Appendix of FCAF3D

Danila Rukhovich, Anna Vorontsova, and Anton Konushin

Samsung AI Center, Moscow
{d.rukhovich, a.vorontsova, a.konushin}@samsung.com

A Additional Comments on Mobius Parametrization

Comments on Eq. 3. The OBB heading angle θ is typically defined as an angle between x-axis and a vector towards a center of one of OBB faces. If a frontal face exists, then θ is defined unambiguously; however, this is not the case for some indoor objects. If a frontal face cannot be chosen unequivocally, there are four possible representations for a single OBB. The heading angle describes a rotation within the xy plane around z-axis w.r.t. the OBB center. Therefore, the OBB center $(x,\ y,\ z)$, height h, and the OBB size $s = w + l$ are the same for all representations. Meanwhile, the ratio $q = \frac{w}{l}$ of the frontal and lateral OBB faces and the heading angle θ do vary. Specifically, there are four options for the heading angle: $\theta,\ \theta + \frac{\pi}{2},\ \theta + \pi,\ \theta + \frac{3\pi}{2}$. Swapping frontal and lateral faces gives two ratio options: q and $\frac{1}{q}$. Overall, there are four different tuples (q, θ) for the same OBB:

$$(q, \theta)\,,\ \left(\frac{1}{q},\ \theta + \frac{\pi}{2}\right),\ (q, \theta + \pi)\,,\ \left(\frac{1}{q}, \theta + \frac{3\pi}{2}\right).$$

Verification of Eq. 4. Here, we prove that four different representations of the same OBB from Eq. 3 map to the same point on a Mobius strip by Eq. 4.

$$\begin{aligned}
(q,\ \theta) &\mapsto (\ln(q)\sin(2\theta),\ \ln(q)\cos(2\theta),\ \sin(4\theta),\ \cos(4\theta)) \\
\left(\frac{1}{q},\ \theta + \frac{\pi}{2}\right) &\mapsto (\ln(\tfrac{1}{q})\sin(2\theta + \pi),\ \ln(\tfrac{1}{q})\cos(2\theta + \pi), \sin(4\theta + 2\pi),\ \cos(4\theta + 2\pi)) \\
&= (\ln(q)\sin(2\theta),\ \ln(q)\cos(2\theta),\ \sin(4\theta),\ \cos(4\theta)) \\
(q,\ \theta + \pi) &\mapsto (\ln(q)\sin(2\theta + 2\pi),\ \ln(q)\cos(2\theta + 2\pi), \sin(4\theta + 4\pi),\ \cos(4\theta + 4\pi)) \\
&= (\ln(q)\sin(2\theta),\ \ln(q)\cos(2\theta),\ \sin(4\theta),\ \cos(4\theta)) \\
\left(\frac{1}{q},\ \theta + \frac{3\pi}{2}\right) &\mapsto (\ln(\tfrac{1}{q})\sin(2\theta + 3\pi),\ \ln(\tfrac{1}{q})\cos(2\theta + 3\pi), \sin(4\theta + 6\pi),\ \cos(4\theta + 6\pi)) \\
&= (\ln(q)\sin(2\theta),\ \ln(q)\cos(2\theta),\ \sin(4\theta),\ \cos(4\theta))
\end{aligned}$$

B Per-category results

ScanNet. Tab. 1 contains per-category AP@0.25 scores for 18 object categories for the ScanNet dataset. For 12 out of 18 categories, FCAF3D outperforms other methods. The largest quality gap can be observed for *window* (60.2 against 53.7), *picture* (29.9 against 18.6), and *other furniture* (65.4 against 56.4) categories.

Method	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	mAP
VoteNet[3]	36.3	87.9	88.7	89.6	58.8	47.3	38.1	44.6	7.8	56.1	71.7	47.2	45.4	57.1	94.9	54.7	92.1	37.2	58.7
GSDN[1]	41.6	82.5	92.1	87.0	61.1	42.4	40.7	51.5	10.2	64.2	71.1	54.9	40.0	70.5	**100**	75.5	93.2	53.1	62.8
H3DNet[4]	49.4	88.6	91.8	90.2	64.9	61.0	51.9	54.9	18.6	62.0	75.9	57.3	57.2	75.3	97.9	67.4	92.5	53.6	67.2
GroupFree[2]	52.1	**92.9**	93.6	88.0	**70.7**	60.7	53.7	62.4	16.1	58.5	**80.9**	**67.9**	47.0	76.3	99.6	72.0	**95.3**	56.4	69.1
FCAF3D	**57.2**	87.0	**95.0**	**92.3**	70.3	**61.1**	**60.2**	**64.5**	**29.9**	**64.3**	71.5	60.1	**52.4**	**83.9**	99.9	**84.7**	86.6	**65.4**	**71.5**

Table 1. Per-category AP@0.25 scores for 18 object categories from the ScanNet dataset.

