Bridged Transformer for Vision and Point Cloud 3D Object Detection (Appendix)

Yikai Wang[1] TengQi Ye[2] Lele Cao[1] Wenbing Huang[3]
Fuchun Sun[1]✉ Fengxiang He[4] Dacheng Tao[4]

[1]Beijing National Research Center for Information Science and Technology (BNRist),
State Key Lab on Intelligent Technology and Systems,
Department of Computer Science and Technology, Tsinghua University [2]ByteDance Inc.
[3]Institute for AI Industry Research (AIR), Tsinghua University [4]JD Explore Academy, JD.com

wangyk17@mails.tsinghua.edu.cn, yetengqi@gmail.com, caolele@gmail.com, hwenbing@126.com,
fuchuns@tsinghua.edu.cn, fengxiang.f.he@gmail.com, dacheng.tao@gmail.com

A. Implementation Details

On both datasets, our implementation for the point cloud mostly follows implementation settings in [1, 2]. This part provides additional implementation details.

Data augmentation for point clouds. For SUN RGB-D, we use 20k points as input for each point cloud, and these points are randomly sampled from the depth image. For ScanNetV2, we adopt 50k points as input, which are randomly sampled from the scanned point cloud. We augment sampled points by flipping along the YZ plane with the probability of 50%. Note that the 3D box labels are also flipped or rotated following the augmentation of points. Besides, to preserve the lifting relations after the augmentation of points, we define an augmented matrix as

$$\boldsymbol{A} = \begin{bmatrix} \cos(\theta)\cdot\mathbb{I}_{\text{flip}}\cdot s & \sin(\theta)\cdot\mathbb{I}_{\text{flip}} & 0 \\ -\sin(\theta) & \cos(\theta)\cdot s & 0 \\ 0 & 0 & s \end{bmatrix}, \quad (1)$$

where θ is the rotation angle of points along the Z-axis; $\mathbb{I}_{\text{flip}}$ denotes the indicator function which is 1 if points are flipped along the YZ plane, otherwise -1; s is the overall scaling ratio of the point cloud.

The augmented extrinsic matrix after data augmentation is then calculated as $\boldsymbol{R}'_t = \boldsymbol{A}^\top \boldsymbol{R}_t$. We do not further apply data augmentations to the image part.

Lifting in the multi-view scenario. As mentioned in our main paper, for ScanNetV2, we use depths to filter out the projected 3D points which should be occluded, but visible due to the sparsity of the point cloud. In Fig. 4, we depict the projections of 3D points to the corresponding 2D image views, where the 3D bounding boxes are treated as 8 points, which are projected to obtain 2D bounding boxes.

✉ Corresponding author: Fuchun Sun.

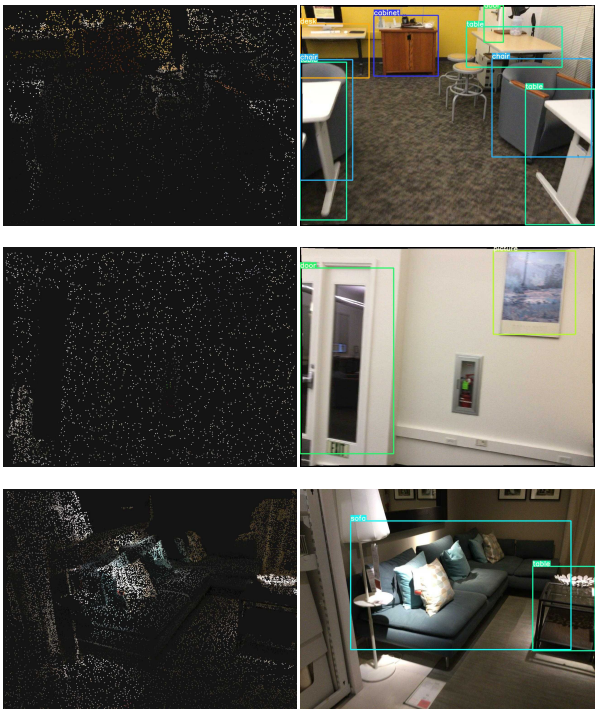

Figure 4. Projecting points and 3D bounding boxes to multi-view images on the ScanNetV2 dataset. Best viewed with zoom in.

B. Additional Results and Visualizations

In Fig. 5, we visualize the predicted 3D detection boxes on the ScanNetV2 to compare, and we also provide predicted 2D boxes on two example image views.

In Fig. 6, we visualize the attention weights w.r.t. corresponding object queries of points and image patches. We observe that the 68-th object token detects the table from 3D

