CN-RMA: Combined Network with Ray Marching Aggregation for 3D Indoor Object Detection from Multi-view Images

CVPR2024

Guanlin Shen, Jingwei Huang, Zhihua Hu, Bin Wang

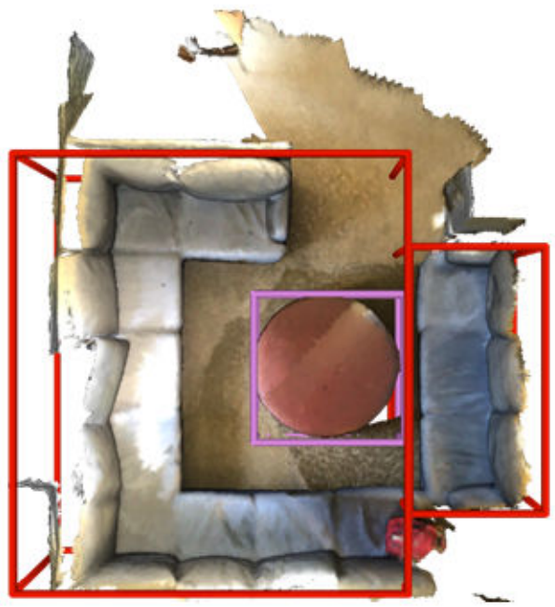
Introduction

Input: Multi-view images of 3D indoor scenes with camera parameters.

Multi-view Images
$\{\boldsymbol{I_i}\}(h_I \times w_I \times 3)$

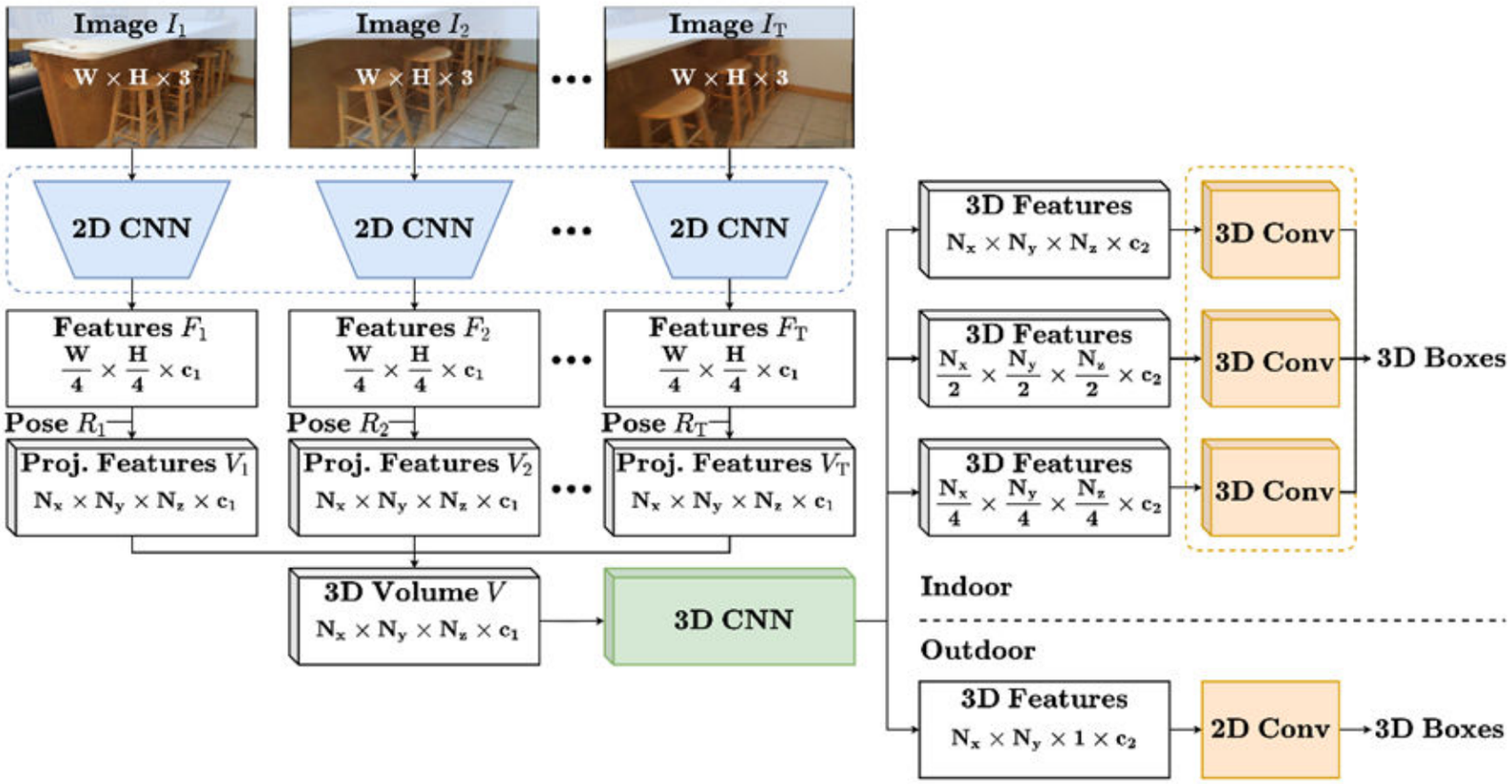
Camera Parameters
$\{\boldsymbol{K_i}, \boldsymbol{R_i}\}(4 \times 4)$

