TUE-AM-096

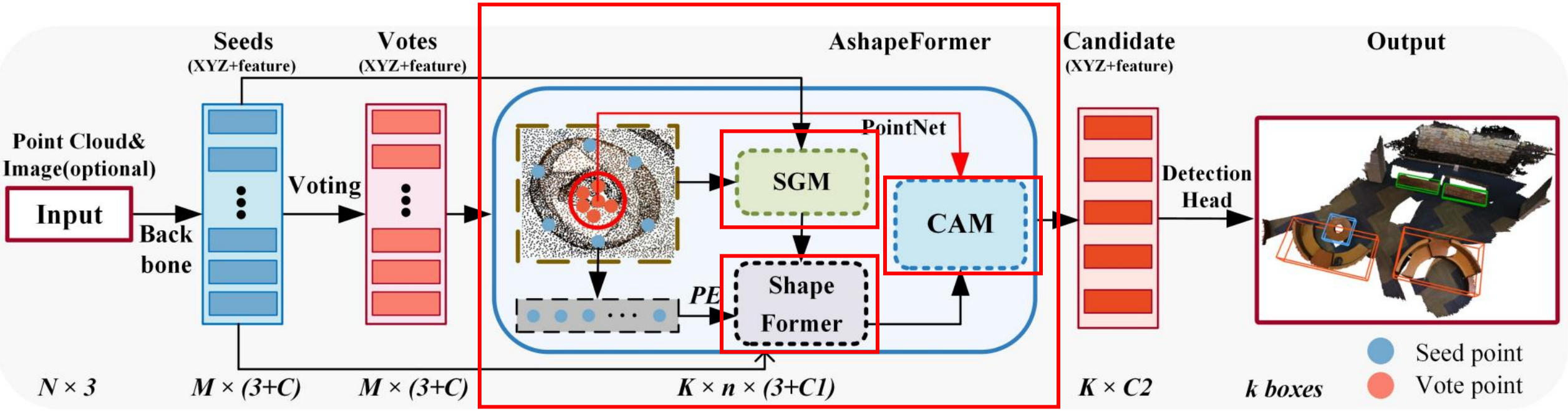
AShapeFormer : Semantics-Guided Object-Level Active Shape Encoding for 3D Object Detection via Transformers

Zechuan Li[1] , Hongshan Yu[1] , Zhengeng Yang[1], Tongjia Chen[1] , Naveed Akhtar[2]

[1]Hunan University [2]The University of Western Australia

AShapeFormer Overview

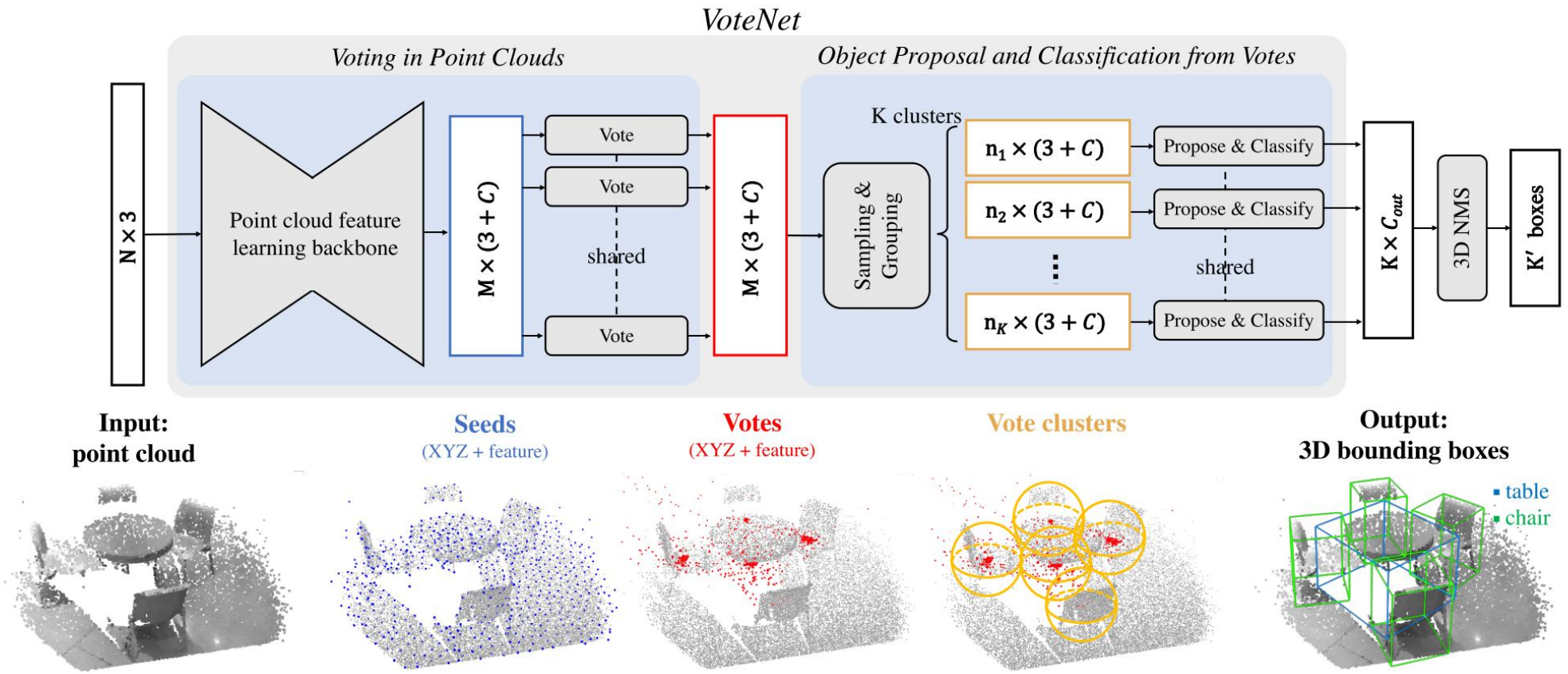
- We propose a plug-and-play active shape encoding module named AShapeFormer, which can be combined with many existing 3D object detection networks to achieve a considerable performance boost.
- To the best of our knowledge, our method is the first to combine multi-head attention and semantic guidance to encode strong object shape features for robust classification and accurate bounding box regression.
- AShapeFormer mainly includes three sub-modules, ShapeFormer ,SGM and CAM.



Introduction

Motivation:

- Most **voting-based** 3D object detection techniques commonly follow a pipeline that **aggregates** predicted object **central point** features to compute candidate points.

