RGB-D Cube R-CNN: 3D Object Detection with Selective Modality Dropout

Supplementary Material

6. Qualitative Examples

Figs. 5 to 7 show additional qualitative results on the ARKitscenes, Hypersim and SUN RGB-D subsets of the $\text{Omni3D}_{\text{IN}}$ test datasets respectively. The illustration expands upon the figure in the paper by offering an additional image plane projection of the detections, as well as a top-down perspective that highlights the differences between the predictions (red) and the actual ground-truth (green). The direct comparison between predictions and ground truth clearly reveals that the RGB-D models not only aligns the bounding boxes more accurately but also captures the overall room layout with greater precision compared to the RGB-only model. While the image plan projections often look good, the shape can still be wrong, *e.g.* the bed in Fig. 7 (bottom row in the middle) is too small. It is worth noting that the RGB-only model frequently fails to detect some annotated objects, while the RGB-D version excels in identifying them. Additionally, in certain instances, both models make accurate predictions for objects that are not included in the annotations (or they are dropped due to occlusions/truncation). This can for example be seen in Fig. 6 (bottom row on the left) where both models confidently detect the night stand and the sofa, showcasing their generalization capabilities.

Figs. 8 to 10 additionally looks at some of these examples in more detail, showing how each of the backbones compare to when using RGB or RGB-D input. In general the Swin-T backbone performs best, followed by DLA34 and finally ViT-S. While in most cases the model with the additional depth input sees improved results, the DLA34-based model clearly has a harder time fully utilizing the extra depth. For DLA34 the improvements in alignment are often less pronounced, whereas the both Transformer-based models can often utilize the additional information better.