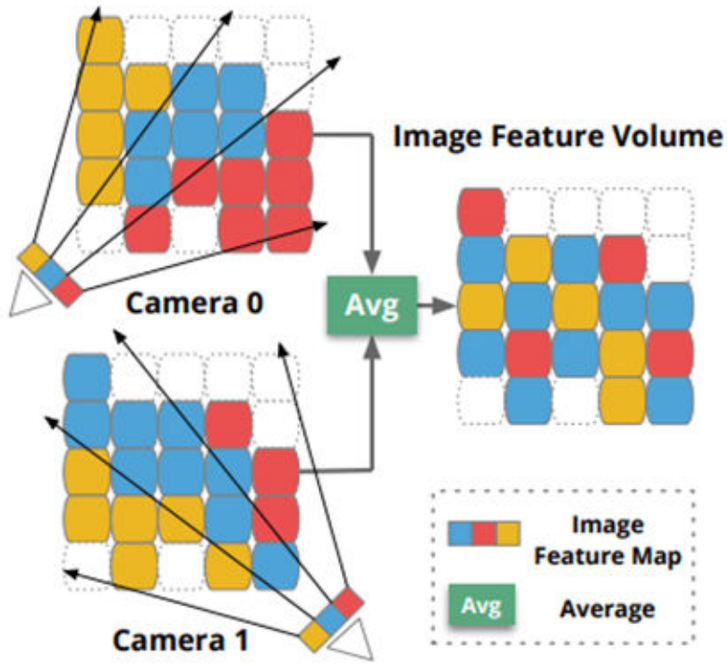
Output: 3D bounding boxes (AABB or OBB) with class labels.



