H^2A^2: Homogeneity-Aware and Heterogeneity-Aware Feature Perception for Unified Indoor 3D Object Detection

Supplementary Material

In this supplementary material, we provide additional details and results to further understand and validate our proposed H^2A^2 framework. Sec. A presents additional training details, including optimization settings and loss functions. Sec. B elaborates on the optimization strategy for the non-differentiable binarization function in our gating module. Sec. C reports and analyzes the per-category detection results on all three benchmarks. Sec. F provides additional qualitative visualizations of the 3D detection results.

A. Additional Training Details

We train a separate detection head for each dataset, following the TR3D [35] implementation on top of MMDetection3D [8]. All models share the same Minkowski-ResNet-34 backbone and separate heads as described in Sec. 4. Our network are optimized on three NVIDIA RTX 4090 GPUs.

Optimization and schedules. We use AdamW with an initial learning rate $\eta_0 = 3 \times 10^{-3}$, weight decay 1×10^{-2}, and gradient clipping at a global norm of 10. The learning rate is linearly warmed up from 0 to η_0 during the first 2,750 iterations and then annealed to 10^{-5} with a cosine schedule, for a total of 30,000 iterations on all three datasets. During joint training on ScanNet v2 [9], SUN RGB-D [36], and S3DIS [1], we use the same per-GPU batch size for all three datasets. To align the effective dataset lengths in this multi-dataset training setting, we wrap ScanNet v2 and SUN RGB-D with RepeatDataset (5 repeats each), while S3DIS is used without repetition, so that each dataset contributes a comparable number of iterations during training. The data augmentation pipeline, which includes point sampling, random flips, global rotation, scaling, translation, and color normalization, is identical to that described in Sec. 4.

Computational Overhead. H^2A^2 is an unified detector while TR3D is trained separately for each dataset. Thus, for parameter count, we report the total parameters required for training the three datasets in the following table. H^2A^2 reduces the total model size by reusing shared parameters. For training time, we report the total time required to train the three datasets on three 4090 GPUs. We report the training GPU memory overhead in the following table. The results are shown in Tab. 6

Method	Params-t (M)	Mem (GB)	Training Time
TR3D	44.0	8.6	4.6 hours
H^2A^2	40.0	10.2	5.9 hours

Table 6. Comparison of parameters, memory consumption, and training time between TR3D and H^2A^2.

Loss functions. We adopt the same loss formulation as the baseline method Tr3D. For the classification branch, we adopt the sigmoid focal loss, using the FocalLoss implementation in MMDetection, in order to better handle the foreground–background class imbalance. Formally, let $\mathbf{s}_i \in \mathbb{R}^C$ denote the classification logits for sample i, $\mathbf{p}_i = \sigma(\mathbf{s}_i)$ the corresponding predicted probabilities, and $\mathbf{y}_i \in 0, 1^C$ the associated one-hot ground-truth label. For each class $c \in 1, \ldots, C$, we define

$$\mathbf{p}_{i,c}^t = \begin{cases} \mathbf{p}_{i,c}, & \mathbf{y}_{i,c} = 1, \\ 1 - \mathbf{p}_{i,c}, & \mathbf{y}_{i,c} = 0, \end{cases} \tag{10}$$

and the classification loss is given by

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N_{\text{pos}}} \sum_{i=1}^{N} \sum_{c=1}^{C} \alpha_c, (1 - \mathbf{p}_{i,c}^t)^\gamma \log \mathbf{p}_{i,c}^t, \tag{11}$$

where N_{pos} denotes the number of positive samples. Following the default MMDetection3D configuration, we set the focusing parameter $\gamma = 2$ and the class-balancing factor $\alpha_c = 0.25$ for positive classes.