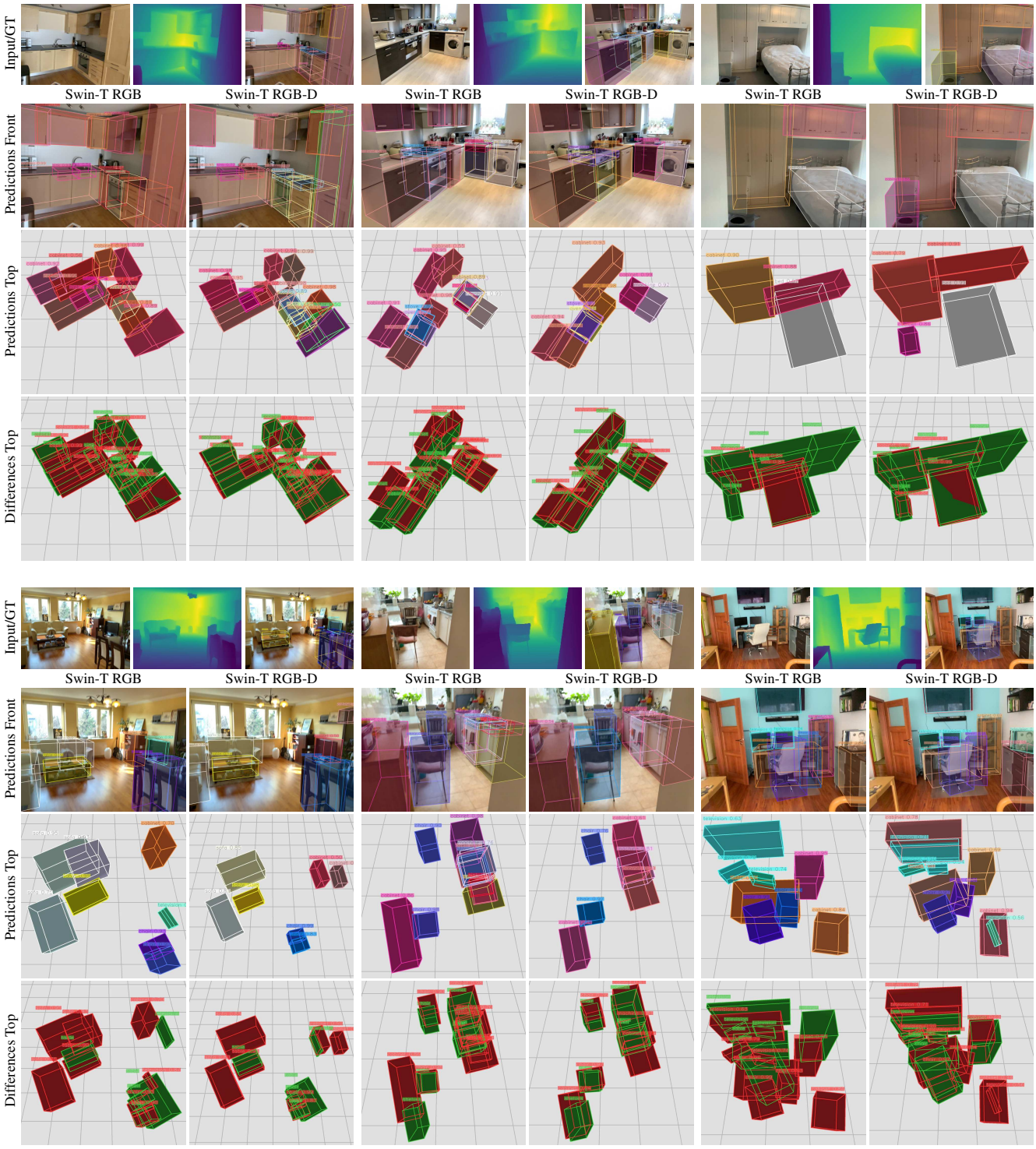

Figure 5. Qualitative results of our Swin-T model using either RGB or RGB-D inputs on the ARKitScenes test dataset. The top row of each example displays the RGB and depth inputs, alongside the ground-truth annotations, viewed from a front perspective. The second row showcases the predictions generated by our RGB and RGB-D models, also from a front perspective. Subsequently, the next two rows exhibit the predictions from a top perspective. In the first row of the top-view predictions, colors represent different classes, whereas in the second row, cuboids are colored as ground-truth annotations (green) or predictions (red). All examples highlight that the predictions of the RGB-D model are much better aligned with the ground-truth compared to the RGB-only model.