Bounding Boxes $\{b_j\}$
and Class Scores $\{s_j\}$

Previous Work

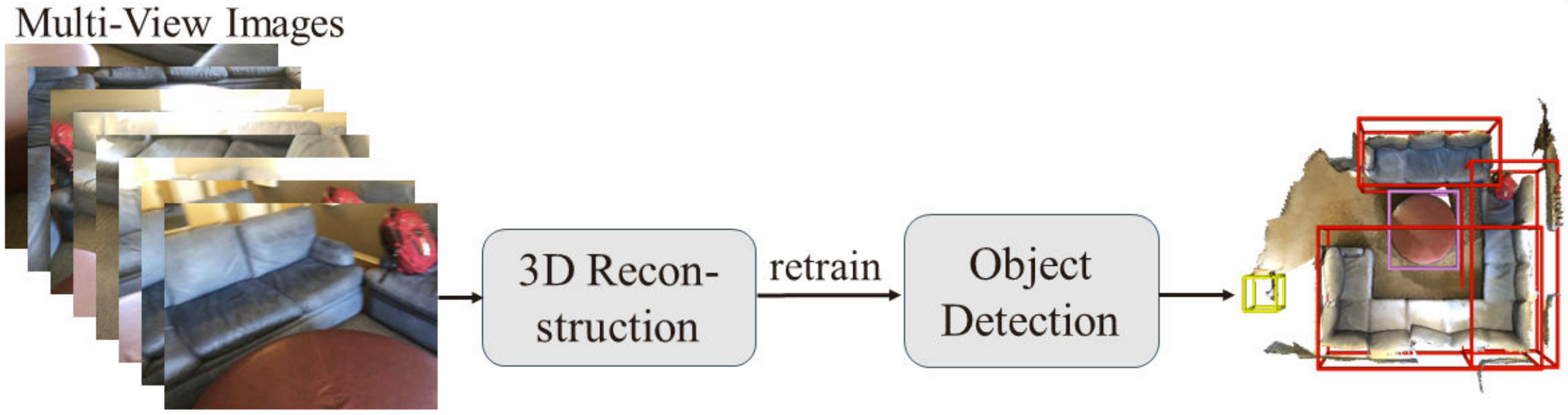




- **1. ImVoxelNet(WACV2022)[1]**
- ImVoxelNet aggregates 2D features extracted from multi-view images into 3D voxel volumes through unprojection.
- Unprojection does not fully exploit explicit geometric information and struggles to effectively address complex occlusion issue.

[1] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2397–2406, 2022.

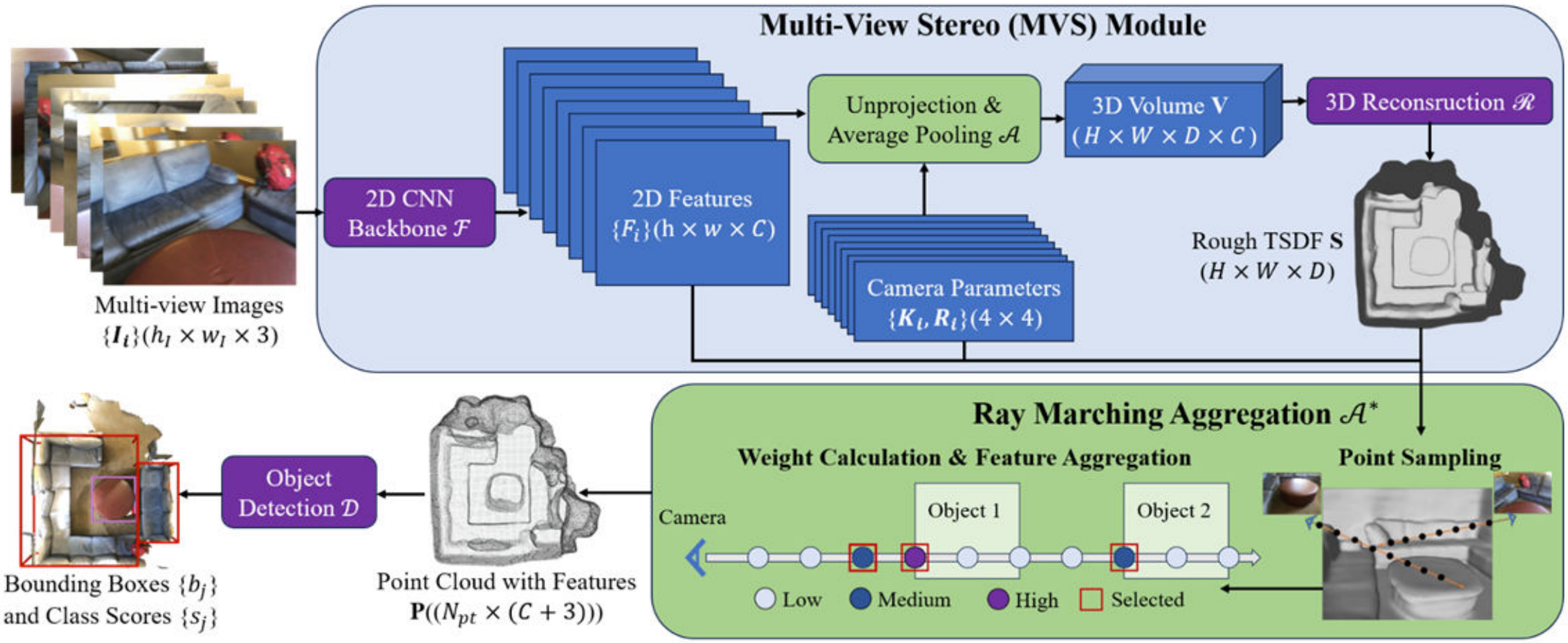
Previous Work



Method	mAP@0.25↑	mAP@0.5↑
ImVoxelNet [25]	46.7	23.4
NeRF-Det [36]	53.5	27.4
ImGeoNet [28]	54.8	28.4
Atlas [19]+FCAF3D [24]	55.4	33.8
NeuralRecon [26]+FCAF3D	51.5	31.6