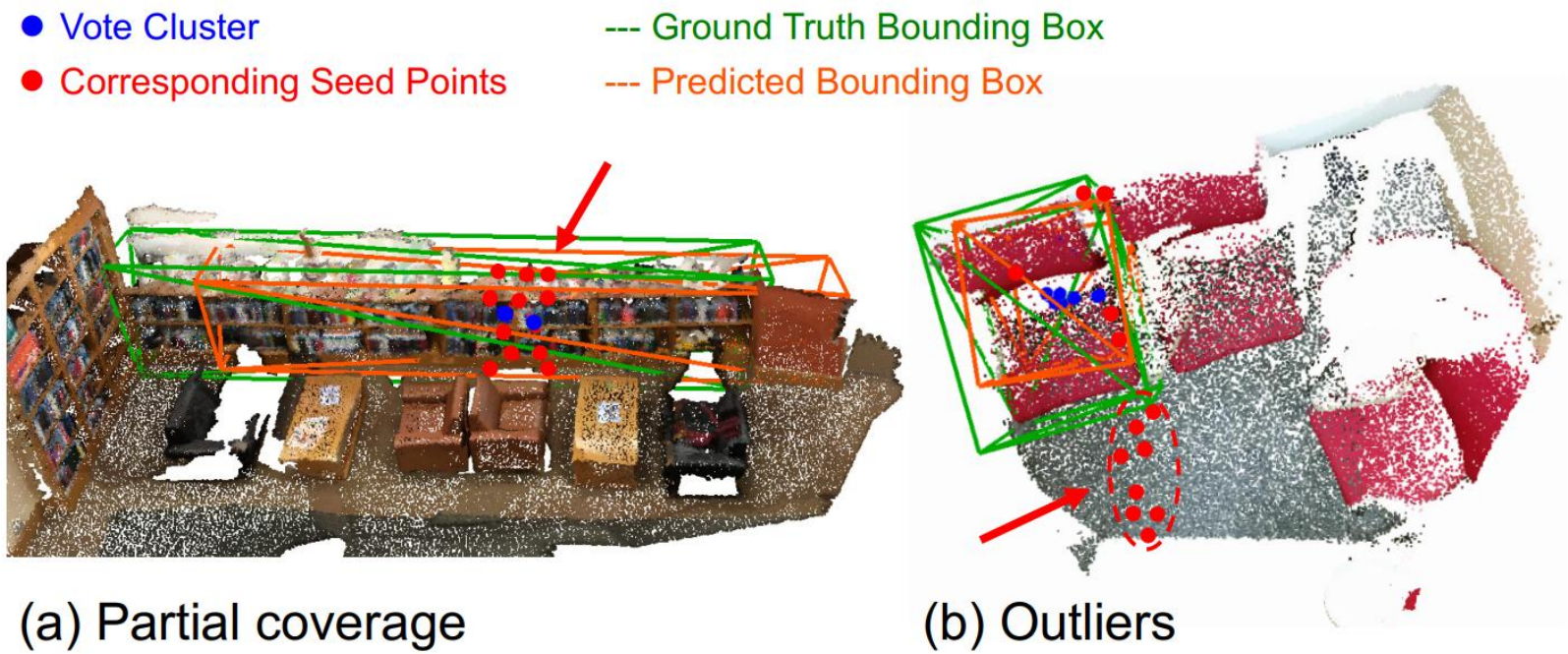


VoteNet: https://arxiv.org/pdf/1904.09664.pdf

Introduction

Motivation:

- The candidate points in the voting-based method only contain position information, largely ignoring the object-level shape information.
- Candidate points contain background points that may affect the performance of 3D object detection.

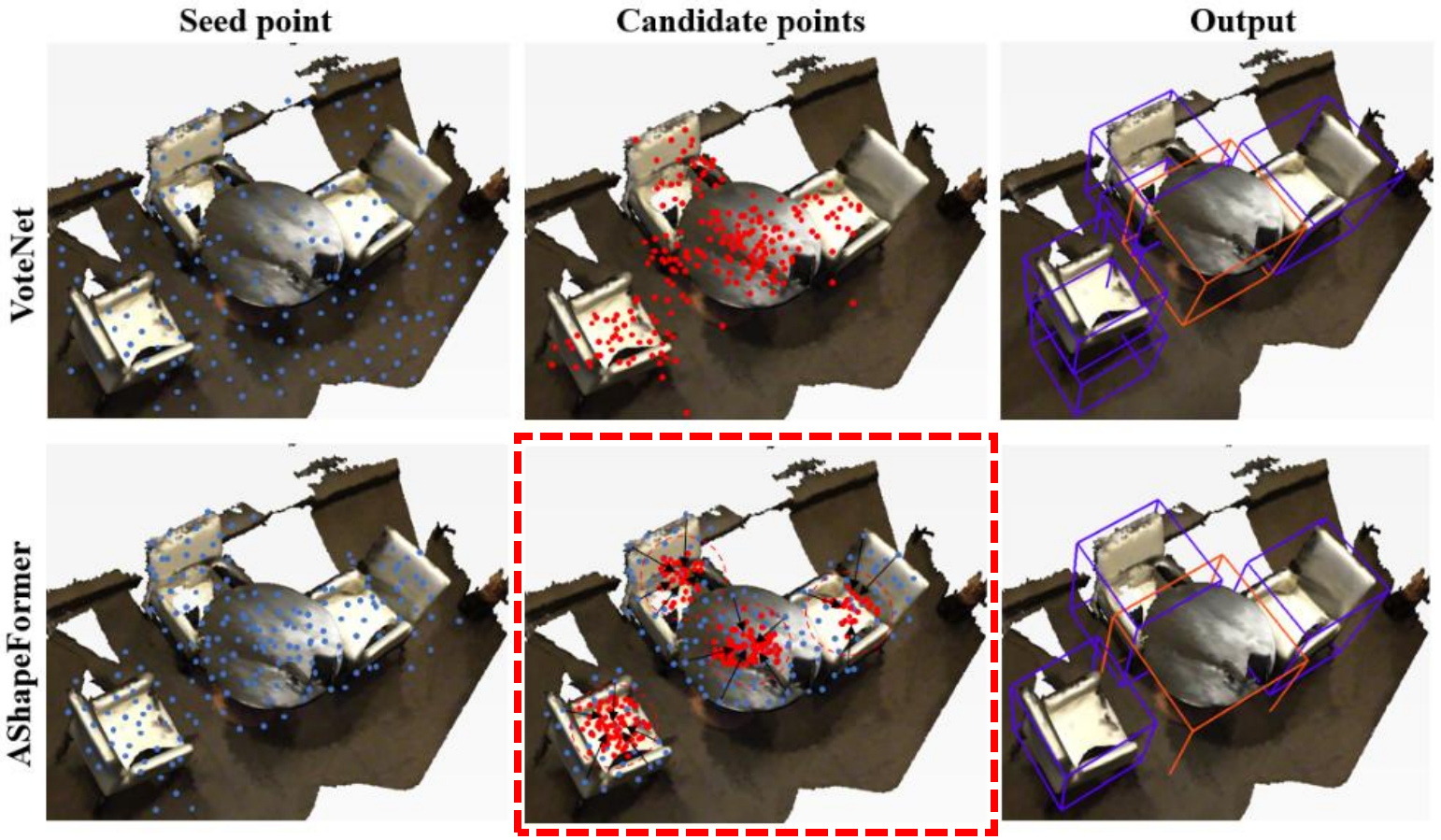


BRNet: https://arxiv.org/pdf/2104.06114.pdf

Introduction

Motivation:

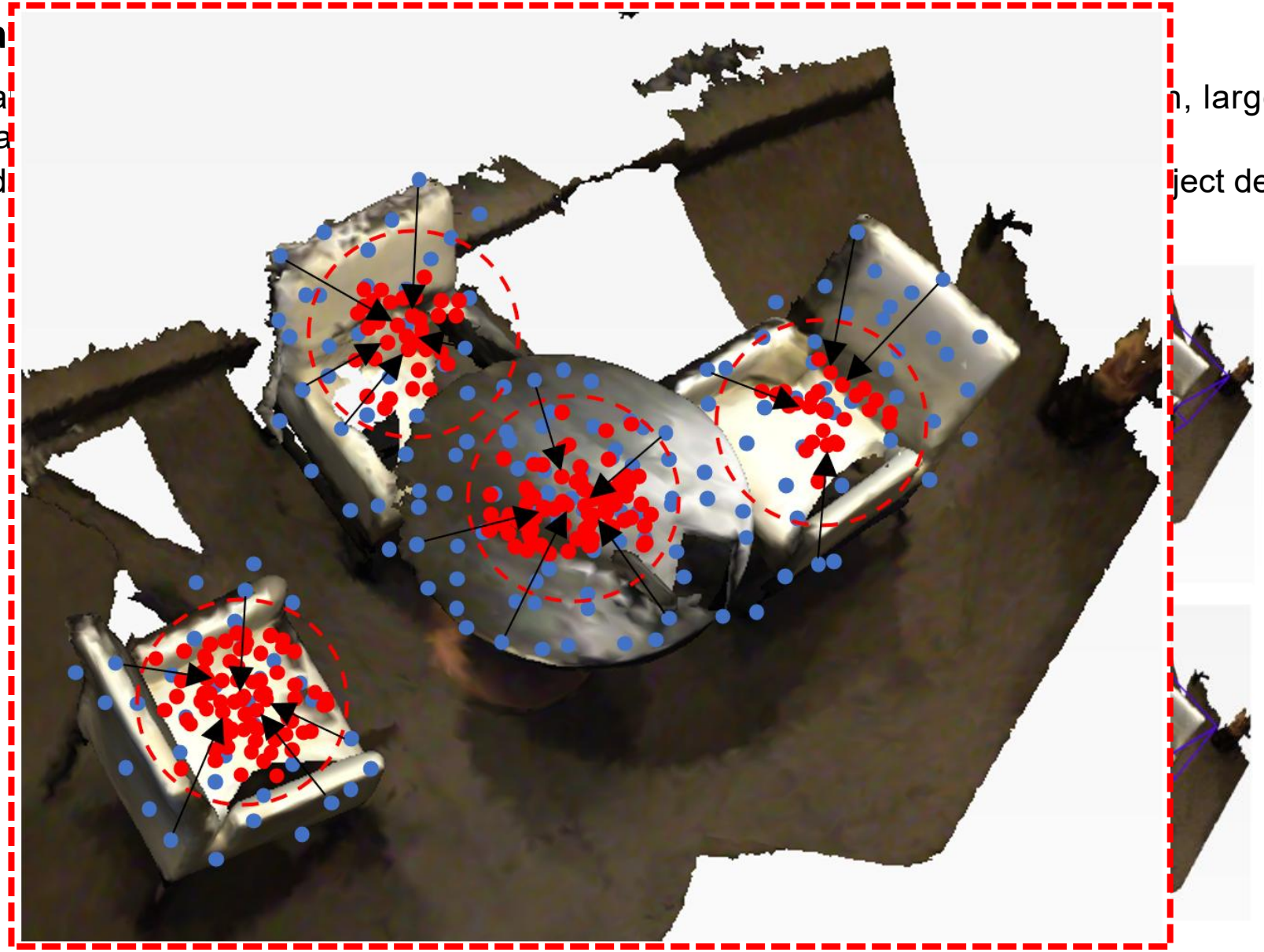
- The candidate points in the voting-based method only contain position information, largely ignoring the object-level shape information.
- Candidate points contain background points that may affect the performance of 3D object detection.



Introduction

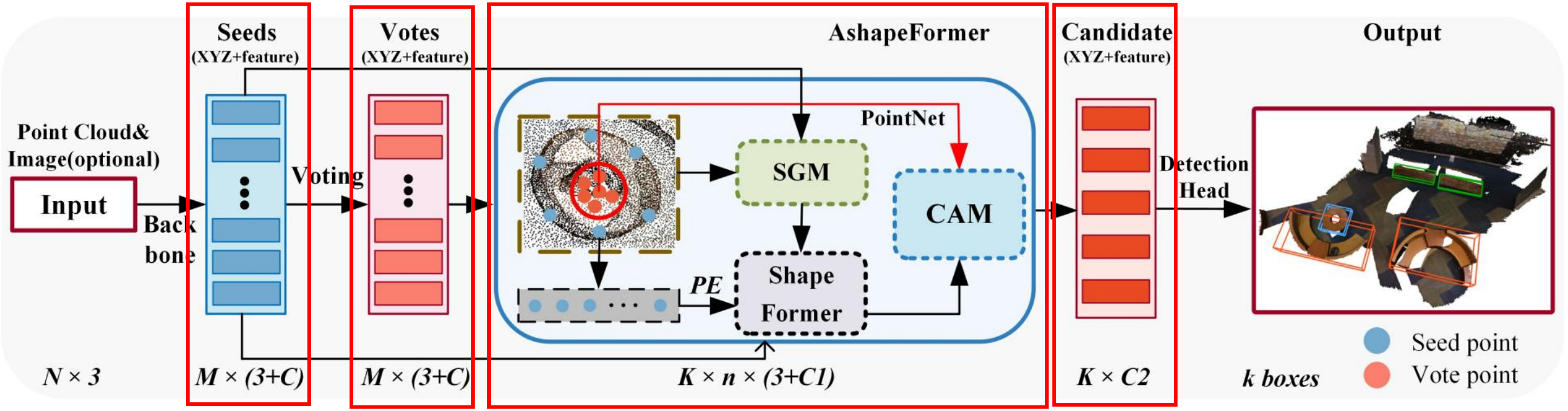
Motiva

- The ca…n, largely ignoring the object-level shape informa…
- Candid…ject detection.



Proposed Approach

AShapeFormer:



Innovations and contributions:

