part of our decoder, the concatenation of clicks and output embedding ($\widehat{C_E}\widehat{O_E}$) will correspond to the queries, while the keys and values will be represented by the scene embedding ($S_E$). Note that, as done in SAM [13], in any attention operation we add to the queries and keys their corresponding positional encoding and label encoding if needed. In case an attention operation involves the use of clicks, we add their corresponding positional encoding and learned label encoding each time, while in case it involves the scene embedding, we add the voxels' positional encoding.

## 8. Details on GS-ScanNet40 and ScanNet++

In order to test the capabilities of our method on a different geometric distribution, we introduced a Gaussian Splatting [12] version of ScanNet40, *i.e*. GS-ScanNet40. To create it, we relied on the *SplatFacto* method available in the popular Nerfstudio repository [28]. We reconstructed all the $\approx$ 1500 ScanNet40 scenes using their corresponding posed DSLR images, following the standard SplatFacto configuration. Once obtained the GS reconstruction, we matched the ScanNet40 mesh-based instance annotations with their gaussian version, by finding the nearest point to the mesh for each gaussian, and assigning the corresponding instance label if their distance was below 5cm. Note that 5cm is the voxel resolution $V_S$ used in all our experiments. Please also note that, being ScanNet a dataset where 3D annotations have been performed and are available only on incomplete and decimated meshes, it was not possible to use images to encode the labels into the gaussians during the scene reconstruction. This process in fact requires the definition of many heuristics, which lead to an unsatisfactory solution.

For what concerns ScanNet++ [35], we used the official repository to obtain vertex-level instance annotations on the meshes. We then obtained our results on the 50 scenes that are part of the official validation set, retaining the objects that are part of the official Instance Segmentation benchmark at the time of submission.

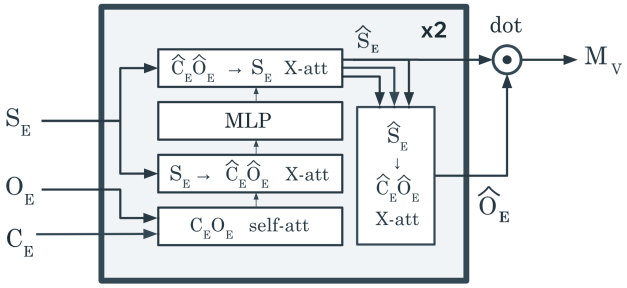


Figure 6. Detail of the attention operations performed inside our decoder.

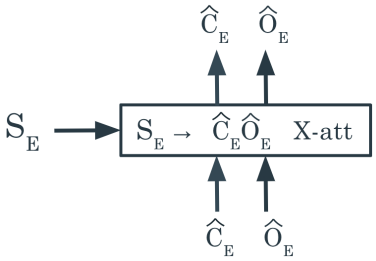


Figure 7. Detail of one of the attention-based [30] operations performed in our decoder.

## 9. Additional implementation details

We trained our method using a batch size of 8 on an NVIDIA A100 with 80GB of memory for $\approx$ 2 days, using an AdamW [18] optimizer with a 0.05 weight decay.

## 10. Analysis on Efficiency

We provide the efficiency comparison in Tab. 6, as originally proposed in AGILE3D [37] and evaluated on an NVIDIA Titan RTX (24GB). We report memory (Mem.) and computation (GFLOPs) requirements, as well as 5 clicks inference time (t@5). From this analysis, it can be seen that our method is shown to be more efficient than AGILE3D and Point-SAM.

| Test Dataset | Train $N_c$ | IoU@1 | IoU@2 | IoU@3 | IoU@5 | IoU@10 |
|---|---|---|---|---|---|---|
| ScanNet40 | 10 | 68.2 | 74.6 | 77.3 | 79.6 | 81.7 |
| | 20 | 67.5 | 75.8 | 76.4 | 80.2 | 81.4 |
| S3DIS | 10 | 65.7 | 76.0 | 80.8 | 84.9 | 87.8 |
| | 20 | 66.3 | 75.5 | 81.9 | 83.6 | 88.1 |
| KITTI-360 | 10 | 46.3 | 58.7 | 66.7 | 76.2 | 83.6 |
| | 20 | 46.0 | 59.5 | 65.8 | 77.2 | 82.3 |
| ScanNet++ | 10 | 54.8 | 61.4 | 64.7 | 67.9 | 71.3 |
| | 20 | 55.0 | 60.5 | 65.2 | 67.1 | 72.3 |