Tab. 2 shows per-category AP@0.5 scores. According to the reported values, FCAF3D is the best at detecting objects of 13 out of 18 categories. The most significant improvement is achieved for *cabinet* (35.8 against 26.0), *sofa* (85.2 against 70.7), *picture* (17.9 against 7.8), shower (64.2 against 44.1), and *sink* (52.6 against 37.4).

Method	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	mAP
VoteNet[3]	8.1	76.1	67.2	68.8	42.4	15.3	6.4	28.0	1.3	9.5	37.5	11.6	27.8	10.0	86.5	16.8	78.9	11.7	33.5
GSDN[1]	13.2	74.9	75.8	60.3	39.5	8.5	11.6	27.6	1.5	3.2	37.5	14.1	25.9	1.4	87.0	37.5	76.9	30.5	34.8
H3DNet[4]	20.5	79.7	80.1	79.6	56.2	29.0	21.3	45.5	4.2	33.5	50.6	37.3	41.4	37.0	89.1	35.1	90.2	35.4	48.1
GroupFree[2]	26.0	81.3	82.9	70.7	**62.2**	41.7	26.5	55.8	7.8	**34.7**	**67.2**	43.9	44.3	44.1	**92.8**	37.4	**89.7**	40.6	52.8
FCAF3D	**35.8**	**81.5**	**89.8**	**85.0**	62.0	**44.1**	**30.7**	**58.4**	**17.9**	31.3	53.4	**44.2**	**46.8**	**64.2**	91.6	**52.6**	84.5	**57.1**	**57.3**

Table 2. AP@0.5 scores for 18 object categories from the ScanNet dataset.

SUN RGB-D. Per-category AP@0.25 scores for the 10 most common object categories for the SUN RGB-D benchmark are reported in Tab. 3. Compared to other methods, FCAF3D is more accurate at detecting objects of 7 out of 10 categories. In this experiment, the quality gap is not so dramatic: it equals 4.1 % for *desk* and 5.2 % for *night stand*; for the rest categories, it does not exceed 2 %. FCAF3D achieves a 1.2 % better mAP@0.25 compared to the closest competitor GroupFree.

Method	bath	bed	bkshf	chair	desk	dresser	nstand	sofa	table	toilet	mAP
VoteNet[3]	74.4	83.0	28.8	75.3	22.0	29.8	62.2	64.0	47.3	90.1	57.7
H3DNet[4]	73.8	85.6	31.0	76.7	29.6	33.4	65.5	66.5	50.8	88.2	60.1
GroupFree[2]	**80.0**	87.8	32.5	79.4	32.6	36.0	66.7	**70.0**	**53.8**	91.1	63.0
FCAF3D	79.0	**88.3**	**33.0**	**81.1**	**34.0**	**40.1**	**71.9**	69.7	53.0	**91.3**	**64.2**

Table 3. AP@0.25 scores for 10 object categories from the SUN RGB-D dataset.

For SUN RGB-D, the superiority of the proposed method is more noticeable when analyzing on per-category AP@0.5. As shown in Tab. 4, FCAF3D outperforms the competitors for 9 out of 10 object categories. For some categories, there is a significant margin: e.g., 30.1 against 21.9 for *dresser*, 59.8 against 49.8 for *night stand*, and 35.5 against 29.2 for *table*. Respectively, FCAF3D surpasses other methods by more than 3.5 % in terms of mAP@0.5.

Method	bath	bed	bkshf	chair	desk	dresser	nstand	sofa	table	toilet	mAP
H3DNet[4]	47.6	52.9	8.6	60.1	8.4	20.6	45.6	50.4	27.1	69.1	39.0
GroupFree[2]	64.0	67.1	**12.4**	62.6	14.5	21.9	49.8	**58.2**	29.2	72.2	45.2
FCAF3D	**66.2**	**69.8**	11.6	**68.8**	**14.8**	**30.1**	**59.8**	**58.2**	**35.5**	**74.5**	**48.9**

Table 4. AP@0.5 scores for 10 object categories from the SUN RGB-D dataset.