For bounding-box regression on ScanNet v2 and S3DIS we use the axis-aligned IoU-based loss TR3D AxisAlignedIoULoss. Given a predicted box B_i^{pred} and its matched ground-truth box B_i^{gt}, the 3D IoU is

$$\text{IoU}(B_i^{\text{pred}}, B_i^{\text{gt}}) = \frac{V(B_i^{\text{pred}} \cap B_i^{\text{gt}})}{V(B_i^{\text{pred}} \cup B_i^{\text{gt}})}, \tag{12}$$

where $V(\cdot)$ denotes the box volume. The regression loss on axis-aligned boxes is

$$\mathcal{L}_{\text{box}} = \frac{1}{N_{\text{pos}}} \sum_{i=1}^{N_{\text{pos}}} \big(1 - \text{IoU}(B_i^{\text{pred}}, B_i^{\text{gt}})\big), \tag{13}$$

which matches the implementation of TR3D AxisAlignedIoULoss in our code.

For SUN RGB-D we instead adopt the rotated 3D IoU-based loss TR3D RotatedIoU3DLoss, operating on oriented boxes:

$$\mathcal{L}_{\text{box}} = \frac{1}{N_{\text{pos}}} \sum_{i=1}^{N_{\text{pos}}} \left(1 - \text{IoU}_{\text{3D}}(B_i^{\text{pred}}, B_i^{\text{gt}}) + \frac{\|\mathbf{c}_i^{\text{pred}} - \mathbf{c}_i^{\text{gt}}\|_2^2}{\|\mathbf{d}_i\|_2^2}\right), \tag{14}$$

Method	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	mAP
VoteNet[32]	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[14]	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	100.0	75.5	93.2	53.1	62.8
H3DNet[45]	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[25]	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[34]	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
TR3D[35]	55.9	87.8	95.6	89.3	73.7	62.9	57.8	65.9	29.8	68.5	82.7	61.9	60.2	83.4	99.9	79.7	91.6	64.8	72.9
Ours	**67.6**	**93.6**	**96.0**	91.6	**77.2**	**68.6**	**61.0**	**68.7**	**47.5**	**71.4**	**86.3**	**65.1**	**70.4**	77.9	98.6	**88.1**	89.6	**75.4**	**77.5**

Table 7. Per-category AP@0.25 scores for 18 object categories from the ScanNet v2 dataset.

Method	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	mAP
VoteNet[32]	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[14]	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[45]	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[25]	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[34]	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
TR3D[35]	38.6	82.9	90.5	83.2	63.5	44.8	31.4	60.1	26.2	36.6	65.3	40.2	55.7	59.5	**96.0**	52.8	80.8	58.9	59.3
Ours	**50.2**	**91.2**	**91.4**	**85.1**	**68.2**	**49.9**	**36.7**	**63.6**	**31.5**	**44.6**	**69.1**	**42.5**	**58.7**	**58.2**	93.2	**55.6**	**89.4**	**69.8**	**63.8**

Table 8. Per-category AP@0.5 scores for 18 object categories from the ScanNet v2 dataset.

where IoU_{3D} is the volumetric IoU of rotated boxes, $\mathbf{c}_i^{\text{pred}}$ and $\mathbf{c}_i^{\text{gt}}$ are their centers, and $\mathbf{d}_i$ is the diagonal of the minimal enclosing box of the two.

The total detection loss used during training is

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}}, \tag{15}$$

with unit weights for the classification and regression terms.

B. Optimization for Non-Differentiable Binarization Function.

In the backward pass of the network, the gradient of the composite descriptor $z_j = \alpha_j \beta_j$ is written by the chain rule as:

$$\frac{\partial L}{\partial z_j} = \frac{\partial L}{\partial \gamma_j} \frac{\partial \gamma_j}{\partial z_j}. \tag{16}$$

The discriminative score is produced by hard binarization,

$$\gamma_j = \text{Binarization}\big(z_j; \tau\big) \in \{0, 1\}, \tag{17}$$

and it modulates a linear combination of two weight branches,

$$W_j = \gamma_j \, W_j^{\text{sh}} + (1 - \gamma_j) \, W_j^{\text{ex}}. \tag{18}$$