Table 5. Results of our model by using a maximum number of clicks ($N_c$) equal to 10 and 20. We followed the same protocol as Tab. 1 of the main paper, training and testing on ScanNet40.

| Method | Mem. [Mb] | GFLOPS | t@5 |
|---|---|---|---|
| InterObject3D | 924 | 1.6 | 1.2 |
| AGILE3D | 710 | 16.6 | 0.5 |
| Point-SAM | 10653 | 65.4 | 0.7 |
| Ours | 425 | 14.3 | 0.3 |

Table 6. Comparison on efficiency as introduced in AGILE3D [37]. We report memory requirement (Mem.) in Megabytes, inference GFLOPS as a metric for computational requirements and inference time for 5 clicks (t@5) in seconds.

## 11. Analysis on Training Number of Clicks

We trained our model with a different number of clicks during training, $N_c = \{10, 20\}$, following the same protocols as Tab. 1 of the main paper. The results in Tab. 5 indicate that the performance in the low-click regime with $N_c = 10$ is similar to the one with $N_c = 20$, showing that $N_c$ does not seem to impact it. Our interpretation is that the most important components of our model are the output embeddings and the decoder, which learn how to combine clicks to predict the output mask. The conditions in which these components operate between 10 and 20 clicks differ significantly from those between 1 and 5 clicks, so training with more clicks does not substantially affect how the model performs in the low-click regime.

## 12. Additional Qualitative Results

In Figs. 8 to 11 we provide additional qualitative results by visualizing the predictions of our method and AGILE3D [37]. Similarly to the main paper, we visualize results for up to 3 user clicks reporting the corresponding IoU@{1,2,3} metric between the predicted mask (red mask) and the ground-truth object (green mask) of models trained only on ScanNet40 [3]. Please note that while the clicks have been visualized with a similar blue sphere, they can represent a positive or negative click depending if they are part of the ground-truth object mask shown on the left (green). If a click is on the mask, then the click will be positive, otherwise it will be negative.