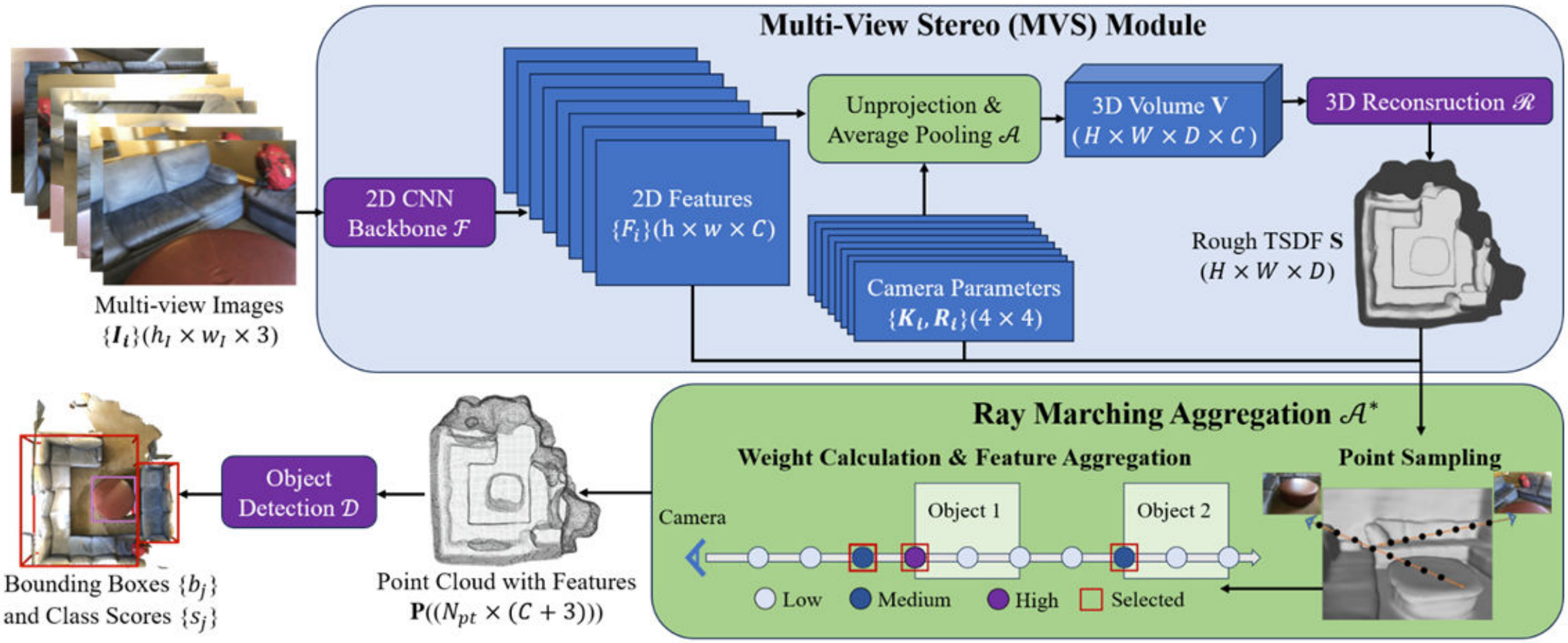
- **2.Two-Stage Methods**
- One solution is to perform 3D scene reconstruction from multi-view images, followed by 3D object detection from reconstructed point clouds.
- Surpasses previous state-of-the-art methods.
- Not ideal due to the lack of connectivity between the two stages, cannot fully exploit rich image features in the reconstruction stage.

Method Overview



- 1. We connect the reconstruction and detection networks seamlessly, enabling better exploitation of image features in 3D space.
- 2. We propose an innovative occlusion-aware aggregation method, RMA, which uses the reconstructed Truncated Signed Distance Function (TSDF) to address the occlusion issues.
- 3. We adopt a pre-training and fine-tuning scheme, and achieve the SOTA performance for indoor 3D object detection from multi-view images.

