DetAny4D: Detect Anything 4D Temporally in a Streaming RGB Video

Supplementary Material

7. Dataset Composition

DA4D is a hybrid 4D detection dataset which includes 12 sub-datasets. The supporting tasks include sequence 4D detection, monocular 3D detection, depth estimation, and reconstruction tasks. DA4D is built on six datasets in Omni3D [5] (ArkitScenes [3], Hypersim [27], KITTI [12], nuScenes [6], Objectron [1], and SUNRGBD [30]) and expanded with a sequence dimension. Six more datasets are introduced for sequence-wise tasks, including Replica [31], MP3D [7], HM3D [26], HSSD [16], Gibson [38], and Scannet [9]. The format is standardized similar to the Omni3D structure with additional sequence information. Each sequence compromises a list of frames. Each frame includes monocular RGB image, camera intrinsics, camera pose, depth map, and object information. The object information includes 3D b-box attribute ($[x, y, z, w, h, l, yaw]$), rotation pose, 2D b-box prompt, category, instance ID, and score.

Dataset Composition. The dataset compromises original single frame 3D detection data as 3D detection capacity validation and multi-frame sequences data for 4D tasks. As shown in Figrue 7, DA4D consists of multi-frame sequences and Omni3D 3D detection data considered as sequences with length of 1.

Dataset Split. Sequences from Omni3D [5] follows the splitting strategy in prior works [5]. For newly compromised datasets to form multi-frame sequences, we split the scenes into the training and validation set, and separately record sequences from the selected scenarios, ensuring all scenes in the validation set have not been visited.

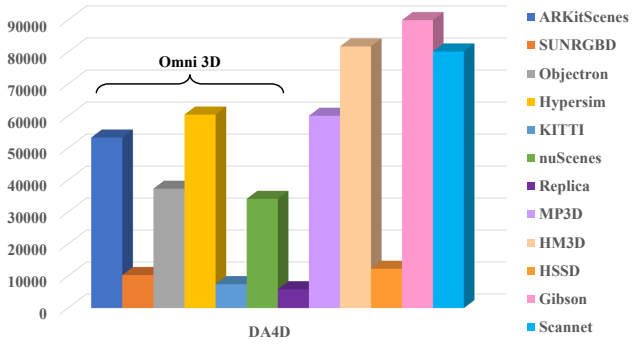

Figure 7. Visualization of the DA4D dataset composition.

8. Training Strategy Details

Here we illustrate the sequence crop and object padding strategy in Section 4.4 with more details. As shown in Figure 8, we crop each clip under a fixed maximum length. For the cropped sequence, we count the total objects number in each sequence and pad the objects in each frame to this number. During training, this strategy ensures each frame has the same object query dimension, and predictions generated with the padded queries will be masked which do not contribute to the loss. During inference, the padded queries enable the model to manage newly appeared objects during forecasting, where new objects are registered to the padded queries and maintained if the prediction result has a high score and differs against objects in the query memory list.

Figure 8. Visualization of the sequence crop and object padding strategy. The object query list maintains the objects and padding status. As frames in the sequence forecast, object status updates. The disappeared objects and padded objects do not contribute to the loss.

9. 3D B-box annotation Pattern

This section illustrates the differences in the b-box annotation pattern for 3D detection task and 4D detection task. As shown in Figure 9, when observing an object from different views, 3D detection annotation pattern turns to consider each view as a singular observation and annotate the box referring only to the current view. On the contrary, for 4D detection, each object is registered with a b-box in the global coordinates and this constant b-box is projected to various views ensuring consistency globally.

Given the annotation formats of the 4D detection task and our goal to leverage powerful prior knowledge from 3D detection, specifically by utilizing pre-trained 3D detection models, we designed a specialized loss function (Section 4.4) to constrain the predictions to align with 4D task