Because this binarization is defined by a discontinuous indicator function, the operation is non-differentiable and the chain rule above does not provide usable gradients for z_j. To cope with this problem, we utilize a straight gradient estimator [2] : the hard binarization is kept unchanged in the forward computation, and its derivative with respect to the composite descriptor is approximated by the identity in the backward pass. Under this approximation, we modify the gradient as

$$\frac{\partial L}{\partial z_j} \approx \frac{\partial L}{\partial \gamma_j}. \tag{19}$$

When the discriminative score is applied in the linear combination above, its gradient is obtained by standard back-propagation,

$$\frac{\partial L}{\partial \gamma_j} = \Big\langle \frac{\partial L}{\partial W_j}, \, W_j^{\text{sh}} - W_j^{\text{ex}} \Big\rangle, \tag{20}$$

where $\langle A, B \rangle$ denotes the Frobenius inner product (sum of element-wise products). Propagating this surrogate gradient to the composite descriptor restores learnability of the underlying factors,

$$\frac{\partial L}{\partial \alpha_j} \approx \beta_j \frac{\partial L}{\partial z_j}, \qquad \frac{\partial L}{\partial \beta_j} \approx \alpha_j \frac{\partial L}{\partial z_j}. \tag{21}$$

This straight-through treatment enables gradient-based optimization while preserving a hard binary score at inference.

C. Per-category results

ScanNet. Tab. 7 contains per-category AP@0.25 scores for 18 object categories on the ScanNet v2 dataset. For 13 out of 18 categories, our method outperforms all previous approaches and achieves the highest mAP. The largest quality gaps over the strongest baseline TR3D can be observed for *cabinet* (67.6 against 55.9), *picture* (47.5 against 29.2), *other furniture* (75.4 against 64.3), *refrigerator* (70.4 against 58.2), and *sink* (88.1 against 77.7).

Method	bath	bed	bkshf	chair	desk	dresser	nstand	sofa	table	toilet	mAP
VoteNet[32]	74.4	83.0	28.8	75.3	22.0	29.8	62.2	64.0	47.3	90.1	57.7
H3DNet[45]	73.8	85.6	31.0	76.7	29.6	33.4	65.5	66.5	50.8	88.2	60.1
GroupFree[25]	80.0	87.8	32.5	79.4	32.6	36.0	66.7	70.0	**53.8**	91.1	63.0
FCAF3D[34]	79.0	**88.3**	33.0	81.1	34.0	40.1	**71.9**	69.7	53.0	**91.3**	64.2
TR3D[35]	88.4	59.8	71.3	82.3	**92.5**	43.8	45.3	71.2	36.8	79.6	67.1
Ours	**88.7**	61.4	**72.1**	**84.6**	92.3	**50.8**	48.1	**73.9**	36.4	72.1	**68.0**

Table 9. Per-category AP@0.25 scores for 10 object categories from the SUN RGB-D dataset.

Method	bath	bed	bkshf	chair	desk	dresser	nstand	sofa	table	toilet	mAP
H3DNet[45]	47.6	52.9	8.6	60.1	8.4	20.6	45.6	50.4	27.1	69.1	39.0
GroupFree[25]	64.0	67.1	12.4	62.6	14.5	21.9	49.8	58.2	29.2	72.2	45.2
FCAF3D[34]	66.2	**69.8**	11.6	68.8	14.8	**30.1**	**59.8**	58.2	**35.5**	**74.5**	48.9
TR3D[35]	68.3	38.3	**60.1**	69.4	**74.4**	21.6	33.7	60.8	12.4	64.8	50.4
Ours	**69.0**	41.3	59.8	**72.5**	73.9	26.9	37.4	**61.8**	14.3	56.8	**51.4**

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

Tab. 8 shows per-category AP@0.5 scores. According to the reported values, our method is the best at detecting objects of 14 out of 18 categories and improves the overall mAP@0.5 from 59.3 of TR3D to 63.8. The most significant improvements over TR3D are achieved for *cabinet* (50.2 against 37.6), *bed* (91.2 against 82.2), *table* (68.2 against 62.2), *counter* (44.6 against 36.1), and *other furniture* (69.8 against 58.5), while maintaining competitive performance on the remaining categories.