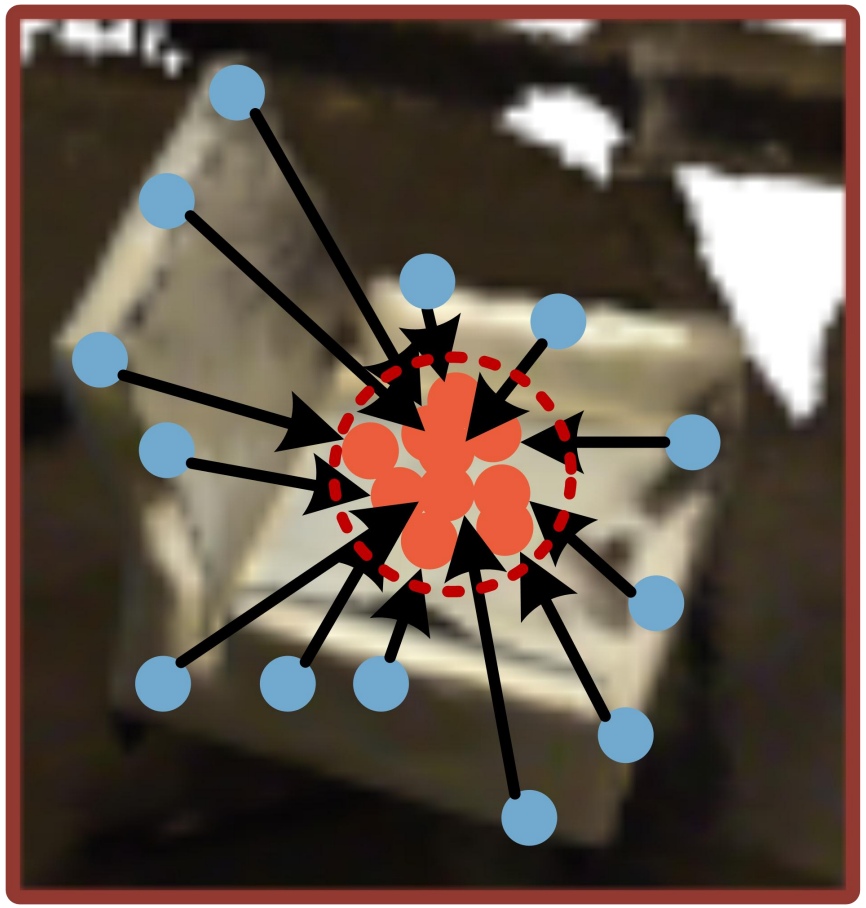
（1）ShapeFormer： Actively encode object-level shape features using the Transformers module.

（2）Semantics Guided Module (SGM)： Pay more attention to the foreground points.

（3）Channel Attention Module (CAM)： Fuse the shape information enriched features with candidate features.

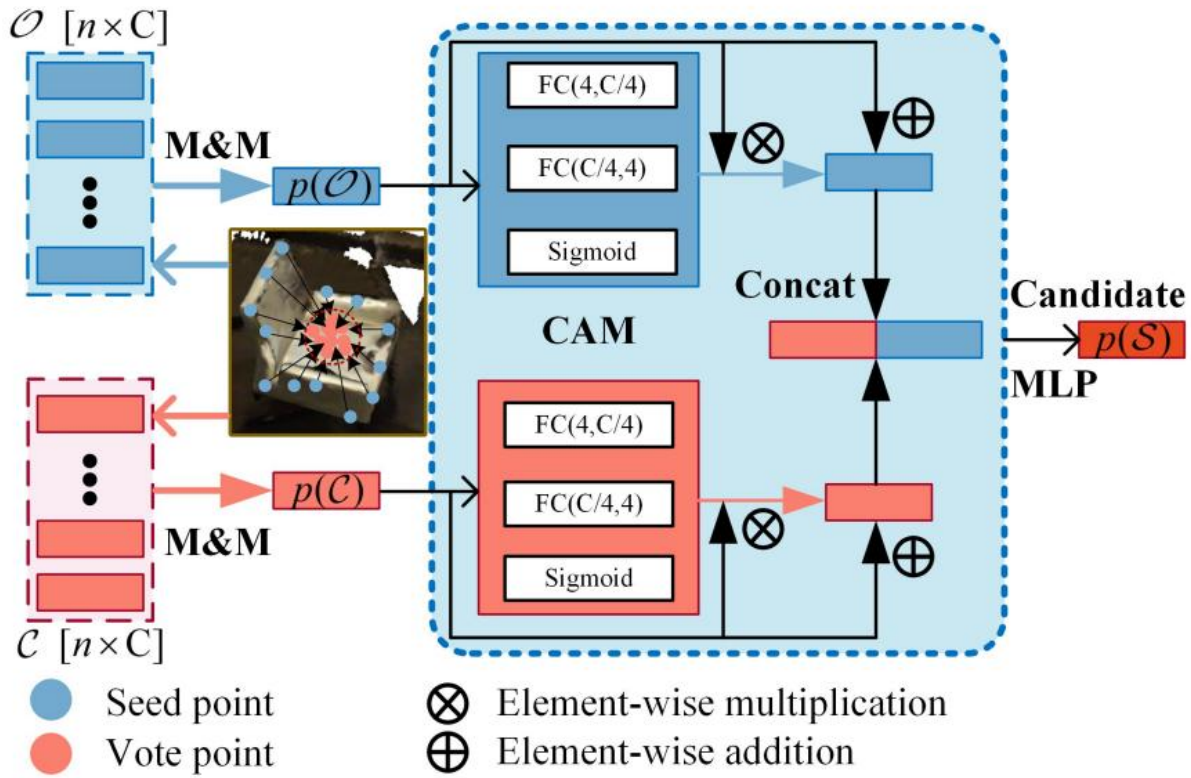
Proposed Approach

Naïve Object-Level Shape Encoding:



Problems:

- loss of fine-grained information
- marred by interference of the background points.