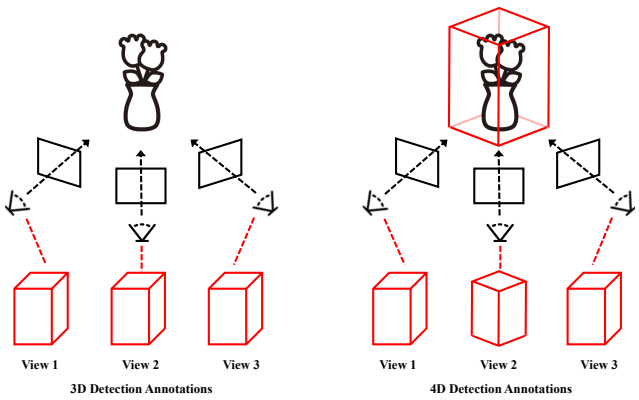

Figure 9. Visualization of the annotation pattern differences. 3D detection task turns to annotate each object reffering only to the current view. 4D detection requires each view predicts objects in the global coordinates.

annotations. Our composite loss function, detailed in Section 4.4, incorporates constraints on the center, dimensions, and rotation angle of the 3D bounding boxes. Figure 10 shows the supervision of the b-box. In contrast to the rigid constraints traditionally used in 3D detection, we employ a softened loss for the dimensions and rotation. This design is motivated by the potential misalignment between the outputs of 3D pre-trained models and the 4D annotations, as shown in Figure 11. Directly applying hard constraints can lead to slow and difficult convergence, whereas the softened loss facilitates a more effective and stable alignment of the predictions to the 4D ground truth.

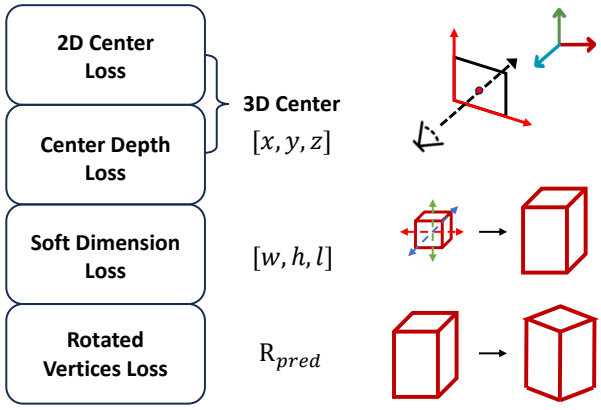

Figure 10. Visualization of the b-box supervision. The 3D bounding box is constrained sequentially by its center, dimensions, and rotation angle.

10. Open-Set Validation

To provide a detailed assessment of our model's open-set detection capabilities, we separately evaluated the 4D detection metrics on objects from categories that were within the training set (seen categories) and those that were outside of it (unseen categories) in the validation set. Table 4 shows the AP and F1 score under threshold IoU@0.5 (the same metrics illustrated in Section 5.1) of open-set categories and none-open-set categories in the validation set. It can be seen that the model performance on open-set highly keeps align with the none-open-set performance.

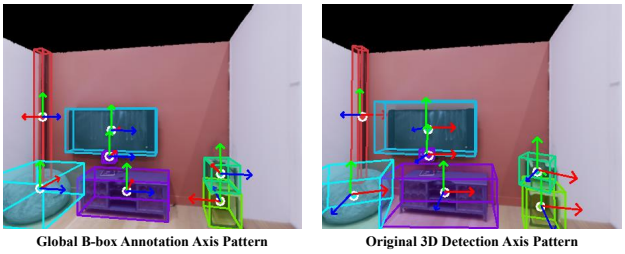

Figure 11. Visualization of the axis pattern comparison between global annotations and 3D prediction results. Global annotated b-box dimensions and rotation differs from the predicted results of pre-trained 3D detection model.

Datasets	Categories Open-Set / All	Objects Num Open-Set / All	**Open-Set**		**None-Open-Set**	
			$AP_{3D}^{\uparrow}$	$F1^{\uparrow}$	$AP_{3D}^{\uparrow}$	$F1^{\uparrow}$
Replica	18 / 62	0.3k / 16k	27.9	46.8	28.0	47.1
MP3D	20 / 330	3k / 98k	24.7	41.9	25.0	43.2
HM3D	75 / 698	11k / 344k	27.2	44.4	26.7	43.7

Table 4. Evaluation of the open-set performance. We separately evaluate the AP and F1 score on three sub-datasets and also report the ratio of the categories and numbers of open-set objects.

11. More Results

We provide more visualization comparison to show the prior performance of our proposed DetAny4D on the spatiotemporal consistency validation.

Figure 12. More qualitative comparison results.