SUN RGB-D. Per-category AP@0.25 scores for the 10 most common object categories on the SUN RGB-D benchmark are reported in Tab. 9. Compared to other methods, our approach attains the highest AP@0.25 for 5 out of 10 categories and achieves the best mAP. The most prominent gains over TR3D are observed for *dresser* (50.8 against 40.8), *chair* (84.6 against 81.3), *sofa* (73.9 against 71.2), and *bookcase* (72.1 against 71.3). For the remaining categories, our method remains competitive with FCAF3D and TR3D.

For SUN RGB-D, the advantages of our method become more noticeable when analyzing per-category AP@0.5, as shown in Tab. 10. Our method delivers the highest mAP@0.5 among all approaches, improving TR3D from 50.4 to 51.3. It also achieves the best performance for *bath* (69.0), *chair* (72.5), and *sofa* (61.8). TR3D and FCAF3D remain strong baselines on the *bed*, *night stand*, *table*, and *desk* categories, indicating that our improvements mainly come from more robust recognition of several key object classes.

S3DIS. The results of our method in comparison with GSDN, FCAF3D, and TR3D are presented in Tab. 12 and Tab. 13. In terms of AP@0.25, our method achieves the best mAP and is markedly more accurate when detecting *table*, *bookcase*, and *whiteboard*. In particular, we obtain an AP@0.25 of 66.9 for the *bookcase* category (against 40.6 of TR3D and 36.7 of FCAF3D) and 75.2 for *whiteboard* (against 60.5 of TR3D), leading to a 4.2 % mAP gain over the strongest baseline TR3D (78.7 against 74.5).

In terms of AP@0.5, our method consistently outperforms all competitors in overall mAP, improving TR3D from 51.7 to 57.2, see Tab. 13. Similar to AP@0.25, the largest improvements are obtained for *table* (60.9 against 51.8), *bookcase* (50.6 against 18.7), and *whiteboard* (35.0 against 22.8). While TR3D still slightly outperforms our method on the *chair* category and FCAF3D retains the lead on *sofa*, our approach provides the highest overall accuracy and the most balanced performance across all categories.

D. Generalization Experiment.

To further demonstrate unification, we compare H^2A^2 with Unidet3D that also takes point as input and is also an unified 3D detector under multiple datasets. As shown in the Tab. 11, H^2A^2 also exhibit competitive performance. To validate the generalization performance of our proposed H^2A^2 across a wider range of datasets, we follow the official dataset setup of UniDet3D [20] and perform joint training on six diverse indoor point cloud scene datasets. We conduct a thorough comparison against two representative methods: TR3D [35], a method tailored for single-scene single-dataset settings, and UniDet3D [20], a unified point-based 3D detector designed for multi-dataset joint training. As shown in the table above, H^2A^2 achieves highly competitive performance compared with both counterparts.

E. Evaluation metrics.

For all datasets, we adopt mean average precision (mAP) as the primary evaluation metric under intersection-over-union

Methods	Input	S3DIS		ScanNet v2		ARKitScenes		MultiScan		3RScan		ScanNet++	
		$mAP_{0.25}$	$mAP_{0.5}$	$mAP_{0.25}$	$mAP_{0.5}$	$mAP_{0.25}$	$mAP_{0.5}$	$mAP_{0.25}$	$mAP_{0.5}$	$mAP_{0.25}$	$mAP_{0.5}$	$mAP_{0.25}$	$mAP_{0.5}$
UniDet3d	Geo	75.2	60.8	77.5	66.1	61.3	47.1	64.2	51.6	64.7	48.6	26.4	17.2
TR3D (trained on each dataset)	Geo	74.5	51.7	77.5	63.8	64.3	47.0	62.6	51.3	62.1	44.6	29.1	18.9
TR3D (jointly training)	Geo	69.7	50.0	70.1	55.7	63.2	46.5	59.4	42.2	63.7	48.1	28.7	18.1
H^2A^2	Geo	**79.8**	**62.4**	**78.9**	**64.1**	**64.8**	**48.0**	**67.3**	**56.4**	**64.6**	**49.3**	**30.6**	**22.2**