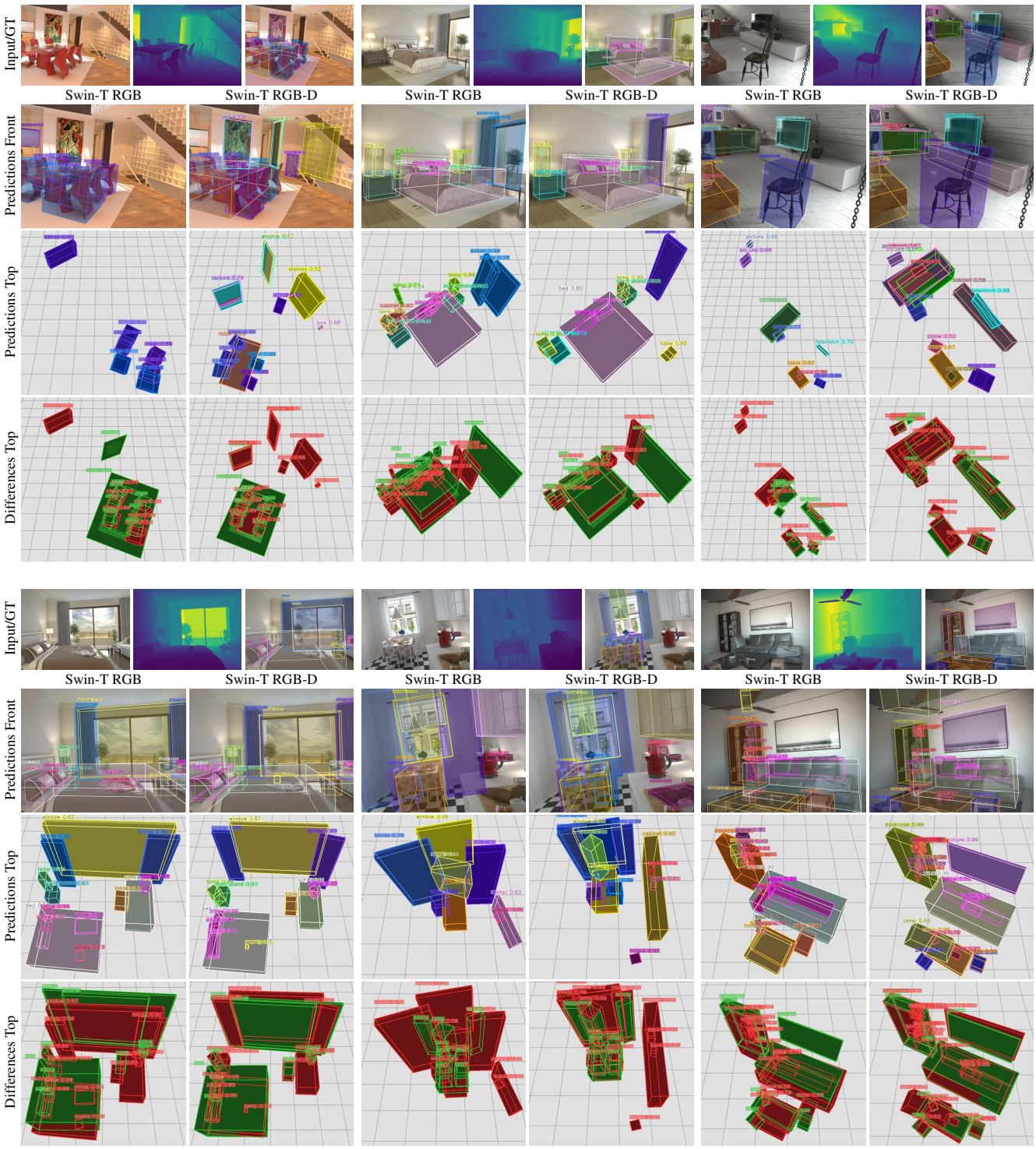

Figure 6. Qualitative results of our Swin-T model using either RGB or RGB-D inputs on the Hypersim test dataset. The top row of each example displays the RGB and depth inputs, alongside the ground-truth annotations, viewed from a front perspective. The second row showcases the predictions generated by our RGB and RGB-D models, also from a front perspective. Subsequently, the next two rows exhibit the predictions from a top perspective. In the first row of the top-view predictions, colors represent different classes, whereas in the second row, cuboids are colored as ground-truth annotations (green) or predictions (red). The difference view reveals that the RGB-D model not only aligns the boxes more accurately but also captures the overall room layout with greater precision compared to the RGB-only model. For instance, in the third example, the RGB model fails to predict the depth of the pictures located in the rear.