S3DIS. The results of the proposed method in comparison with GSDN are presented in Tab. 5 and Tab. 6. In terms of AP@0.25, FCAF3D is far more accurate when detecting *sofas*, *bookcases*, and *whiteboards*. Most notably, FCAF3D achieves an impressive AP@0.25 of 92.4 for the *sofa* category, leaving GSDN with AP@0.25 of 20.8 far behind. The difference in mAP in favor of the proposed method is almost 19 %.

Method	table	chair	sofa	bkcase	board	mAP
GSDN[1]	**73.7**	**98.1**	20.8	33.4	12.9	47.8
FCAF3D	69.7	97.4	**92.4**	**36.7**	**37.3**	**66.7**

Table 5. Per-category AP@0.25 scores for 5 object categories from the S3DIS dataset.

In terms of AP@0.5, FCAF3D outperforms GSDN by a large margin for each category. Similar to AP@0.25, the accuracy gap for the *sofa* category is the most dramatic: with an AP@0.25 of 70.1, FCAF3D is an order of magnitude more accurate than GSDN, which has only 6.1. Accordingly, FCAF3D has an approximately 1.8 times larger mAP compared to GSDN.

Method	table	chair	sofa	bkcase	board	mAP
GSDN[1]	36.6	75.3	6.1	6.5	1.2	25.1
FCAF3D	**45.4**	**88.3**	**70.1**	**19.5**	**5.6**	**45.9**

Table 6. AP@0.5 scores for 5 object categories from the S3DIS dataset.

C Visualization

This section contains additional visualizations of the results of 3D object detection for all three benchmarks. The ground truth and estimated 3D object bounding boxes are drawn over the corresponding point clouds. Objects of different categories are marked with different colors.

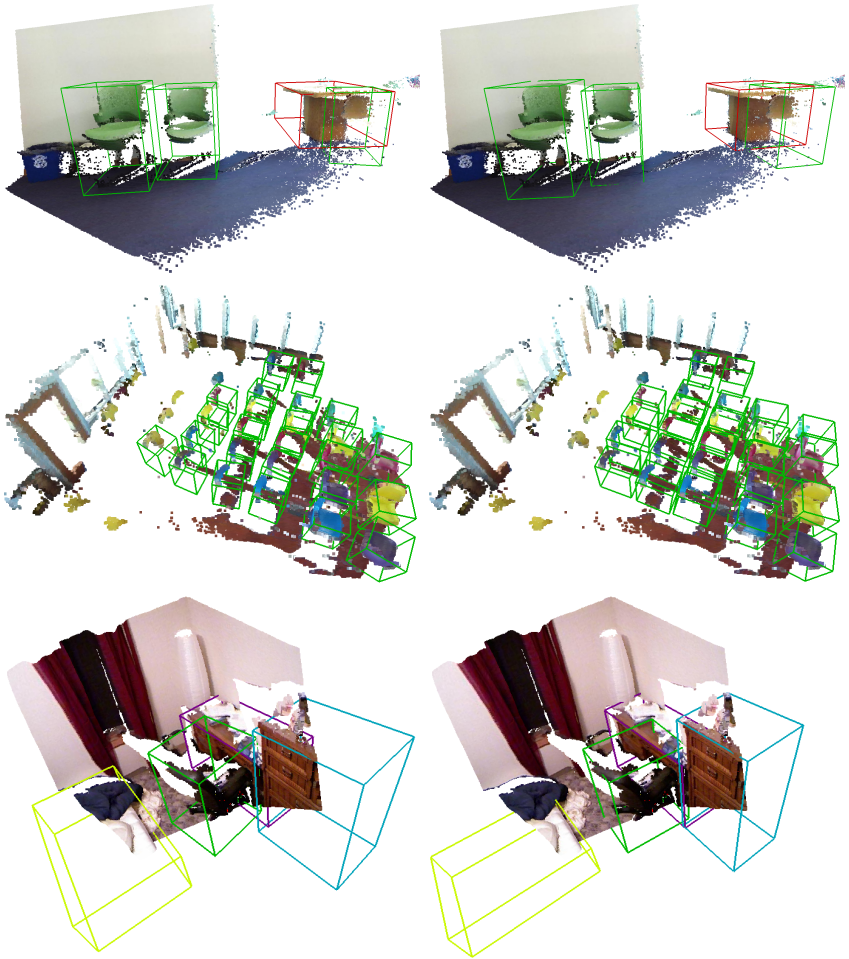

Fig. 1. The point cloud from SUN RGB-D with OBBs. The color of a bounding box denotes the object category: **bed**, **chair**, **desk**, **dresser**, **table** (only categories that are present in the pictures are listed). Left: estimated with FCAF3D, right: ground truth.