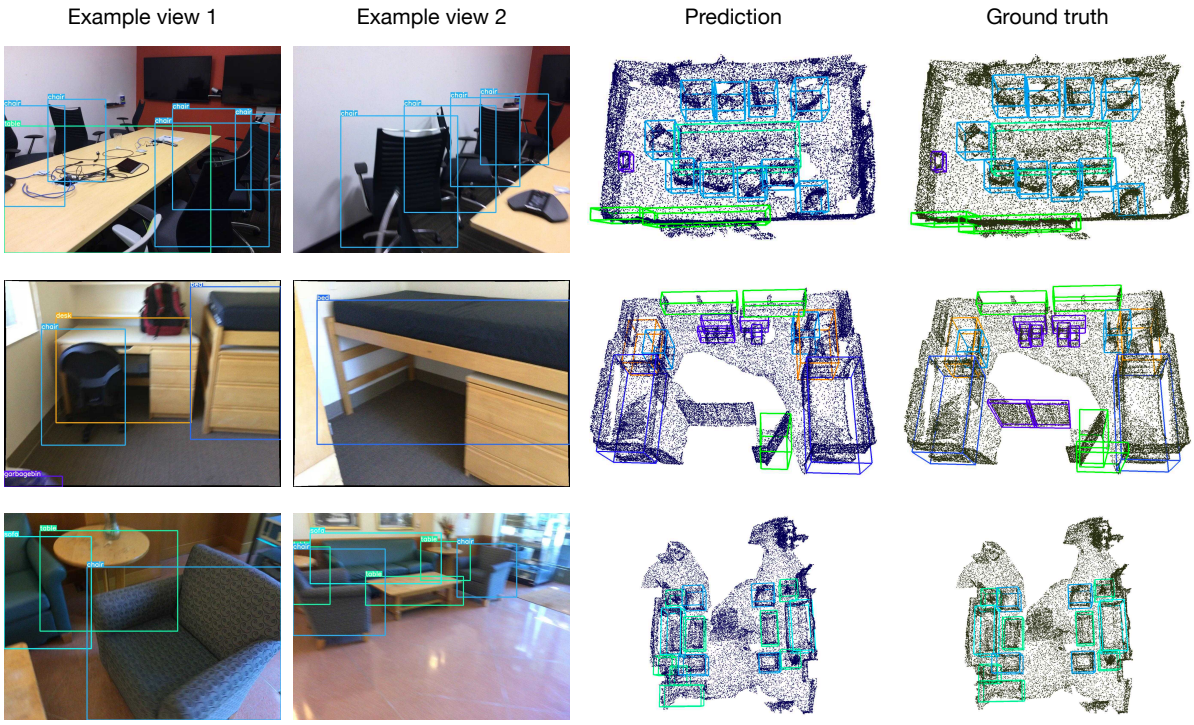

Figure 5. Additional visualizations on the ScanNetV2 dataset based on multi-view images and the point cloud as input. Two predicted image views are provided as examples. Best viewed with zoom in.

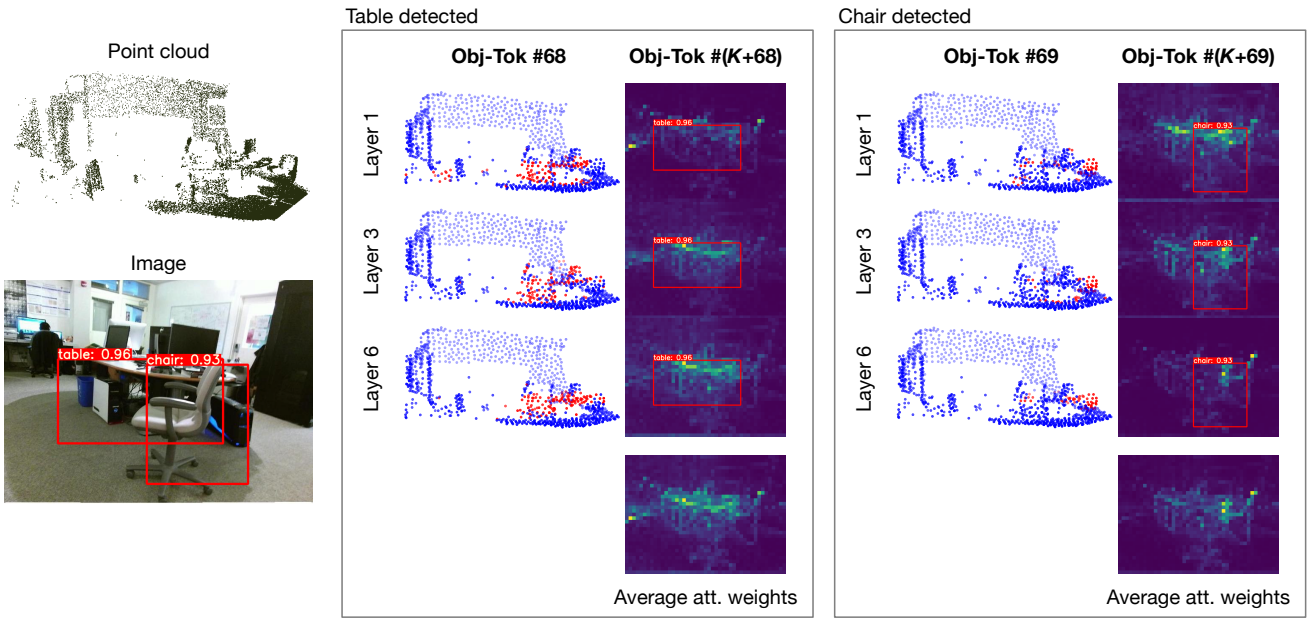

Figure 6. Attention weights w.r.t. corresponding object queries of points and image patches. The points with larger attention weights than 2×10^{-3} are colored as red, otherwise blue. Best viewed in color with zoom in.

points, and the $K+68$-th object token detects table from 2D image patches. Similarly, the 69-th and $K+69$-th object tokens both detect the chair from 3D points and 2D image patches, respectively. Such alignment demonstrates the effectiveness of our bridging technique by using conditional object queries (proposed in Sec. 3.3).

References

[1] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. *arXiv preprint arXiv:2104.00678*, 2021. 1

[2] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, 2019. 1

[3] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers and distillation through attention. In *ICML*, 2021.