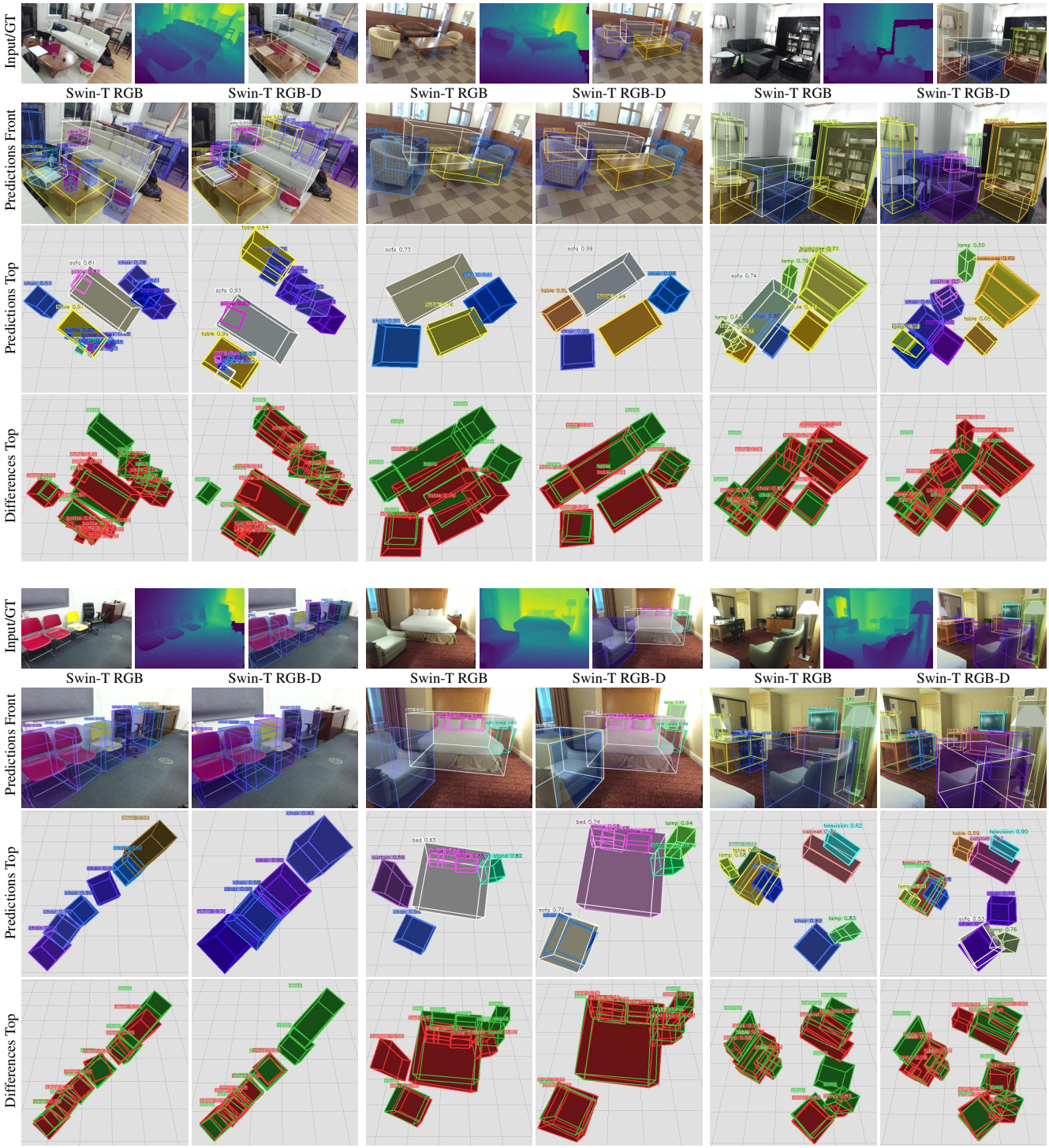

Figure 7. Qualitative results of our Swin-T model using either RGB or RGB-D inputs on the SUN RGB-D test dataset. The top row of each example displays the RGB and depth inputs, alongside the ground-truth annotations, viewed from a front perspective. The second row showcases the predictions generated by our RGB and RGB-D models, also from a front perspective. Subsequently, the next two rows exhibit the predictions from a top perspective. In the first row of the top-view predictions, colors represent different classes, whereas in the second row, cuboids are colored as ground-truth annotations (green) or predictions (red).