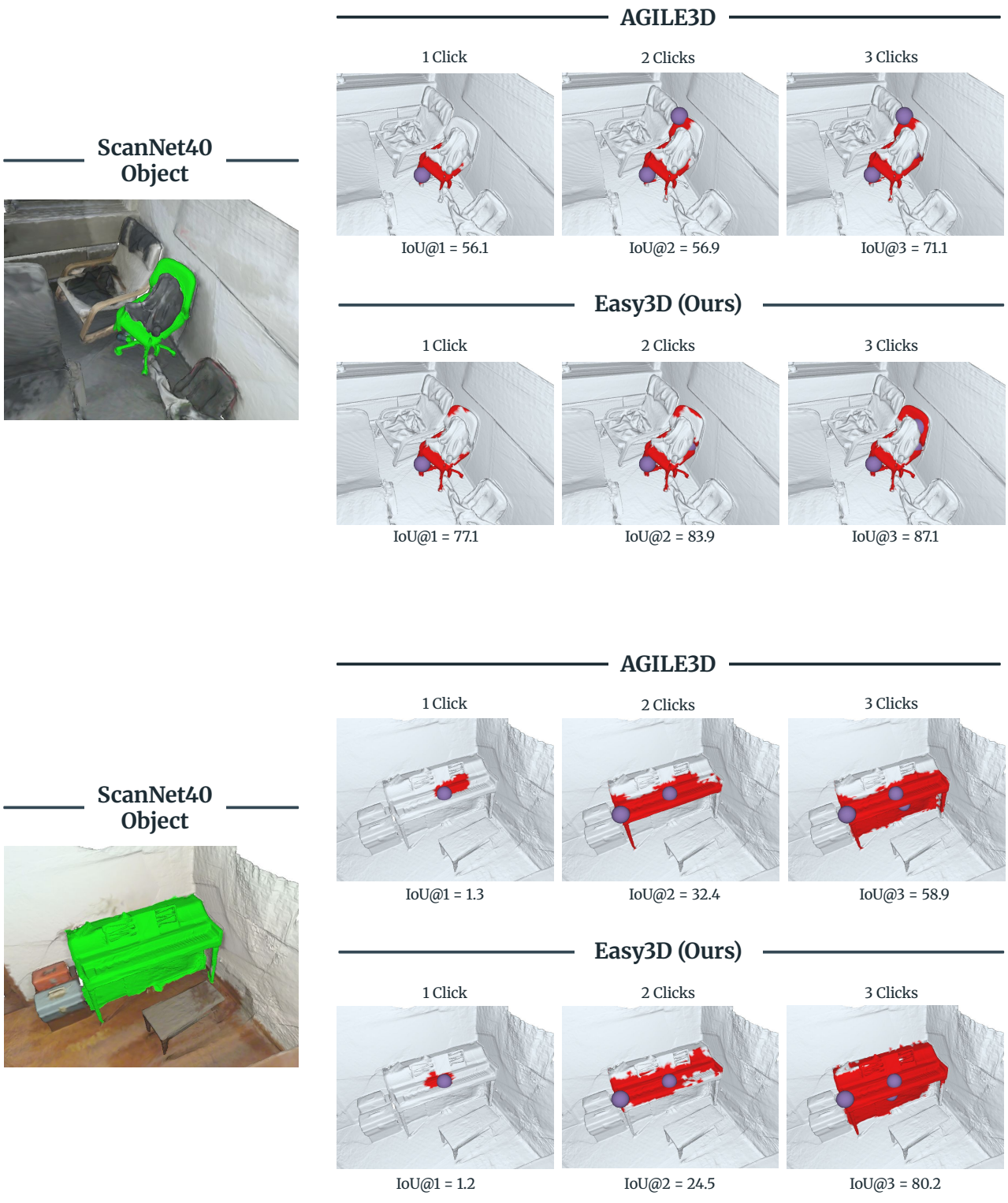


Figure 8. Additional qualitative results on ScanNet40 [3].