Table 11. Comparison of detection accuracy ($mAP_{0.25}$ and $mAP_{0.5}$) across multiple indoor point cloud datasets for different methods.

Method	table	chair	sofa	bkcase	board	mAP
GSDN[14]	73.7	98.1	20.8	33.4	12.9	47.8
FCAF3D[34]	69.7	97.4	92.4	36.7	37.3	66.7
TR3D[35]	73.9	**98.8**	**95.8**	41.9	61.9	74.5
Ours	**75.0**	96.3	80.4	**66.9**	**75.2**	**78.7**

Table 12. Per-category AP@0.25 scores for 5 object categories from the S3DIS dataset.

Method	table	chair	sofa	bkcase	board	mAP
GSDN[14]	36.6	75.3	6.1	6.5	1.2	25.1
FCAF3D[34]	45.4	88.3	70.1	19.5	5.6	45.9
TR3D[35]	54.3	89.5	**70.1**	19.2	25.3	51.7
Ours	**61.7**	**90.2**	57.4	**51.6**	**35.7**	**59.3**

Table 13. Per-category AP@0.5 scores for 5 object categories from the S3DIS dataset.

Method	ScanNet v2	SunRGBD	S3DIS
	$mAP_{0.05:0.5:0.05}$	$mAP_{0.05:0.5:0.05}$	$mAP_{0.05:0.5:0.05}$
TR3D	70.2	62.9	67.0
H^2A^2	**75.2**	**63.5**	**73.5**

Table 14. Accuracy results under the evaluation metric mAP_{3D}.

(IoU) thresholds of 0.25 and 0.5. To ensure statistical reliability and mitigate stochastic variations, each model is trained and evaluated five times independently.

To further validate the performance stability of our network for indoor 3D object detection, we conduct supplementary experiments using the AP3D evaluation metric proposed by Omni3D [5], which averages precision across IoU3D thresholds from 0.05 to 0.50 with a step of 0.05. This metric mitigates the noise sensitivity of conventional single-threshold metrics and enables a more comprehensive assessment of the model's overall detection capability. As shown in Tab. 14, our H2A2 network achieves the best AP3D performance on ScanNet v2, SUN RGB-D, and S3DIS datasets, fully demonstrating the detection accuracy and performance stability of our network.

F. Visualization

This section provides supplementary visualizations of the 3D object detection results across all three benchmark datasets. Both the ground truth labels and predicted 3D object bounding boxes are overlaid on the corresponding point clouds. Specifically, objects of different categories in the ground truth are uniformly marked in red, while the predicted 3D object bounding boxes are distinguished using distinct colors.