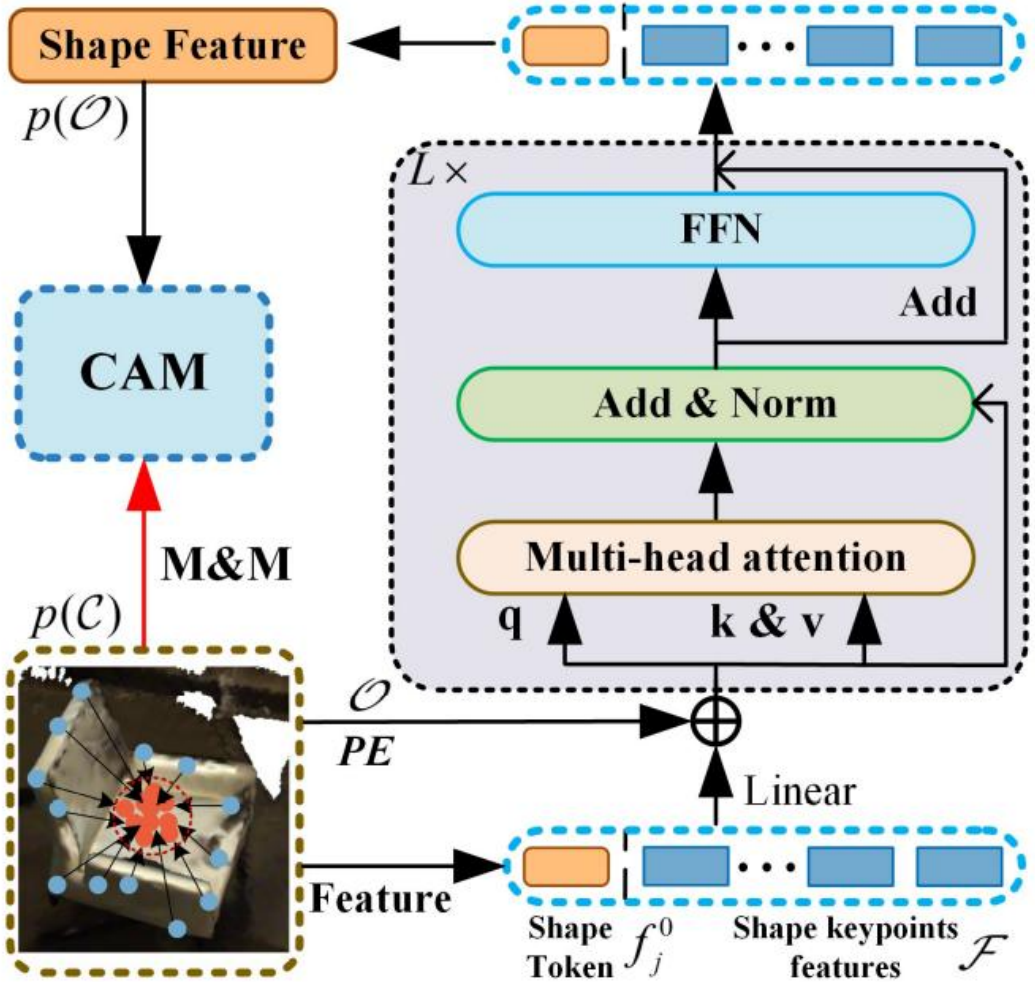


$$p(\mathcal{C}) = \text{MaxPool}\left(\text{MLP}\left(\mathcal{C}\right)\right) \quad p(\mathcal{O}) = \text{MaxPool}\left(\text{MLP}\left(\mathcal{O}\right)\right)$$

$$p(\mathcal{S}) = \text{Concat}\left(\text{CAM}(p(\mathcal{C}), \text{CAM}\left(p(\mathcal{O})\right)\right)$$

Proposed Approach

ShapeFormer:



ShapeFormer：

$$q = zW_q, k = zW_k, v = zW_v$$

$$y' = \sum \text{softmax}\left(\frac{qk^\intercal}{\sqrt{d}}\right) v$$

$$y = \text{Concat}\left(y'^{(0)}, y'^{(1)}, \ldots, y'^{(m-1)}\right)$$

$$\mathbf{o} = \mathcal{A}\left(\mathcal{F}\left(\mathcal{A}\left(y\right)\right)\right)$$

$$\mathbf{o}^l = \left[f_j^{0(l)}, f_j^{1(l)}, f_j^{2(l)}, \ldots, f_j^{n(l)}\right]$$

$$p(\mathcal{O}) = \text{MLP}\left(f_j^{0(l)}\right)$$

Object-Scene Positional Encoding：

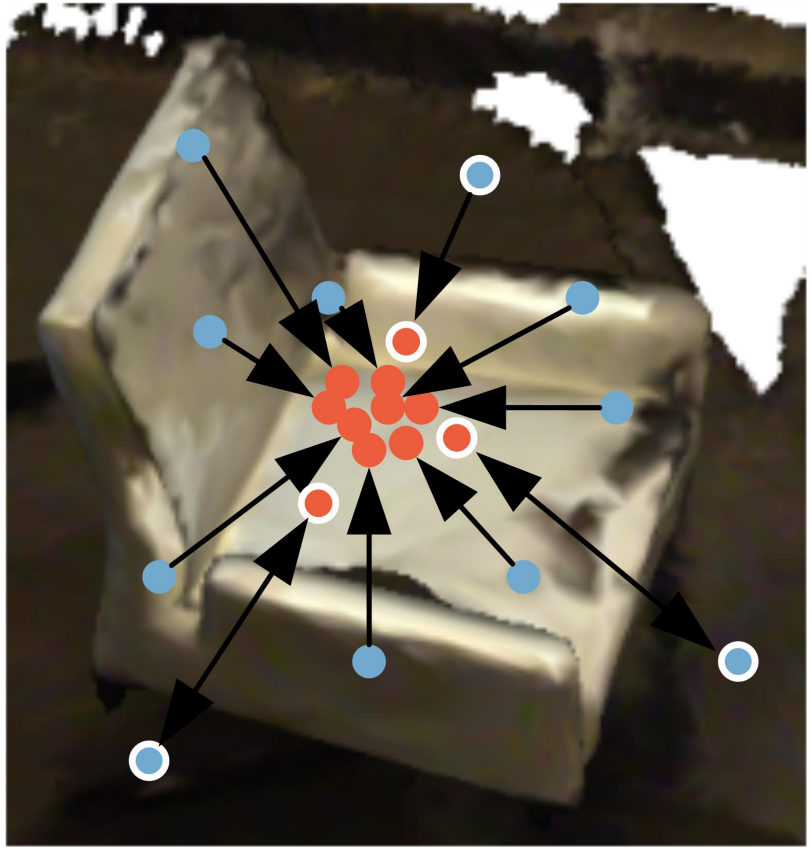
$$\text{PE} = \text{PE}_s + \text{PE}_o,$$

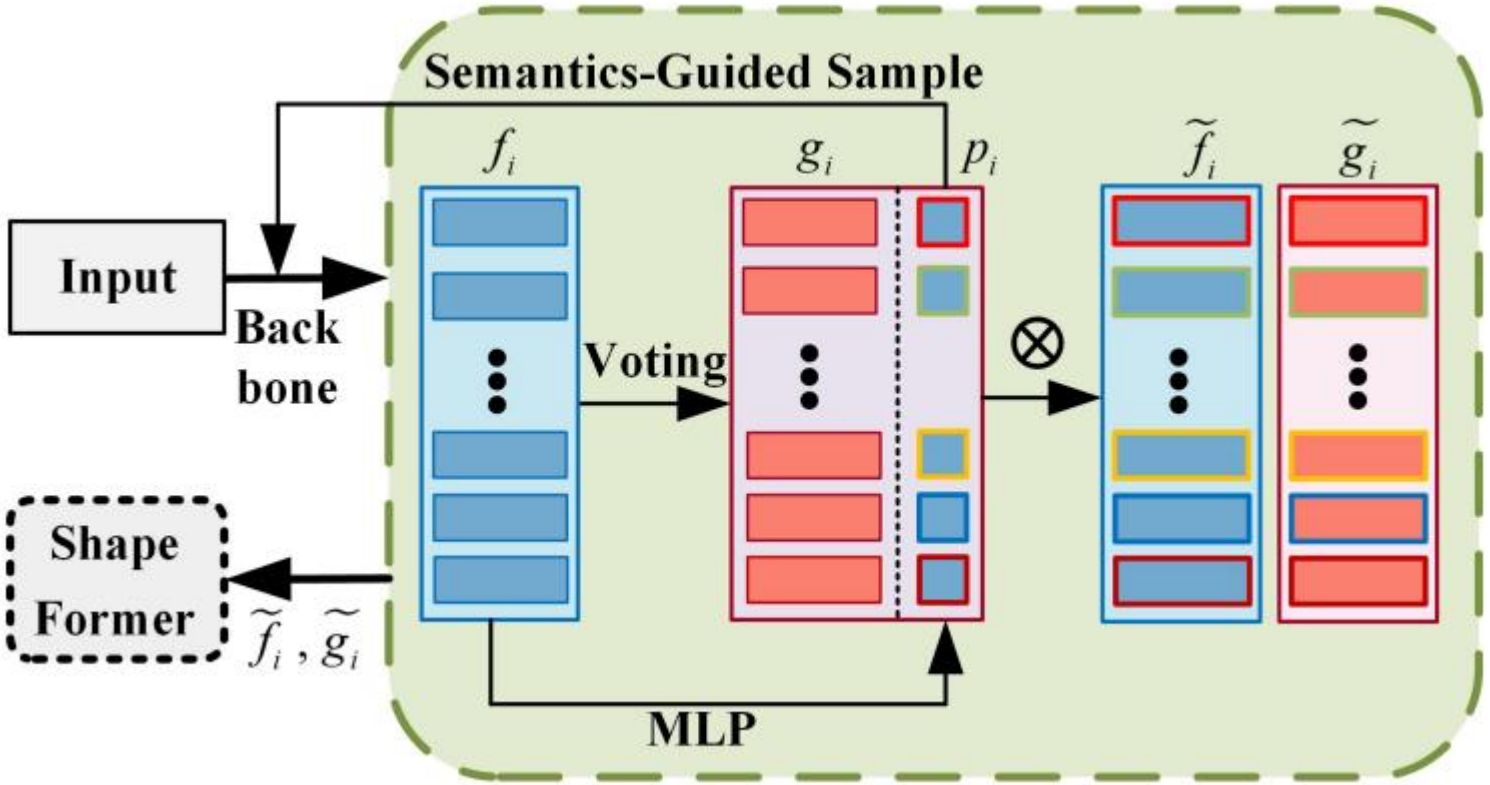
$$\text{PE}_s = \text{MLP}\left(\left[x_c; x_1, x_1 \ldots, x_n\right]\right),$$