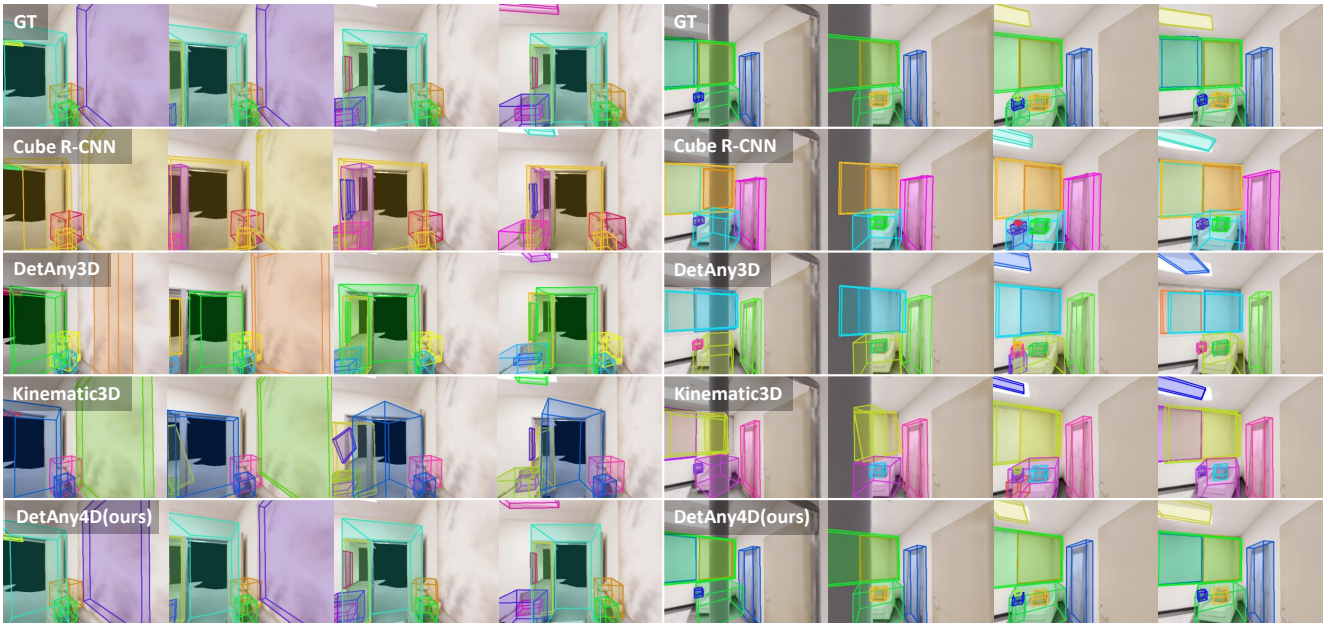

Figure 13. More qualitative comparison results.