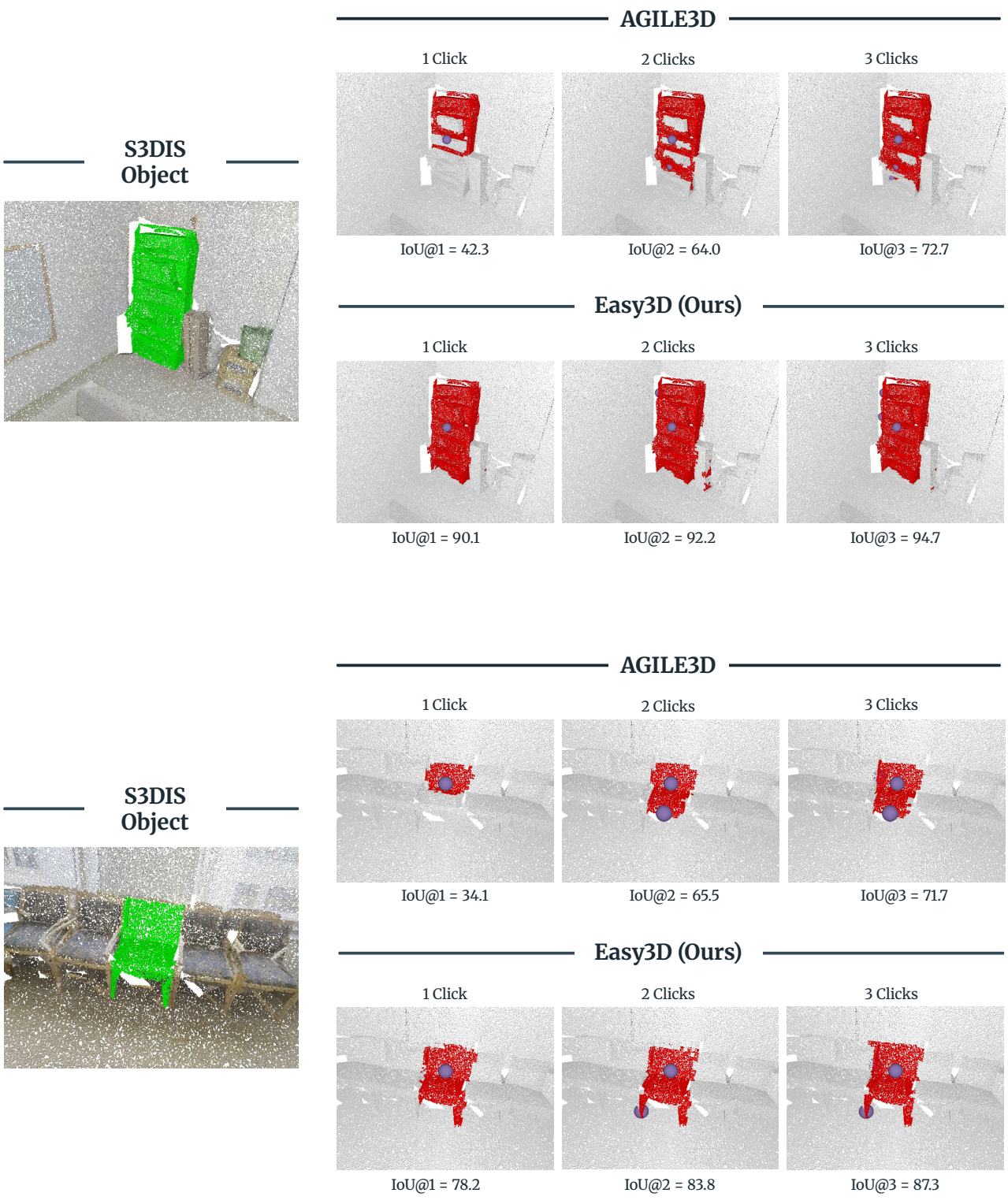


Figure 9. Additional qualitative results on S3DIS [1].

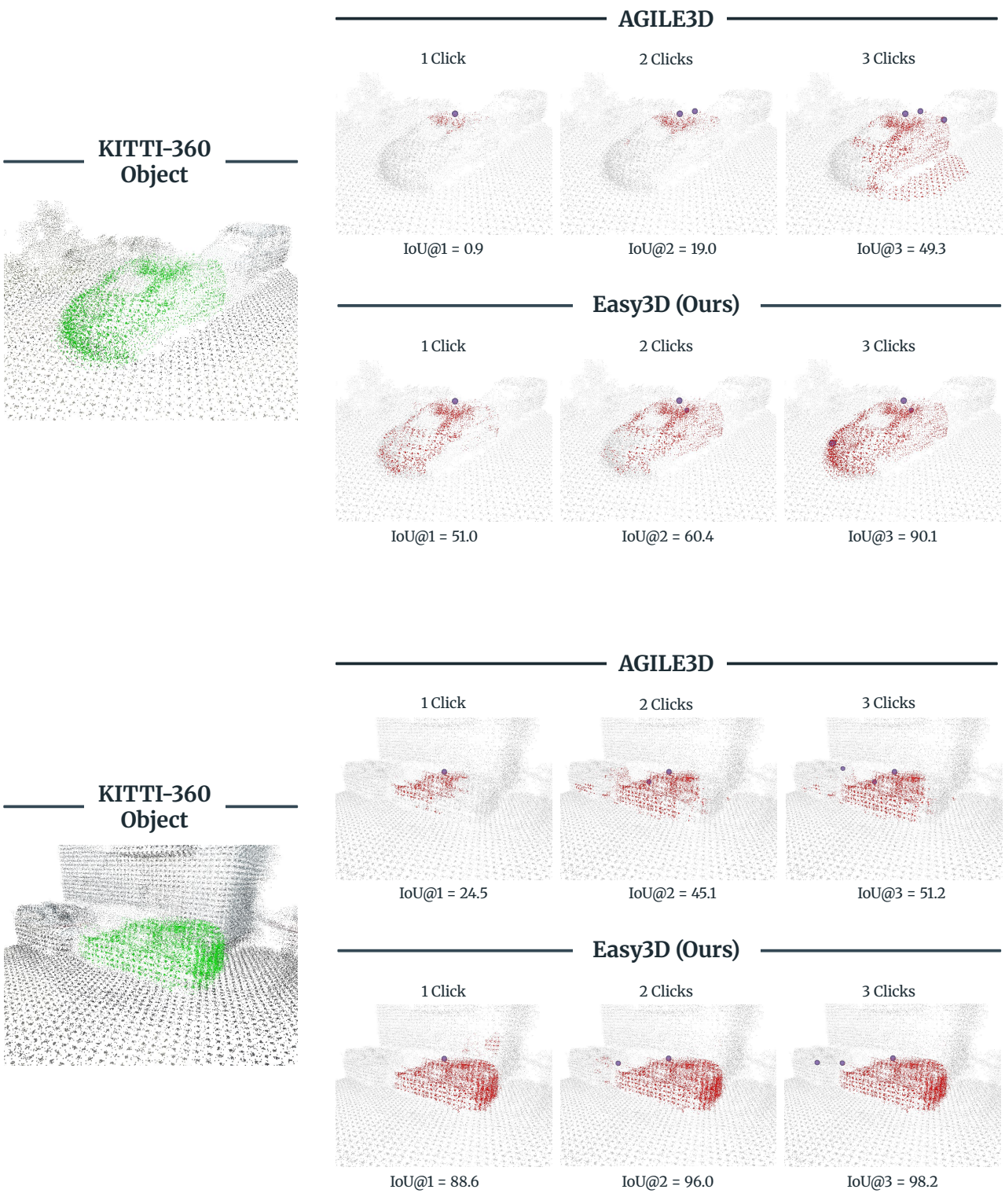


Figure 10. Additional qualitative results on KITTI-360 [17].