$$\text{PE}_o = \text{MLP}\left(\left[\mathbf{0}; x_1 - x_c, x_1 - x_c \ldots, x_n - x_c\right]\right),$$

Proposed Approach

Semantics Guided Module：





Experiments

Quantitative comparison:

- We evaluate the performance of the proposed AShapeFormer on two popular datasets of indoor scenes, namely; ScanNet dataset [9] and SUN RGB-D dataset [38].

Model	mAP@0.25	bed	table	sofa	chair	toilet	desk	dresser	nightstand	shelf	bathtub
VoteNet [29]	57.7	83.0	47.3	64.0	75.3	90.1	22.0	29.8	62.2	28.8	74.4
VoteNet∗ [29]	59.7	84.8	49.6	67.8	77.6	87.4	24.3	29.3	61.9	32.1	82.1
BRNet [7]	61.1	86.9	51.8	66.4	77.4	91.3	29.6	35.9	65.9	29.7	76.2
Groupfree3D [23]	63.0	87.8	53.8	70.0	79.4	91.1	32.6	36.0	66.7	32.5	80.0
imVoteNet [28]	63.4	87.6	51.1	70.7	76.7	90.5	28.7	41.4	69.9	41.3	75.9
imVoteNet∗ [28]	64.5	88.5	51.6	73.2	79.2	90.2	30.9	38.0	67.3	46.4	79.7
RBGNet [42]	64.1	88.4	54.5	71.0	82.7	91.3	32.1	38.7	66.7	34.5	80.6
FCAF3D [34]	64.2	88.3	53.0	69.7	81.1	91.3	34.0	40.1	71.9	33.0	79.0
TokenFusion [44]	64.9	-	-	-	-	-	-	-	-	-	-
DisARM [11]	65.3	87.5	52.7	74.1	80.7	91.6	33.3	39.8	69.5	43.7	79.9
Ours (VoteNet)	**61.2(+3.5)**	86.9	51.5	67.8	78.8	91.2	29.0	33.6	65.0	31.3	76.6
Ours (VoteNet∗)	**62.2(+2.5)**	86.9	51.3	69.3	78.9	90.2	28.2	34.6	65.9	35.6	80.7
Ours (imVoteNet)	**65.1(+1.7)**	89.2	53.7	72.9	78.3	90.8	30.2	43.2	70.0	46.5	76.1
Ours (imVoteNet∗)	**65.8(+1.3)**	87.6	55.2	72.8	80.9	92.5	31.2	45.8	67.7	43.7	80.9

We achieve an absolute gain of more than 3.5% and 1.7% for VoteNet and imVoteNet, respectively.

Experiments

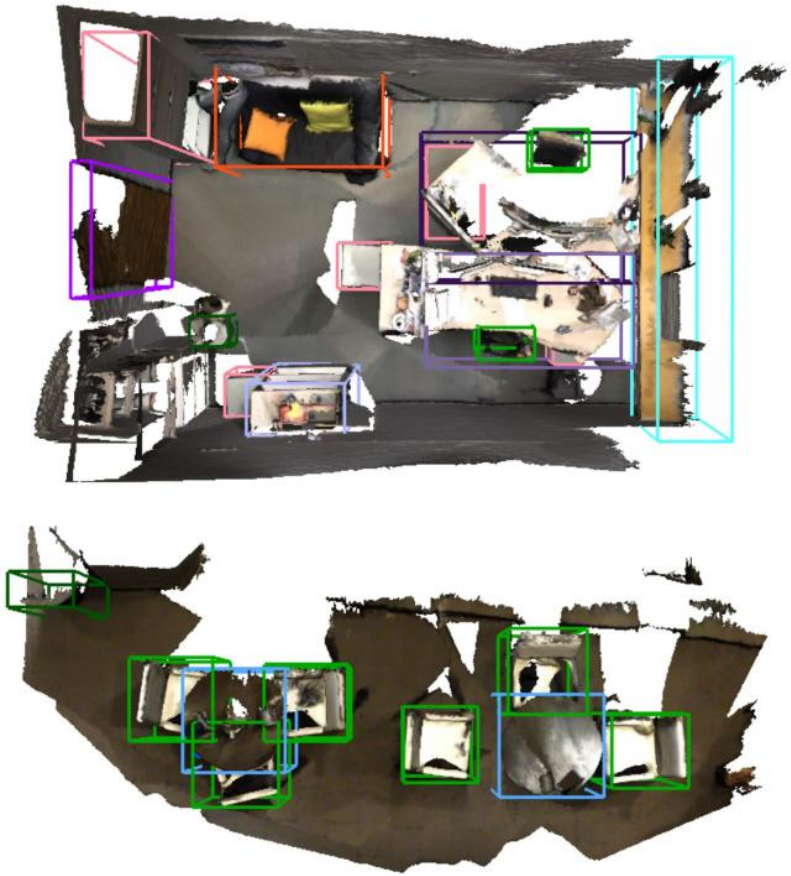
Quantitative comparison:

Model	mAP	cab	bed	chair	sofa	table	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn
3D-SIS [7]	22.5	5.7	50.2	52.5	55.4	21.9	10.8	0.0	13.1	0.0	0.0	23.6	2.6	24.5	0.8	71.7	8.9	56.4	6.8
HGNet [1]	34.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
VoteNet [17]	33.5	8.0	76.0	67.2	68.8	42.3	15.3	6.4	28.0	1.25	9.5	37.5	11.5	27.8	9.9	86.5	16.7	78.8	11.6
VoteNet∗ [17]	44.2	23.1	77.7	76.7	70.6	46.9	30.4	15.7	45.7	4.6	27.4	49.8	30.0	36.9	20.9	90.7	32.5	83.3	28.4
MLCVNet [20]	42.1	16.6	83.3	78.1	74.7	55.1	28.1	17.0	51.7	3.7	13.9	47.7	28.6	36.3	13.4	70.9	25.6	85.7	27.5
3DETR [11]	47.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GroupFree3D [9]	52.8	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
RBGNet [19]	54.2	30.6	80.9	86.5	84.8	66.4	40.3	29.5	48.6	7.9	44.7	59.1	40.8	44.8	39.7	92.9	45.3	90.9	41.5
Ours(VoteNet)	**41.6(+8.1)**	20.1	80.8	76.1	70.2	53.6	31.5	14.7	30.4	5.4	25.5	33.1	27.3	35.5	13.6	89.1	30.5	89.7	22.2
Ours(VoteNet∗)	**47.8(+3.6)**	25.3	81.1	80.8	71.3	56.8	33.3	20.9	53.8	6.1	33.8	56.9	29.2	38.0	33.4	86.5	40.1	80.3	30.9
Ours(GroupFree3D)	**53.4(+0.6)**	30.3	82.5	82.5	74.2	64.6	39.8	26.7	56.2	6.8	34.2	69.9	47.3	41.6	43.2	89.9	39.9	91.1	40.8
Ours(RBGNet)	**56.6(+1.4)**	31.0	82.2	86.9	87.7	67.6	43.1	35.4	57.2	13.2	34.8	59.8	37.2	52.8	50.0	97.6	43.7	91.0	46.4

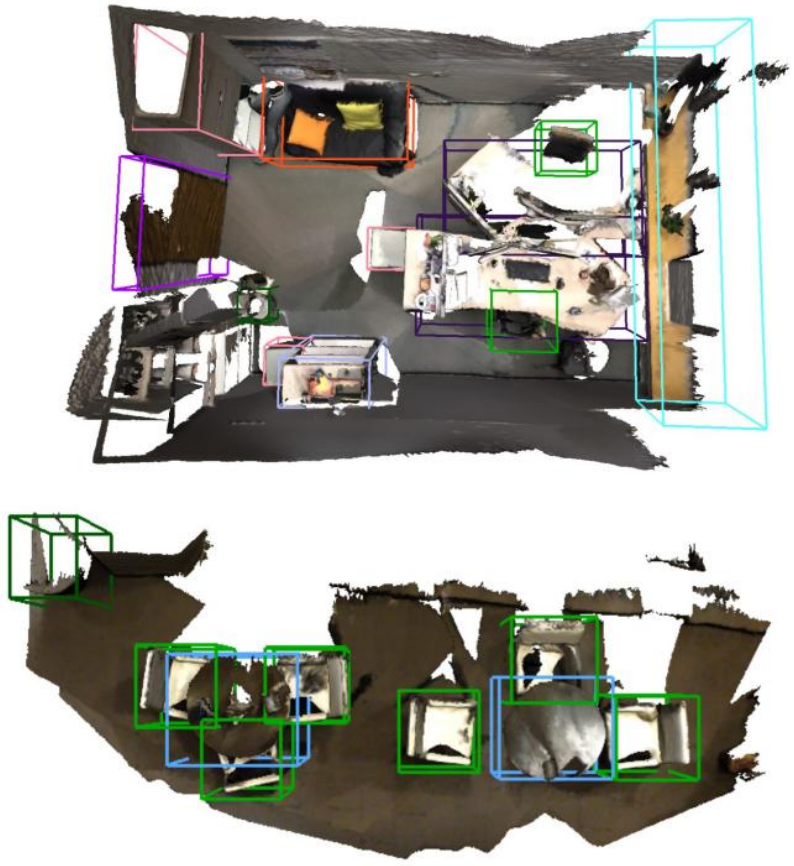
Obtained 8.1% performance improvement on the ScanNet V2 dataset