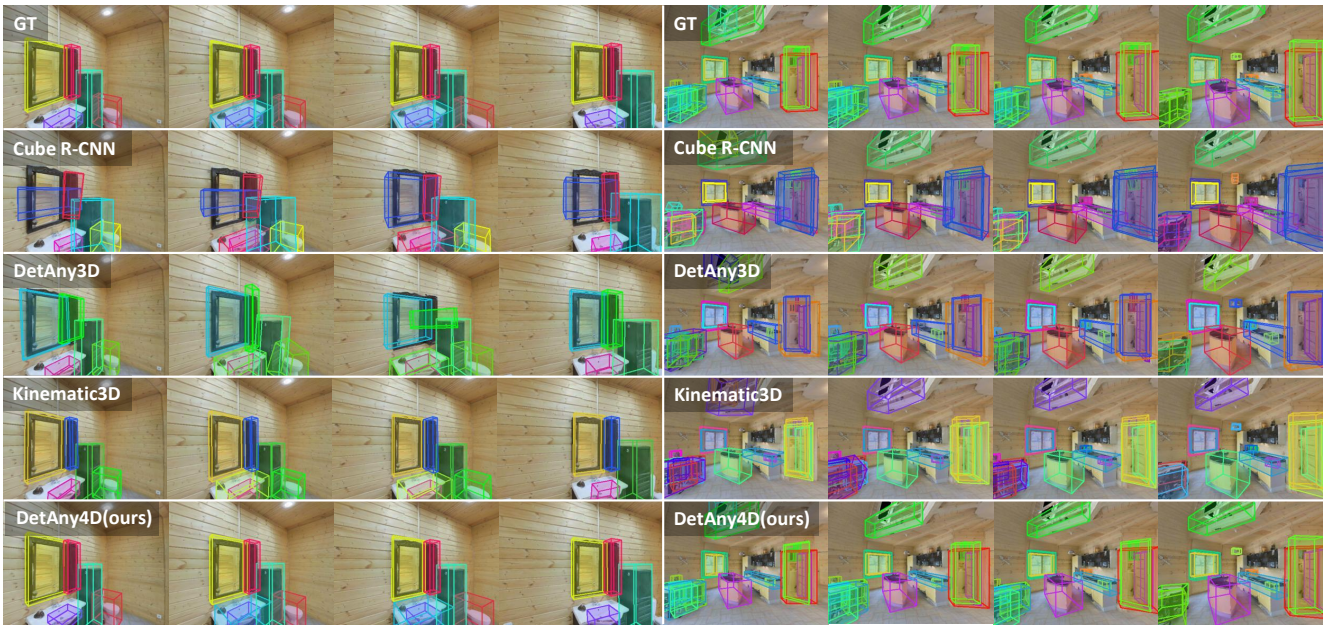

Figure 14. More qualitative comparison results.

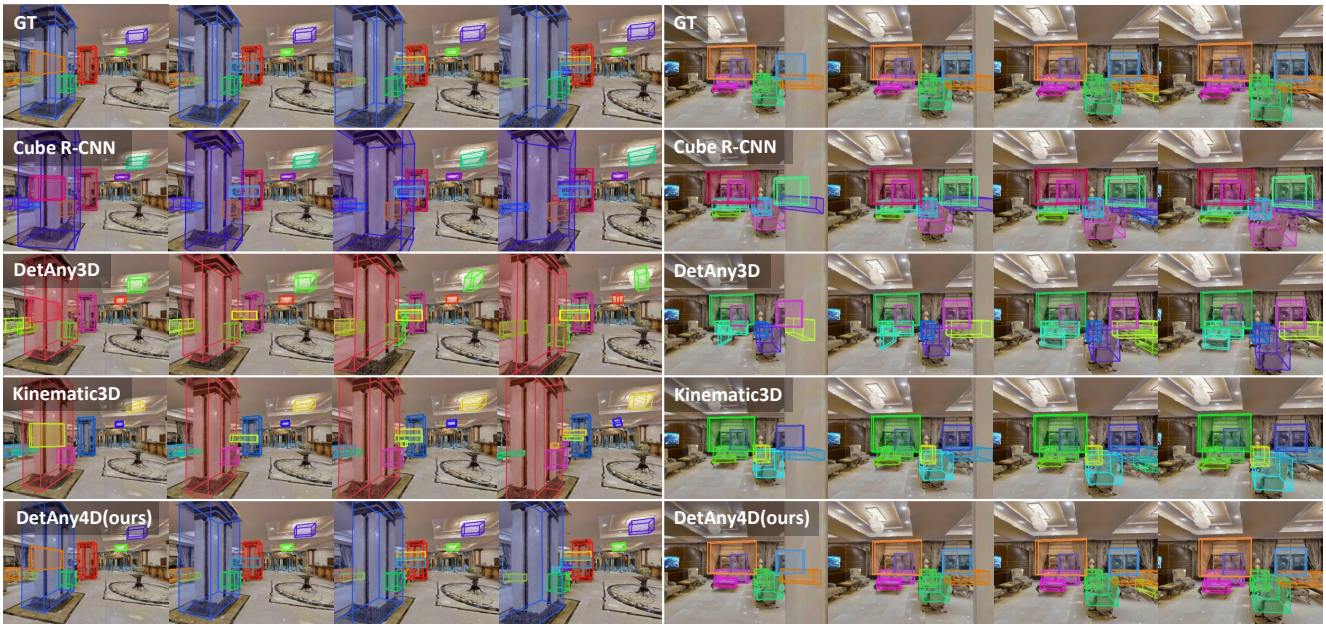

Figure 15. More qualitative comparison results.