Method Overview



- 1.Extract image features and predict rough scene TSDF through our Multi-View Stereo (MVS) module.
- 2.Aggregate the image features in 3D point cloud with features through our Ray Marching Aggregation (RMA) method.
- 3.Detect the final bounding boxes and class scores through the 3D object detection network.

Ray Marching Aggregation (RMA)

- 1. For each pixel on each feature map, emit a ray $\boldsymbol{p}(t) = o + t \cdot d$ and unitedly sample points along each ray.
- 2. Following NeuS[2], obtain the opacity $\alpha\big(\boldsymbol{p}(t_i)\big)$, accumulated transmittance $T\big(\boldsymbol{p}(t_i)\big)$, and finally the weight $\mathrm{W}\big(\boldsymbol{p}(t_i)\big)$ of each point.

$$W(\mathbf{p}(t_i)) = T(\mathbf{p}(t_i)) \cdot \alpha(\mathbf{p}(t_i)),$$

$$T(\mathbf{p}(t_i)) = \prod_{i=0}^{i-1} (1 - \alpha(\mathbf{p}(t_j)))$$

$$\alpha(\mathbf{p}(t_i)) = \max\left(\frac{\Phi(\mathbf{S}(\mathbf{p}(t_i))) - \Phi(\mathbf{S}(\mathbf{p}(t_{i+1})))}{\Phi(\mathbf{S}(\mathbf{p}(t_i)))}, 0\right).$$

- Where $\mathbf{S}\big(\boldsymbol{p}(t_i)\big)$ denotes the TSDF value of sample point $\big(\boldsymbol{p}(t_i)\big)$, and $\Phi(\mathbf{x})$ denotes the sigmoid function.
- 3. Select sample points with the threshold θ_{rma}, concat the coordinates with features to form the 3D point cloud with features.

[2] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689, 2021.

Ray Marching Aggregation (RMA)

- The opacity $\alpha(\boldsymbol{p}(t_i))$ carries scene geometry information.

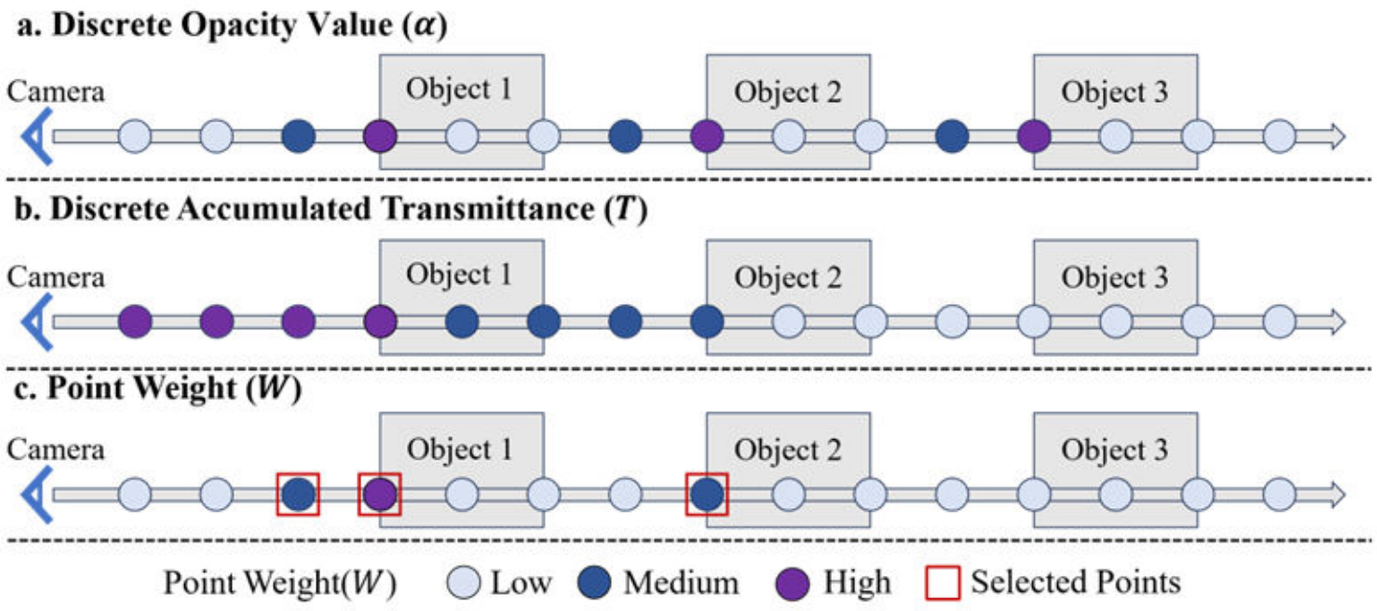
$$\alpha(\mathbf{p}(t_i)) = \max\left(\frac{\Phi(\mathbf{S}(\mathbf{p}(t_i))) - \Phi(\mathbf{S}(\mathbf{p}(t_{i+1})))}{\Phi(\mathbf{S}(\mathbf{p}(t_i)))}, 0\right).$$