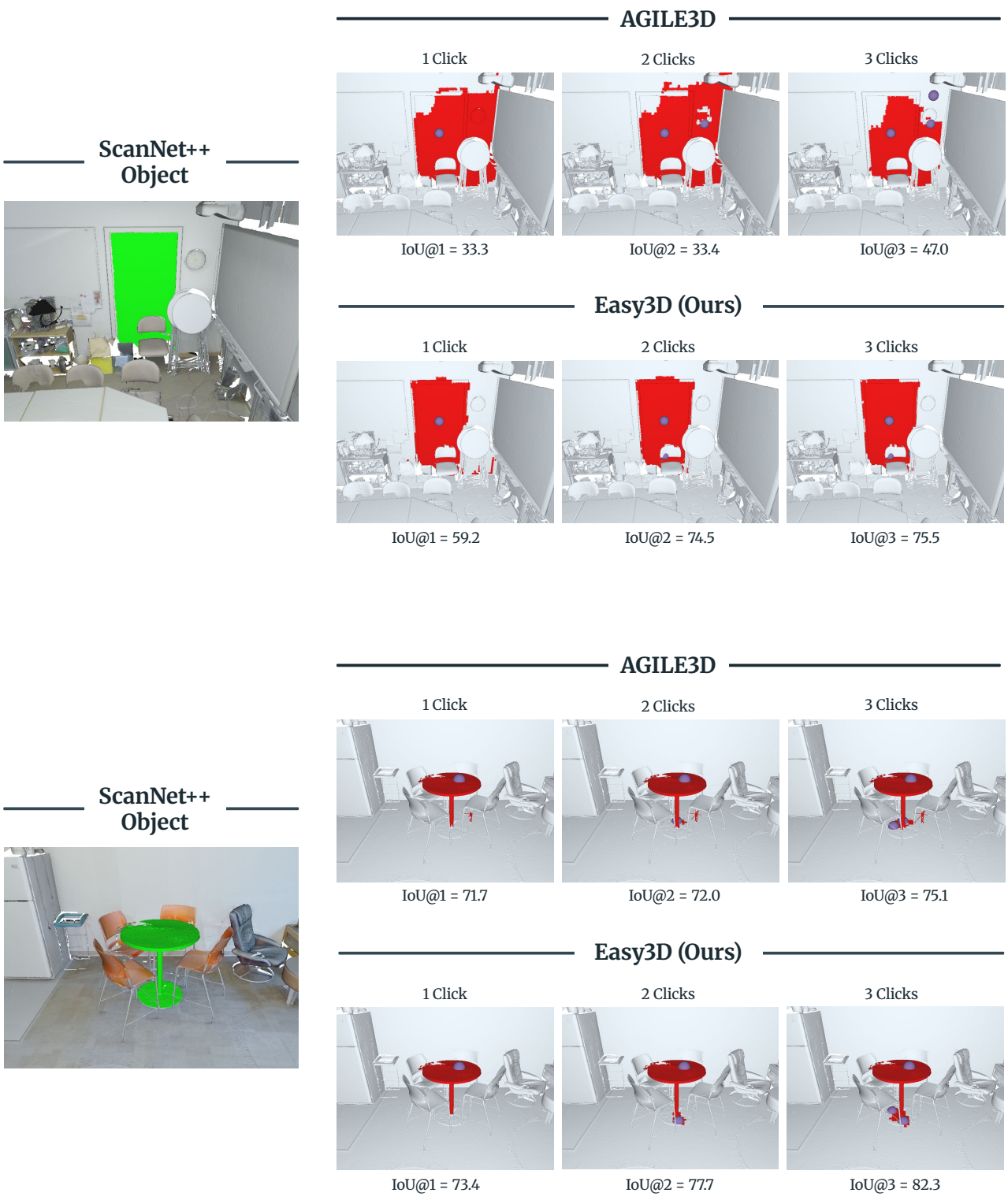


Figure 11. Additional qualitative results on ScanNet++ [35].