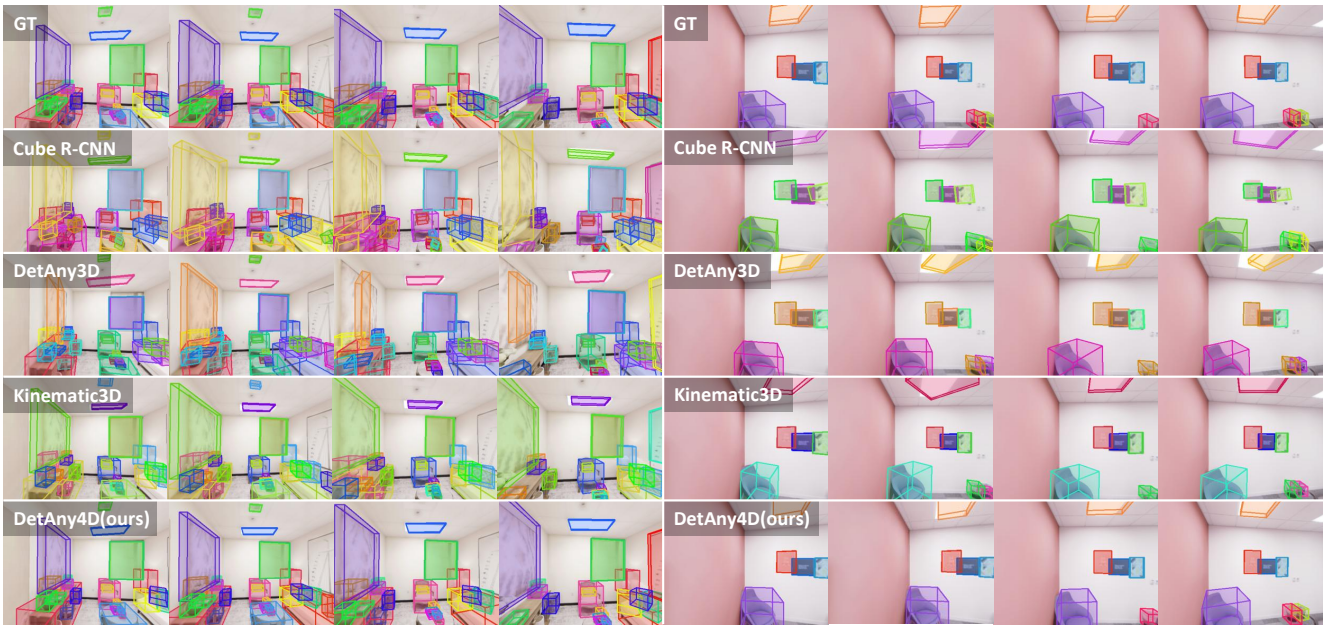

Figure 16. More qualitative comparison results.