- The accumulated transimittance $T(\boldsymbol{p}(t_i))$ considers occlusion.

$$T(\mathbf{p}(t_i)) = \prod_{j=0}^{i-1}(1 - \alpha(\mathbf{p}(t_j)))$$

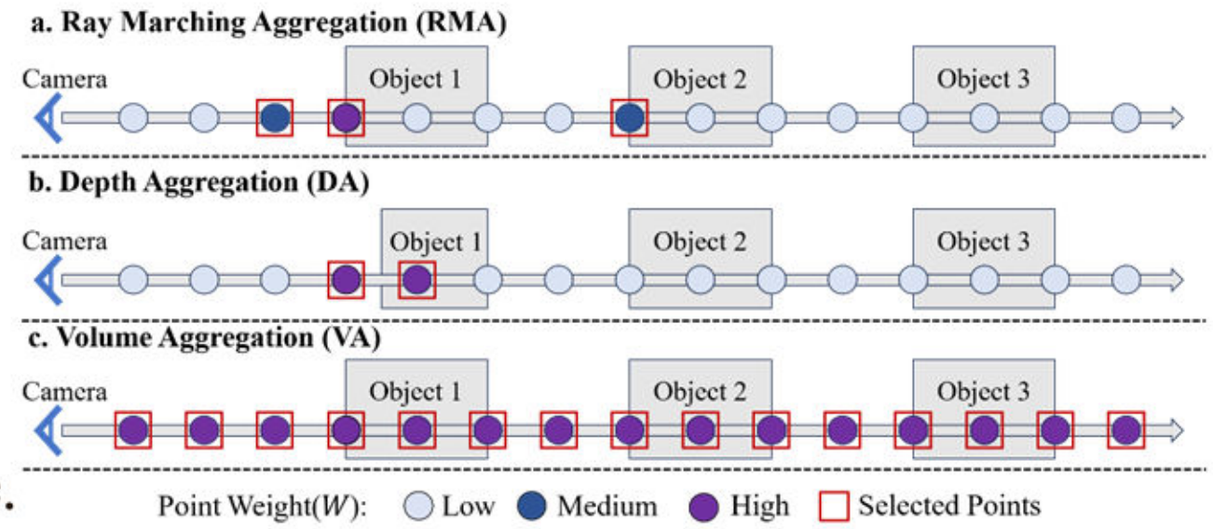
- The weight $\mathrm{W}(\boldsymbol{p}(t_i))$ combines them together

$$W(\mathbf{p}(t_i)) = T(\mathbf{p}(t_i)) \cdot \alpha(\mathbf{p}(t_i)).$$



Ray Marching Aggregation (RMA)

- **Compare with other methods:**
- 1.Compared to the volume-based aggregation method (VA) in ImVoxelNet[1], our RMA method fully utilizes the reconstructed scene geometry and considers the complex occlusion issue.