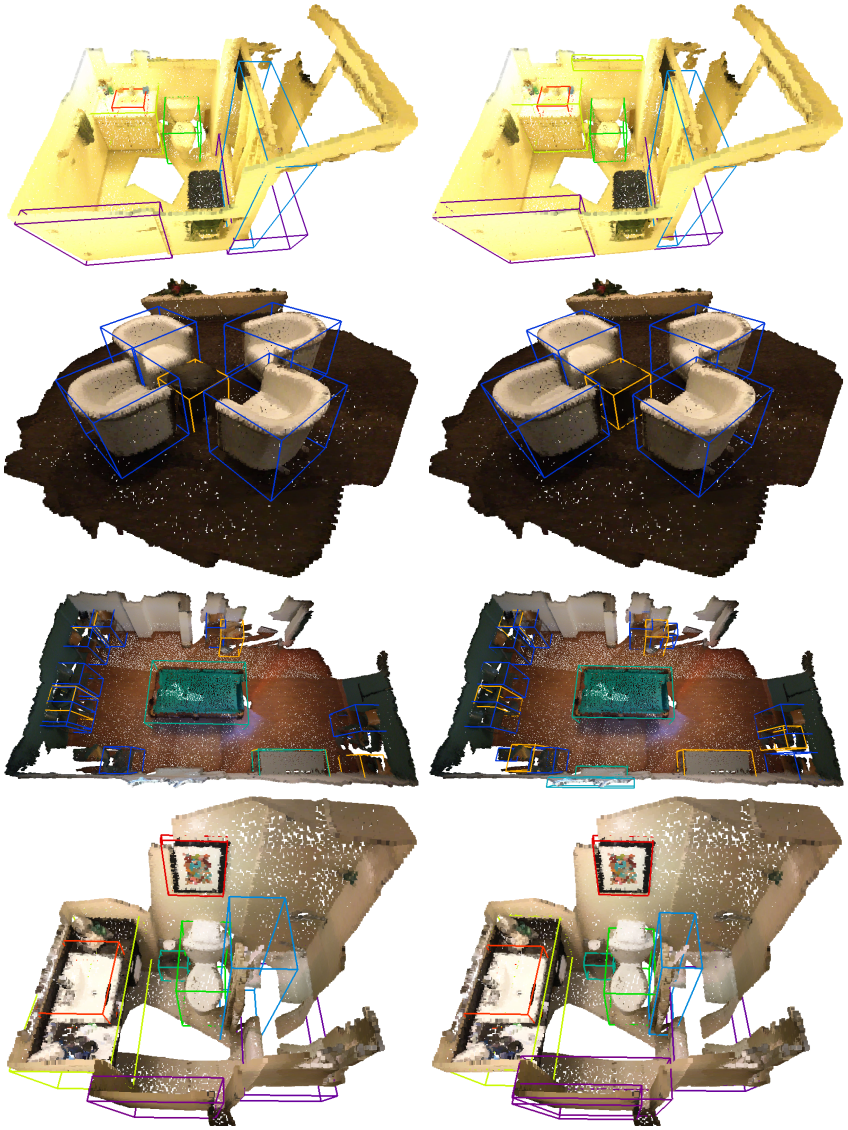

Fig. 2. The point cloud from ScanNet with AABBs. The color of a bounding box denotes the object category: **cabinet**, **chair**, **sofa**, **table**, **door**, **window**, **bookshelf**, **picture**, **counter**, **desk**, **shower curtain**, **toilet**, **sink**, **bathtub**, **other furniture** (only categories that are present in the pictures are listed). Left: estimated with FCAF3D, right: ground truth.

Fig. 3. The point cloud from S3DIS with AABBs. The color of a bounding box denotes the object category: **table**, **chair**, **sofa**, **bookcase**, **whiteboard**. Left: estimated with FCAF3D, right: ground truth.

References

1. Gwak, J., Choy, C., Savarese, S.: Generative sparse detection networks for 3d single-shot object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 297–313. Springer (2020) 2, 3
2. Liu, Z., Zhang, Z., Cao, Y., Hu, H., Tong, X.: Group-free 3d object detection via transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2949–2958 (2021) 2, 3
3. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9277–9286 (2019) 2
4. Zhang, Z., Sun, B., Yang, H., Huang, Q.: H3dnet: 3d object detection using hybrid geometric primitives. In: European Conference on Computer Vision. pp. 311–329. Springer (2020) 2, 3