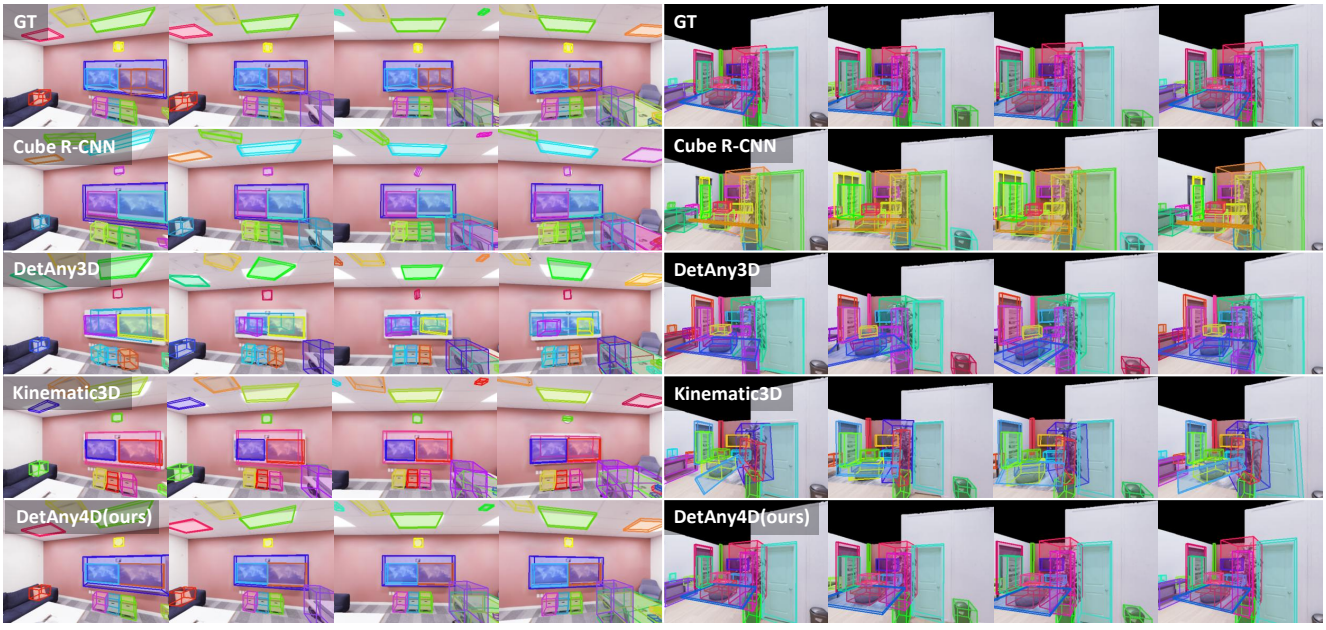

Figure 17. More qualitative comparison results.

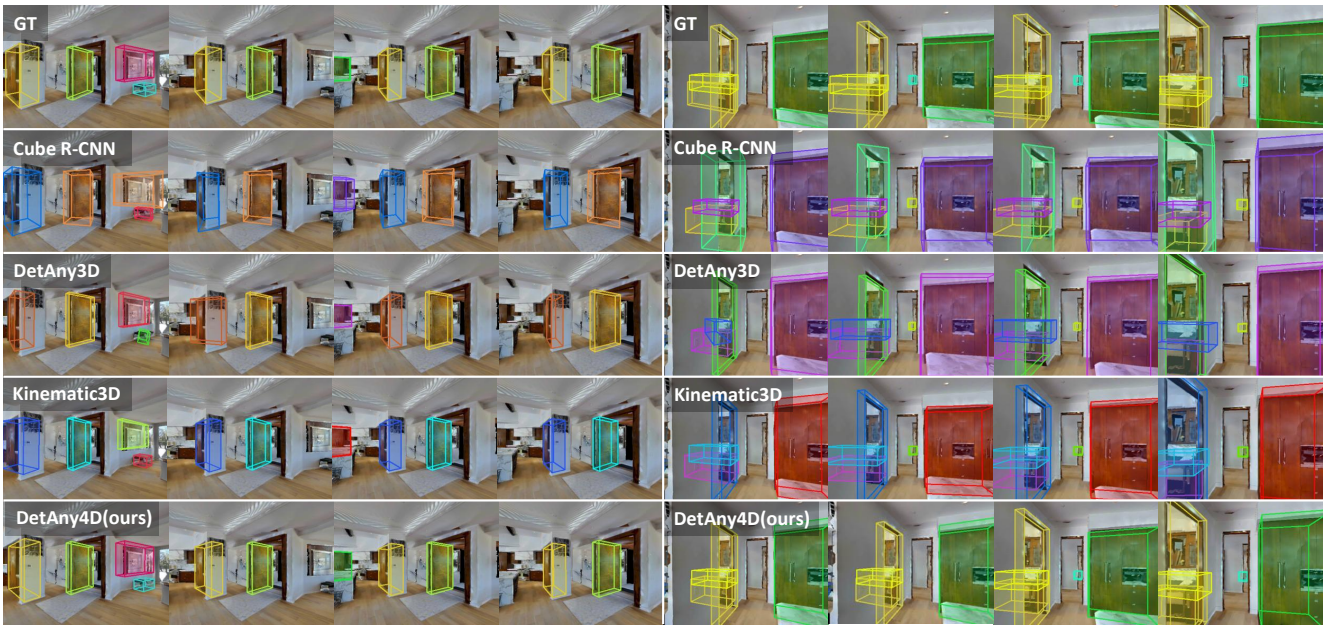

Figure 18. More qualitative comparison results.