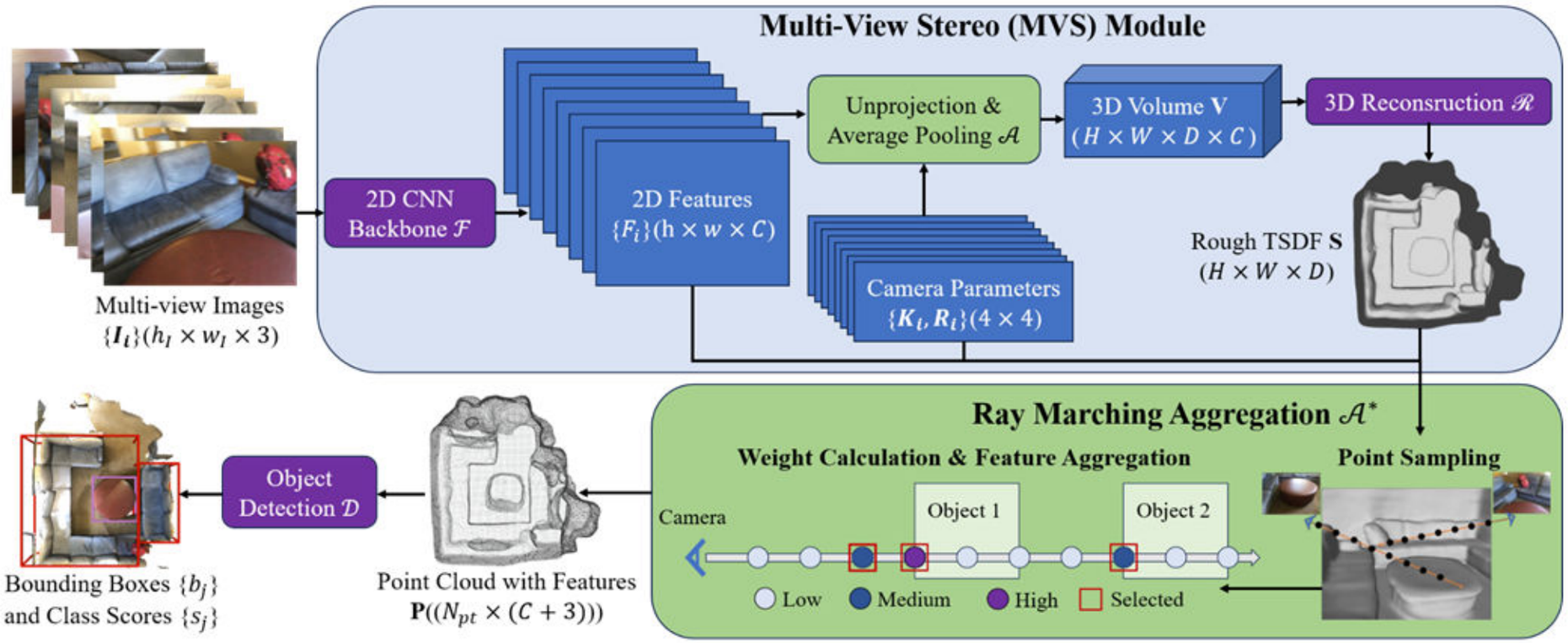


- 2.Compared to the depth-based aggregation method (DA), which directly lifts 2D features to point clouds through depth maps of each view obtained from the reconstruction results, our RMA method is more robust with inaccurate reconstructed scene geometry.

Method	parameter	mAP@0.25↑	mAP@0.5↑
VA	–	31.1	11.3
RMA	$\theta_{rma} = 0.02$	**58.6**	**37.0**
RMA	$\theta_{rma} = 0.05$	**58.6**	36.8
RMA	$\theta_{rma} = 0.10$	57.3	35.4
DA	$k = 1$	57.1	33.8
DA	$k = 2$	56.9	34.1
DA	$k = 3$	56.3	34.2
DA	$k = 4$	57.9	34.7

[1] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2397–2406, 2022.

Method Details



- 1. We select Atlas[3] as our MVS module, and FCAF3D[4] as our detection module for simplicity and fine performance.
- 2. We first pre-train our MVS and our detection modules separately, and then jointly fine-tune the entire network.

Training scheme	mAP@0.25↑	mAP@0.5↑
Joint Train From Scratch	48.2	28.8
P-MVS + JFT	50.3	30.9
P-MVS + P-Det	55.8	34.7
P-MVS + P-Det + JFT	**58.6**	**36.8**

P-MVS: Pre-train the MVS module
P-Det: Pre-train the detection module
JFT: Jointly fine-tune the entire network

[3] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, pages 414–431. Springer, 2020.

[4] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. In European Conference on Computer Vision, pages 477–493. Springer, 2022.

Experiments

- We conduct experiments on ScanNet[5] and ARKitScenes[6].
- ScanNet is a dataset with Axis-Aligned Bounding Boxes (AABB).
- ARKitScenes is a dataset with Oriented Bounding Boxes (OBB).
- We compare our methods with previous works of ImVoxelNet[1], NeRF-Det[7] and ImGeoNet[8], as well as the two-stage method.