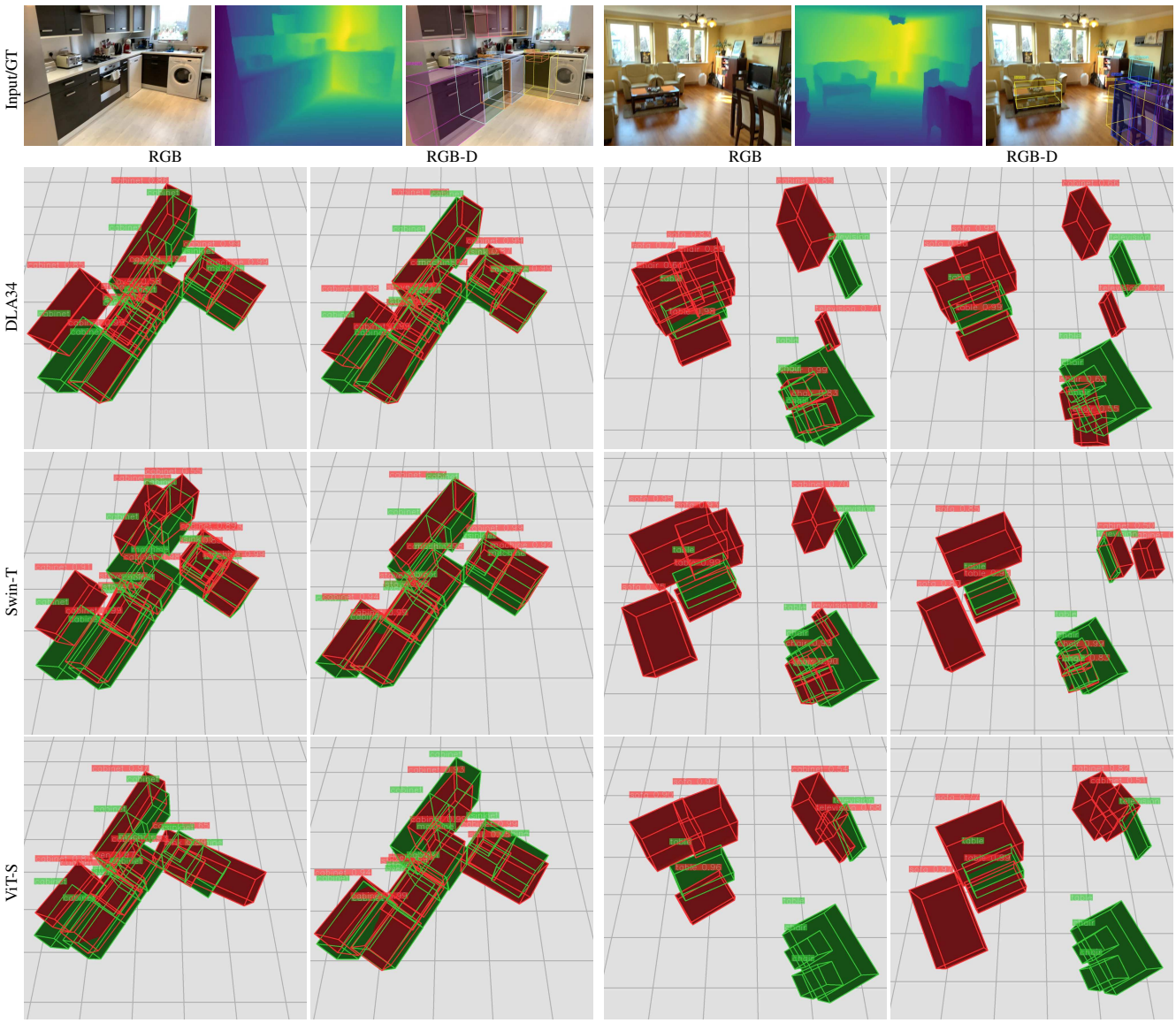

Figure 8. Qualitative results on the ARKitScenes test dataset. The top row of each example displays the RGB and depth inputs, alongside the ground-truth annotations, viewed from a front perspective. Followed by top perspective views of the predictions of DLA34 (top row), Swin-T (middle row) and ViT-S (bottom row) using either RGB or RGB-D inputs. The cuboids are colored as ground-truth annotations (green) or predictions (red). It becomes apparent that the DLA34 predictions do not improve significantly compared to Swin-T when incorporating depth.