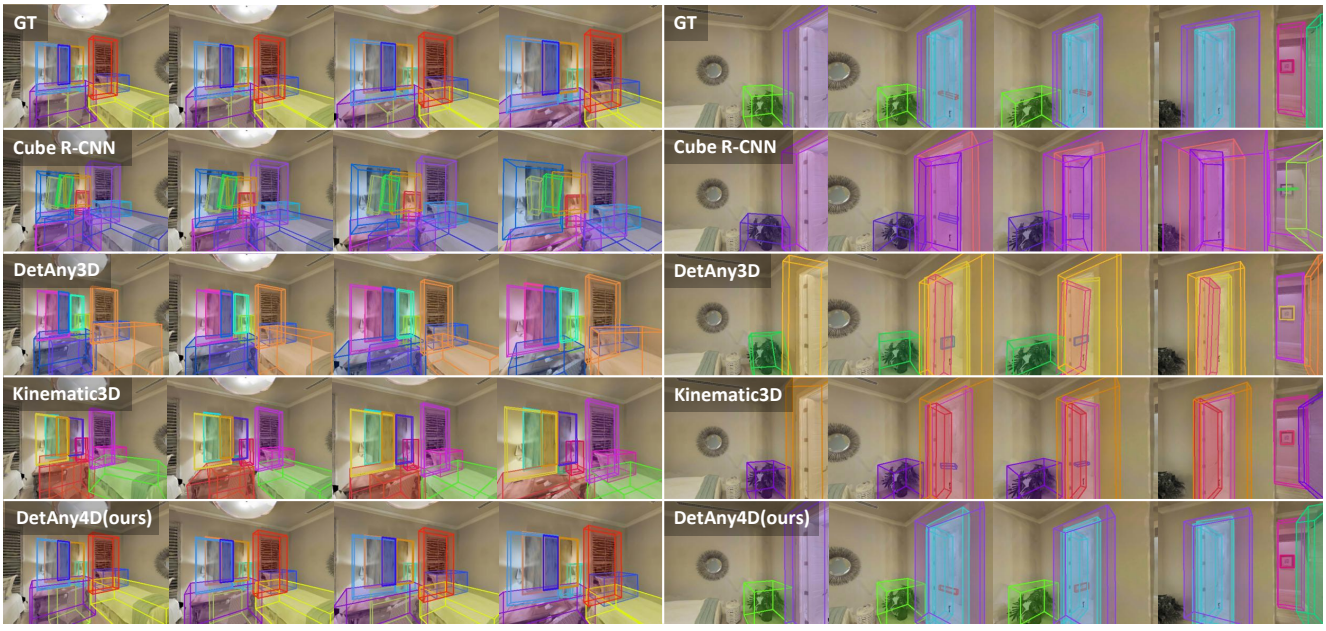

Figure 19. More qualitative comparison results.