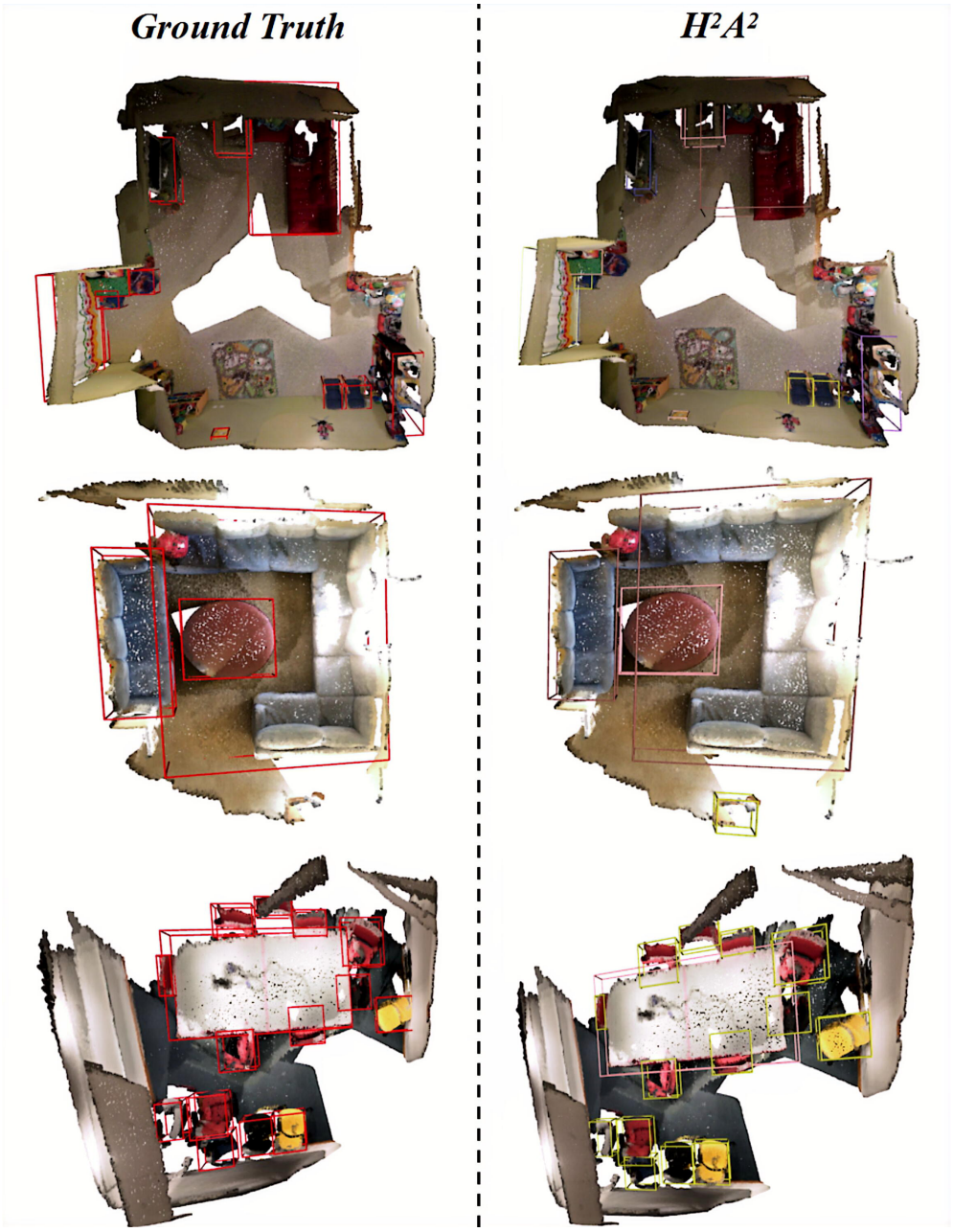

Figure 8. The point cloud from ScanNet v2 with bounding boxes. The color of each bounding box indicates the object category, including cabinet, chair, sofa, table, window, picture, counter and other furniture (only categories present in the figure are listed). Left: Ground truth annotations, Right: Predicted results using our proposed H^2A^2.