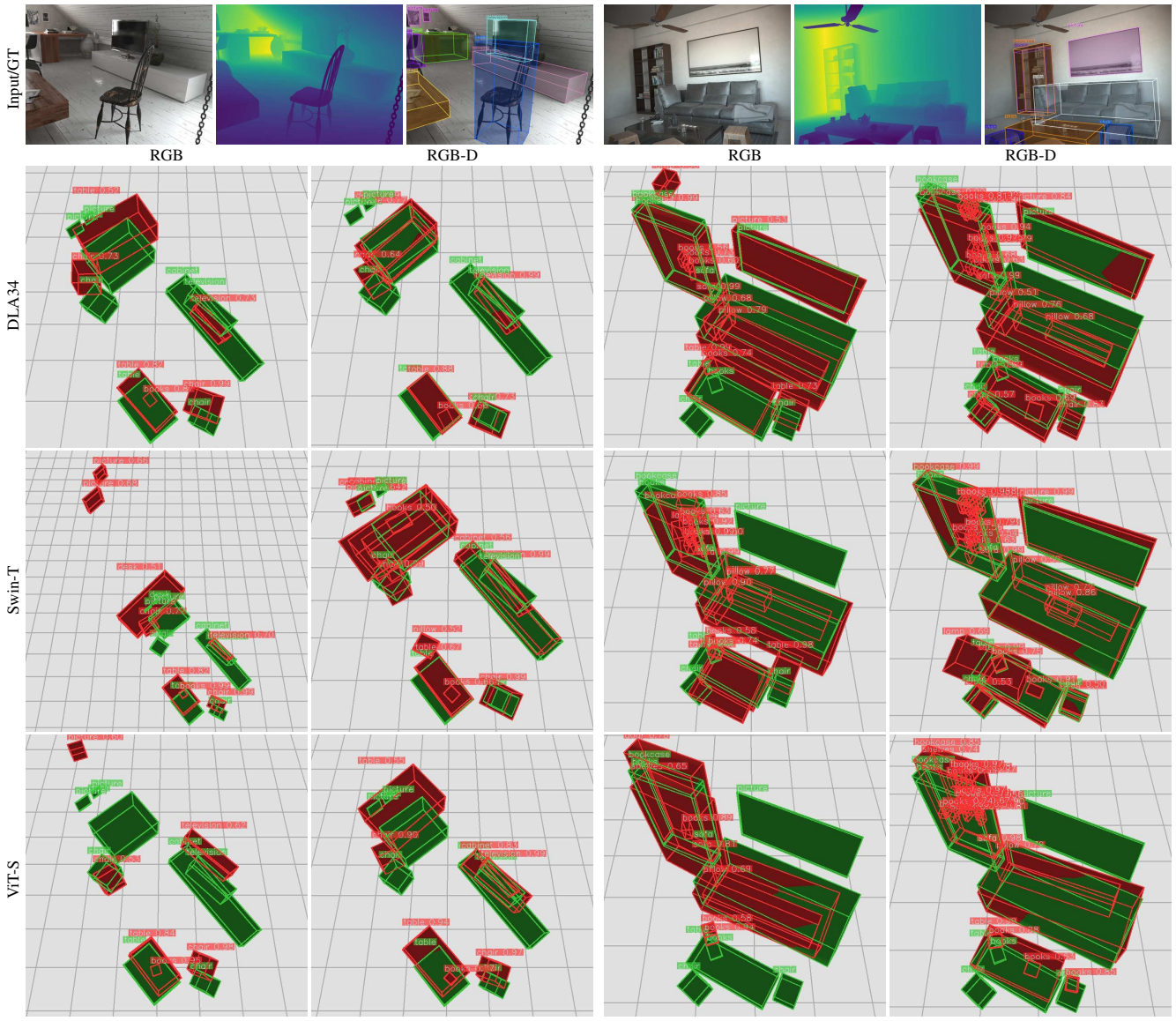

Figure 9. Qualitative results on the Hypersim test dataset. The top row of each example displays the RGB and depth inputs, alongside the ground-truth annotations, viewed from a front perspective. Followed by top perspective views of the predictions of DLA34 (top row), Swin-T (middle row) and ViT-S (bottom row) using either RGB or RGB-D inputs. The cuboids are colored as ground-truth annotations (green) or predictions (red).