CVPR
#8235

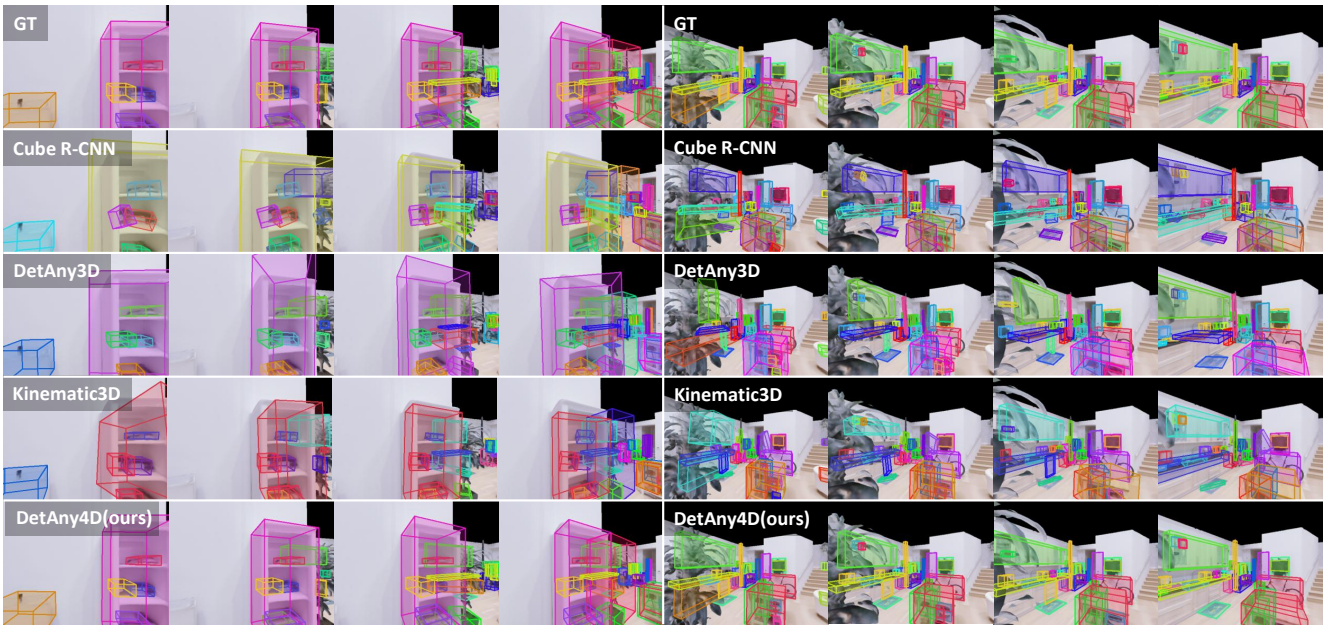

Figure 20. More qualitative comparison results.

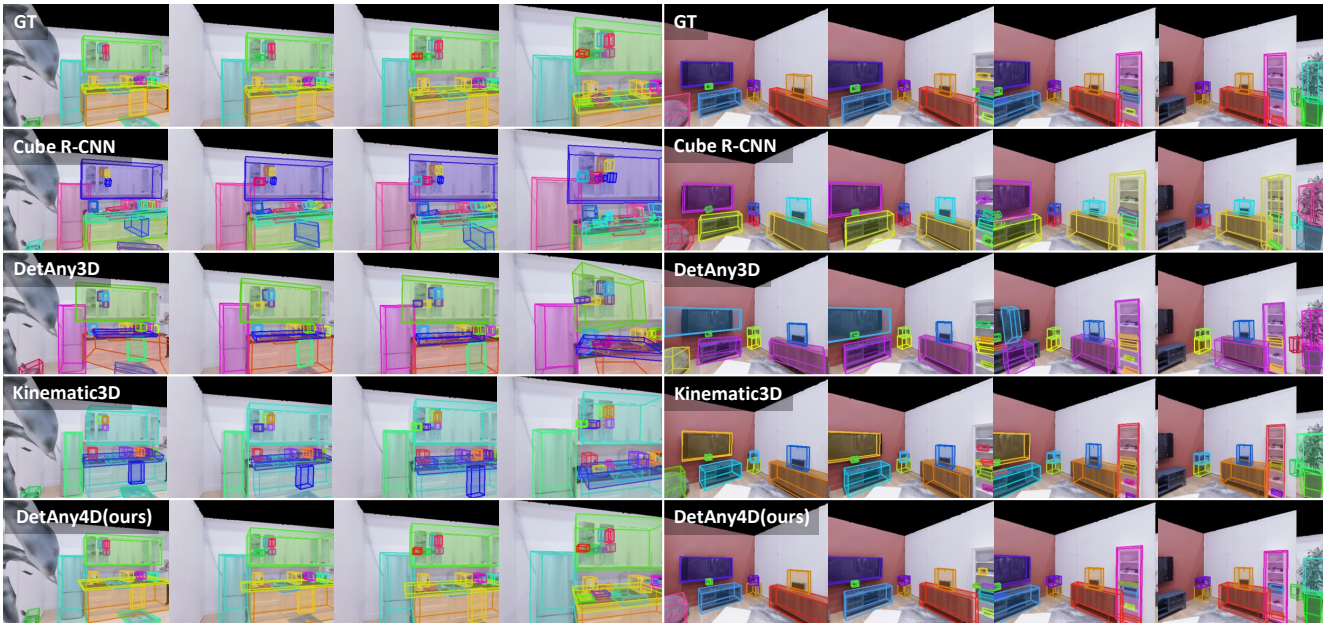

Figure 21. More qualitative comparison results.