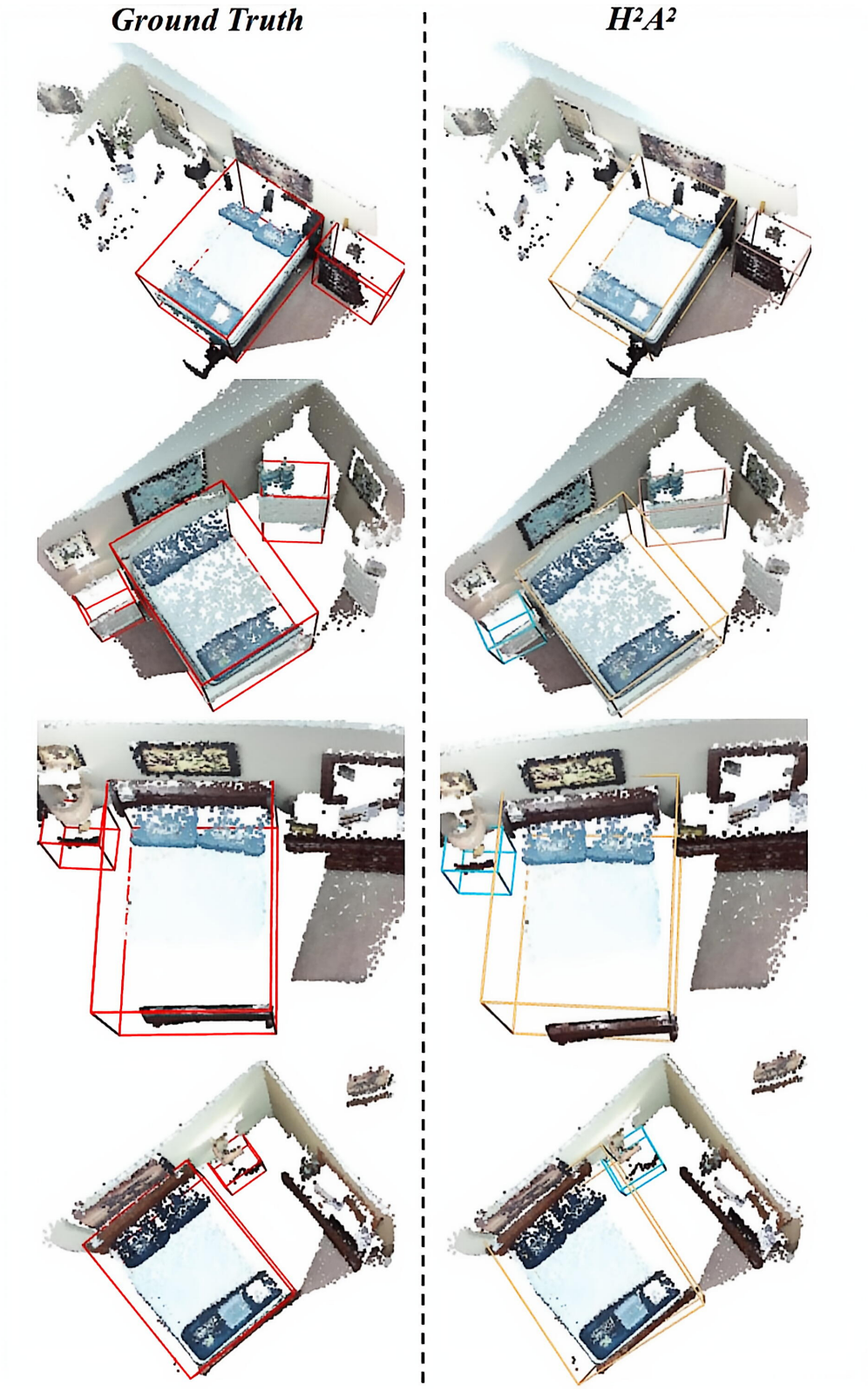

Figure 9. The point cloud from SUN RGB-D with bounding boxes. The color of each bounding box indicates the object category, including bed, dresser, table and other furniture (only categories present in the figure are listed). Left: Ground truth annotations, Right: Predicted results using our proposed H^2A^2.