Experiments

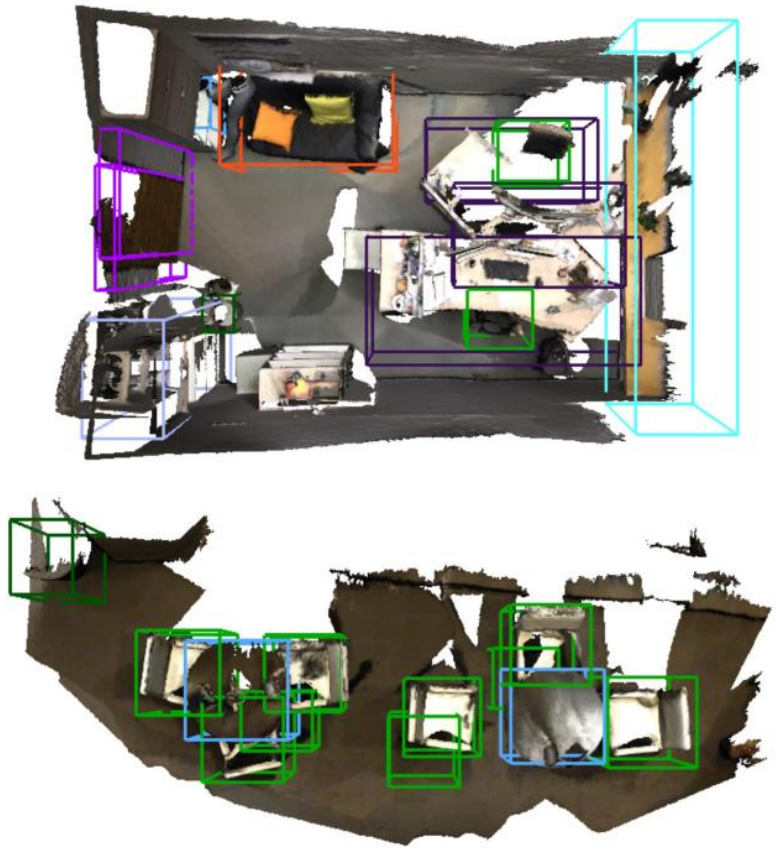
Qualitative comparison：



GT



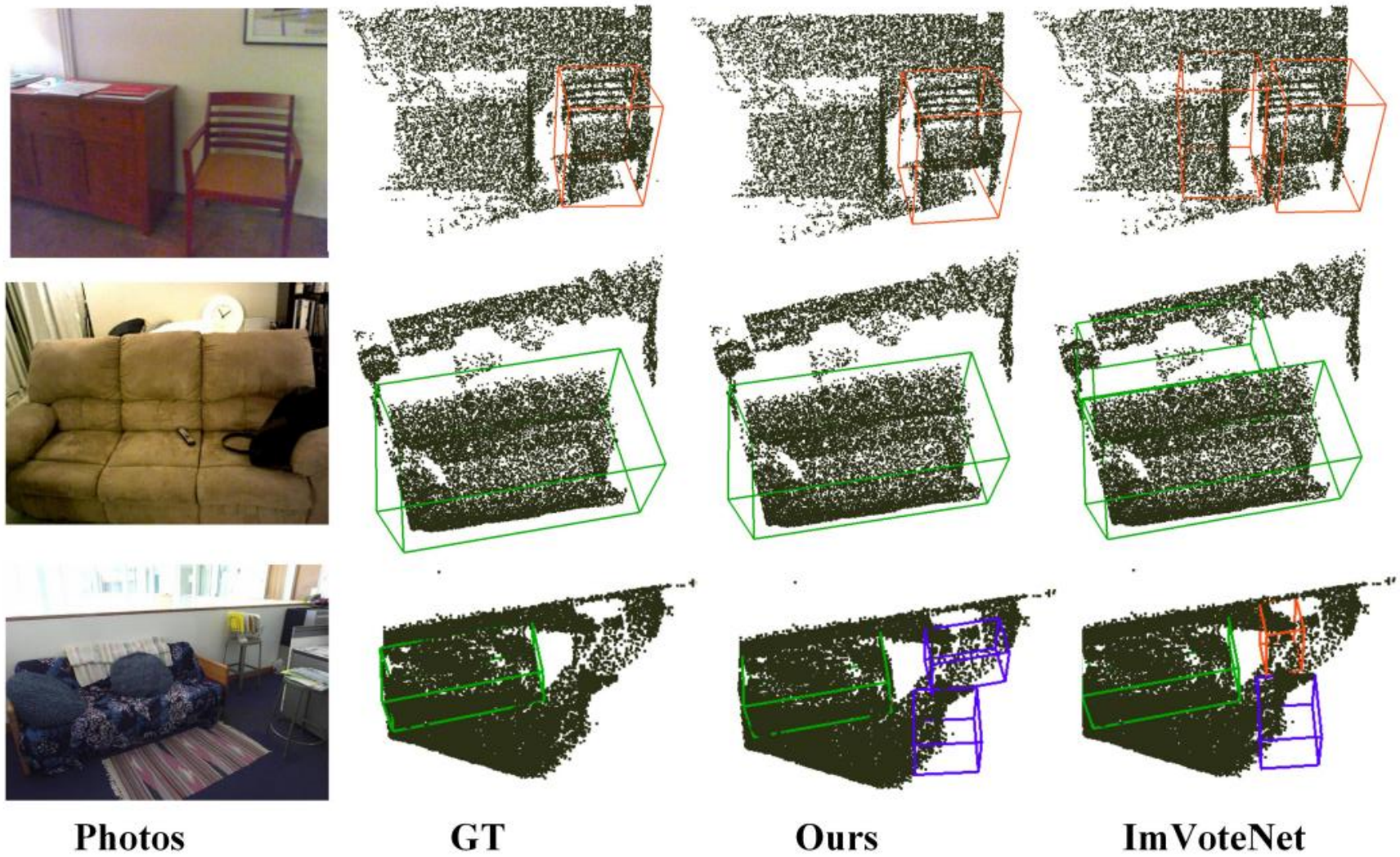
Ours



VoteNet

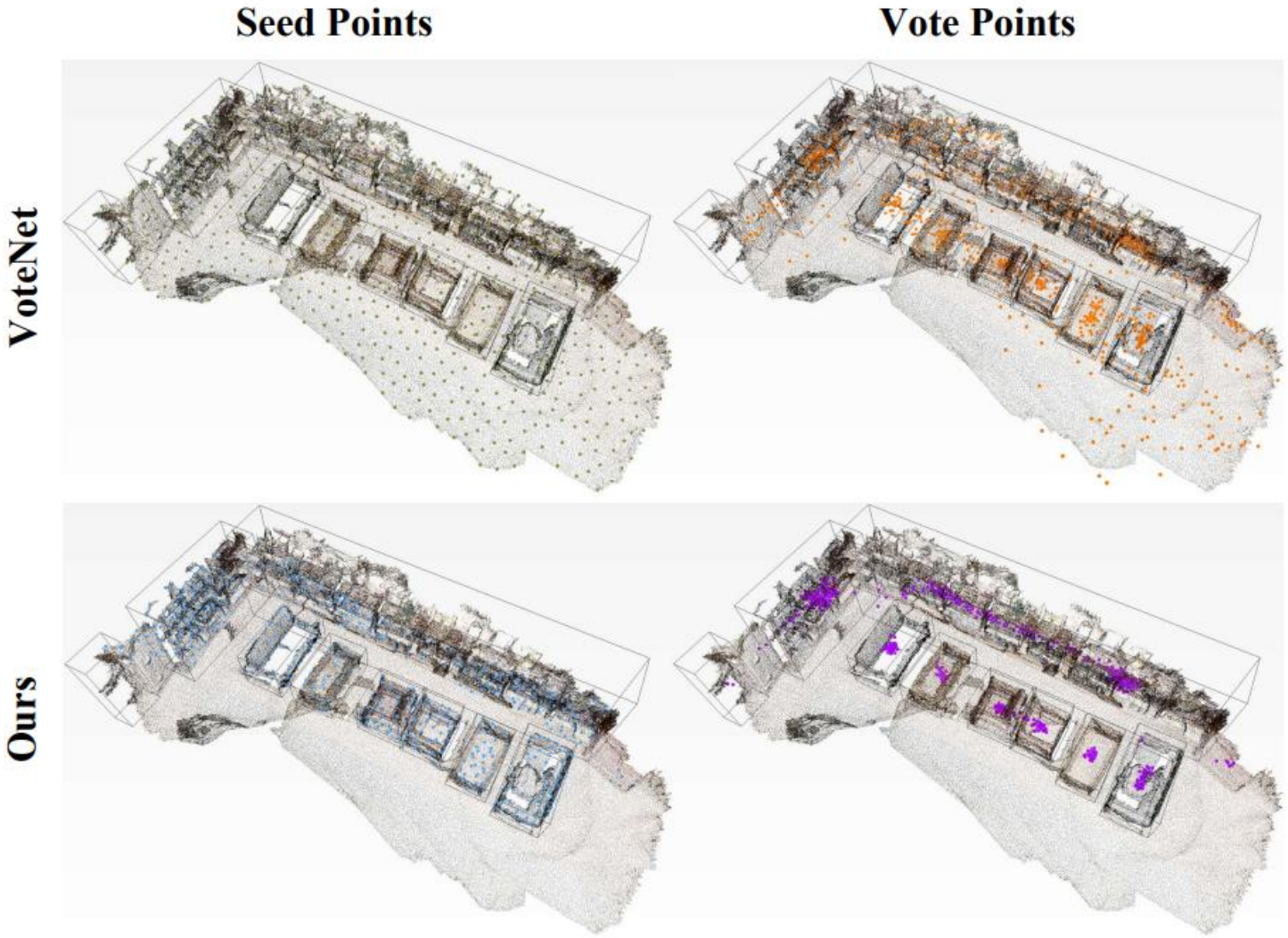
Experiments

Qualitative comparison：



Experiments

Qualitative comparison：



Conclusion

- We proposes a plug-n-play module to improve the performance of indoor 3D object detection by actively encoding shape information of the object.
- We sample the shape key points of the object and re-weight their features by guiding them with semantic information.
- we utilize multi-head attention to encode object shape features to avoid the loss of fine-grained information.
- Results show that our model achieves state-of-the-art performance when assembled with existing methods.

TUE-AM-096

Thank you

谢谢