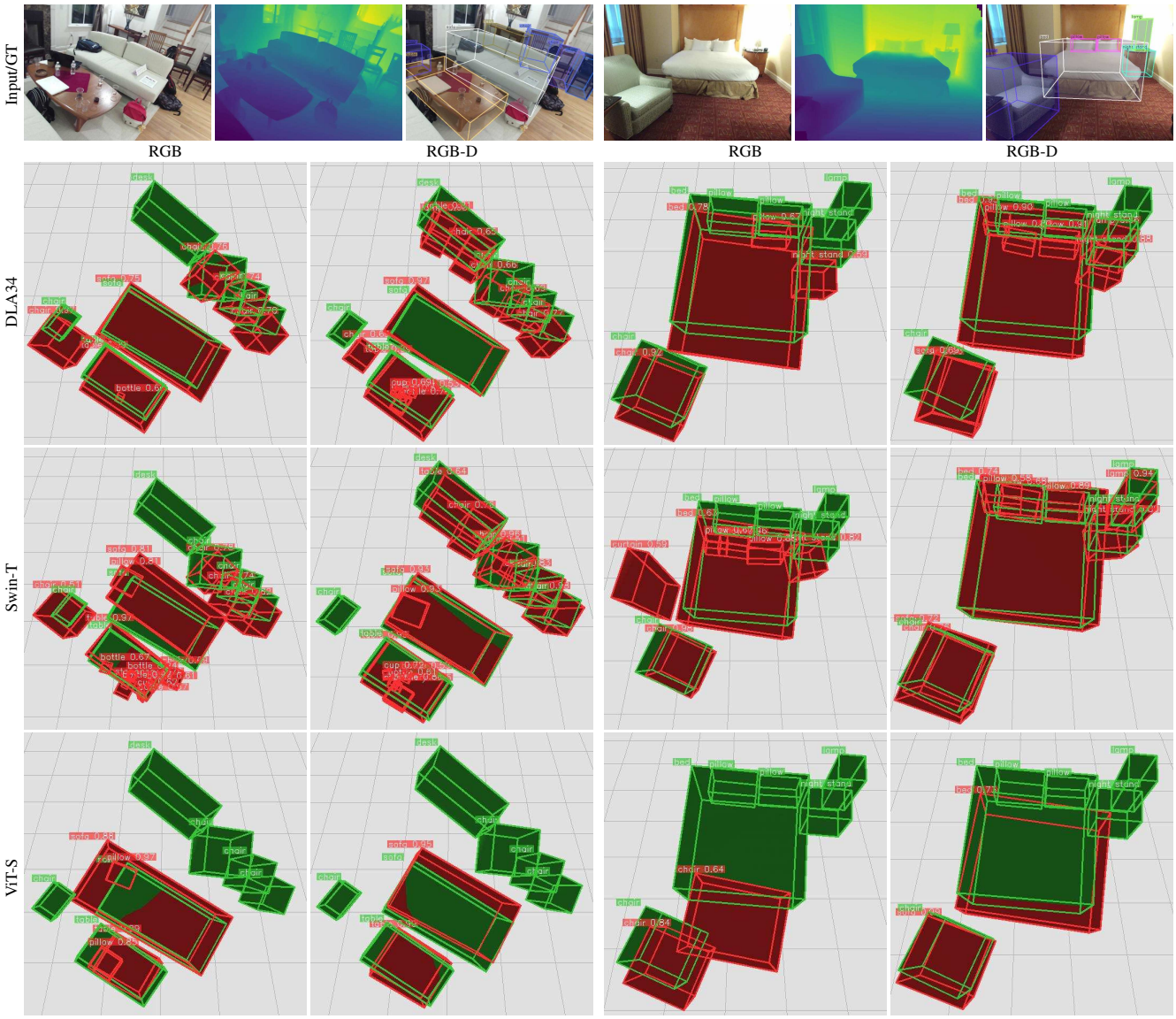

Figure 10. Qualitative results on the SUN RGB-D test dataset. The top row of each example displays the RGB and depth inputs, alongside the ground-truth annotations, viewed from a front perspective. Followed by top perspective views of the predictions of DLA34 (top row), Swin-T (middle row) and ViT-S (bottom row) using either RGB or RGB-D inputs. The cuboids are colored as ground-truth annotations (green) or predictions (red). While DLA34 and Swin-T yield reasonable predictions, it is noticeable that the predictions generated by ViT-S are of inferior quality.