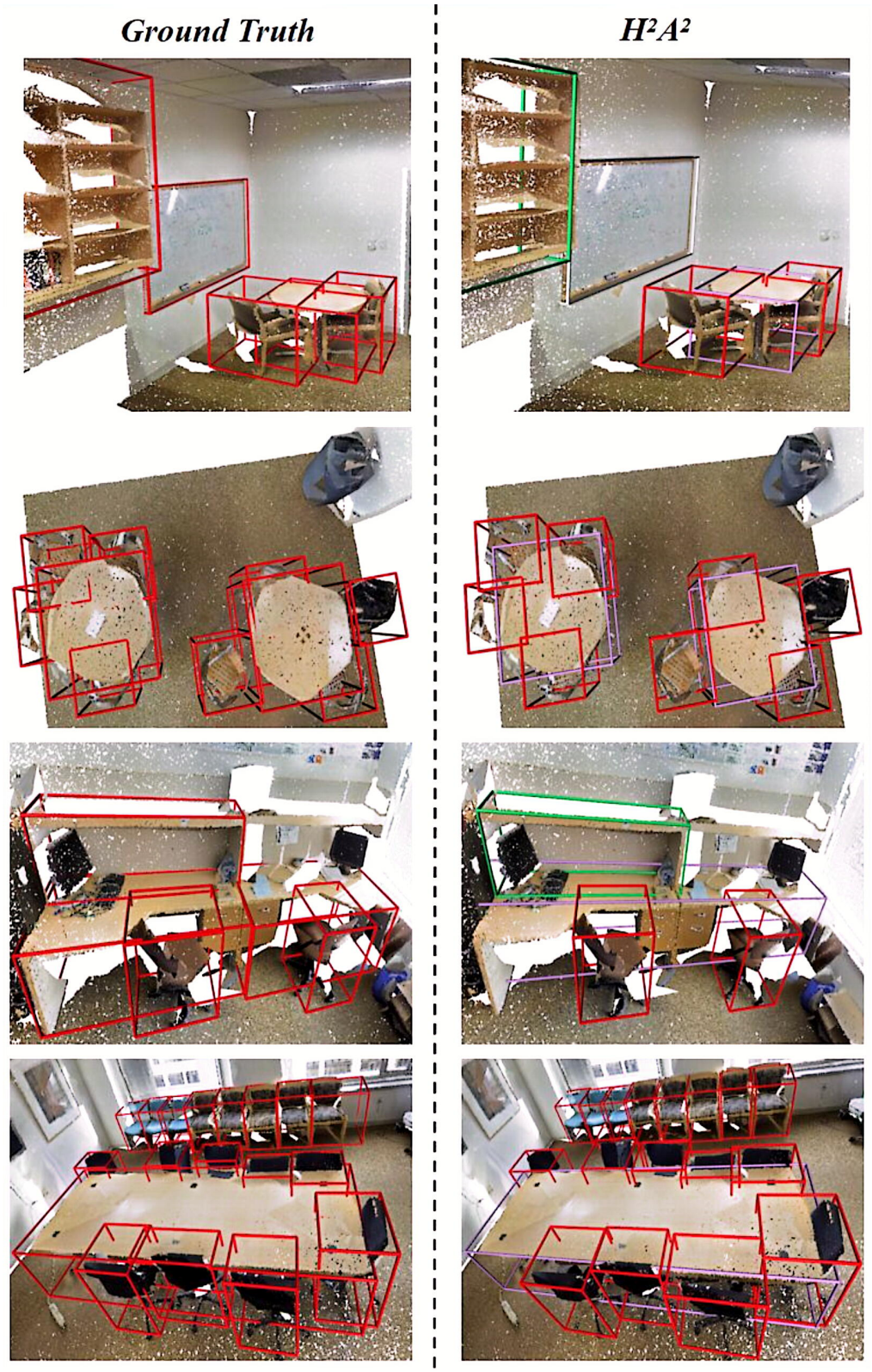

Figure 10. The point cloud from S3DIS with bounding boxes. The color of each bounding box indicates the object category, including table, chair, bookcase, whiteboard and other furniture (only categories present in the figure are listed). Left: Ground truth annotations, Right: Predicted results using our proposed H^2A^2.