Method	mAP@0.25↑	mAP@0.5↑
ImVoxelNet [25]	46.7	23.4
NeRF-Det [36]	53.5	27.4
ImGeoNet [28]	54.8	28.4
Atlas [19]+FCAF3D [24]	55.4	33.8
NeuralRecon [26]+FCAF3D	51.5	31.6
Ours (CN-RMA)	**58.6**	**36.8**

ScanNet Results

Method	mAP@0.25↑	mAP@0.5↑
ImVoxelNet [25]	27.3	4.3
NeRF-Det [36]	39.5	21.9
ImGeoNet [28]	60.2	43.4
Atlas [19]+FCAF3D [24]	51.3	40.6
NeuralRecon [26]+FCAF3D	36.3	24.9
Ours (CN-RMA)	**67.6**	**56.5**

ARKitScenes Results

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2017.
[6] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021.
[7] Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In ICCV, 2023.
[8] Tao Tu, Shun-Po Chuang, Yu-Lun Liu, Cheng Sun, Ke Zhang, Donna Roy, Cheng-Hao Kuo, and Min Sun. Imgeonet: Image-induced geometry-aware voxel representation for multi-view 3d object detection. In Proceedings of the IEEE international conference on computer vision, 2023.

Experiments

Method	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	mAP
ImVoxelNet [25]	31.7	83.4	71.8	67.2	55.1	31.7	15.2	36.8	2.0	33.0	63.2	24.0	53.0	20.0	91.3	53.2	76.1	32.3	46.7
NeRF-Det [36]	**42.3**	84.6	75.9	78.5	56.3	33.4	21.4	49.9	2.4	50.6	**73.9**	21.3	**54.3**	**62.5**	90.9	57.7	75.5	32.3	53.5
ImGeoNet [28]	38.7	**86.5**	76.6	75.7	**59.3**	42.0	28.1	**59.2**	4.3	42.8	71.5	36.9	51.8	44.1	95.2	58.0	79.6	36.8	54.8
Atlas+FCAF	41.6	85.4	**80.2**	81.6	54.7	38.3	27.3	50.1	7.6	58.9	73.3	16.8	36.6	61.9	94.4	58.8	**92.3**	37.4	55.4
Neucon+FCAF	38.7	82.2	78.3	81.4	56.2	30.5	12.5	42.1	6.0	54.2	64.6	20.8	34.6	41.3	89.1	67.8	89.1	37.3	51.5
Ours(CN-RMA)	**42.3**	80.0	79.4	**83.1**	55.2	**44.0**	**30.6**	53.6	**8.8**	**65.0**	70.0	**44.9**	44.0	55.2	**95.4**	**68.1**	86.1	**49.7**	**58.6**

Per-category mAP@0.25 results on ScanNet

Method	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	mAP
ImVoxelNet [25]	10.2	71.5	37.2	32.8	33.0	4.3	0.8	11.7	0.2	4.8	34.0	5.9	16.1	2.1	73.0	20.5	50.8	11.9	23.4
NeRF-Det [36]	15.8	73.1	45.3	40.6	39.5	8.1	2.0	20.3	0.2	13.8	42.5	5.3	25.3	**10.0**	63.0	26.0	49.1	12.7	27.4
ImGeoNet [28]	14.3	74.2	47.4	46.9	41.0	8.1	2.0	26.9	0.5	6.6	44.7	4.4	28.2	3.9	71.0	25.9	48.3	17.2	28.4
Atlas+FCAF	20.3	**74.8**	47.9	65.0	**44.0**	9.7	5.0	37.0	1.1	**25.3**	**51.4**	3.5	23.4	3.0	69.3	31.8	74.1	22.6	33.8
Neucon+FCAF	15.8	74.7	45.6	**68.8**	43.2	8.0	3.5	26.1	**1.2**	15.8	40.4	1.3	21.9	1.6	**74.4**	28.3	**77.8**	21.3	31.6
Ours(CN-RMA)	**21.3**	69.2	**52.4**	63.5	42.9	**11.1**	**6.5**	**40.0**	**1.2**	24.9	**51.4**	**19.6**	**33.0**	6.6	73.3	**36.1**	76.4	**31.5**	**36.8**

Per-category mAP@0.5 results on ScanNet

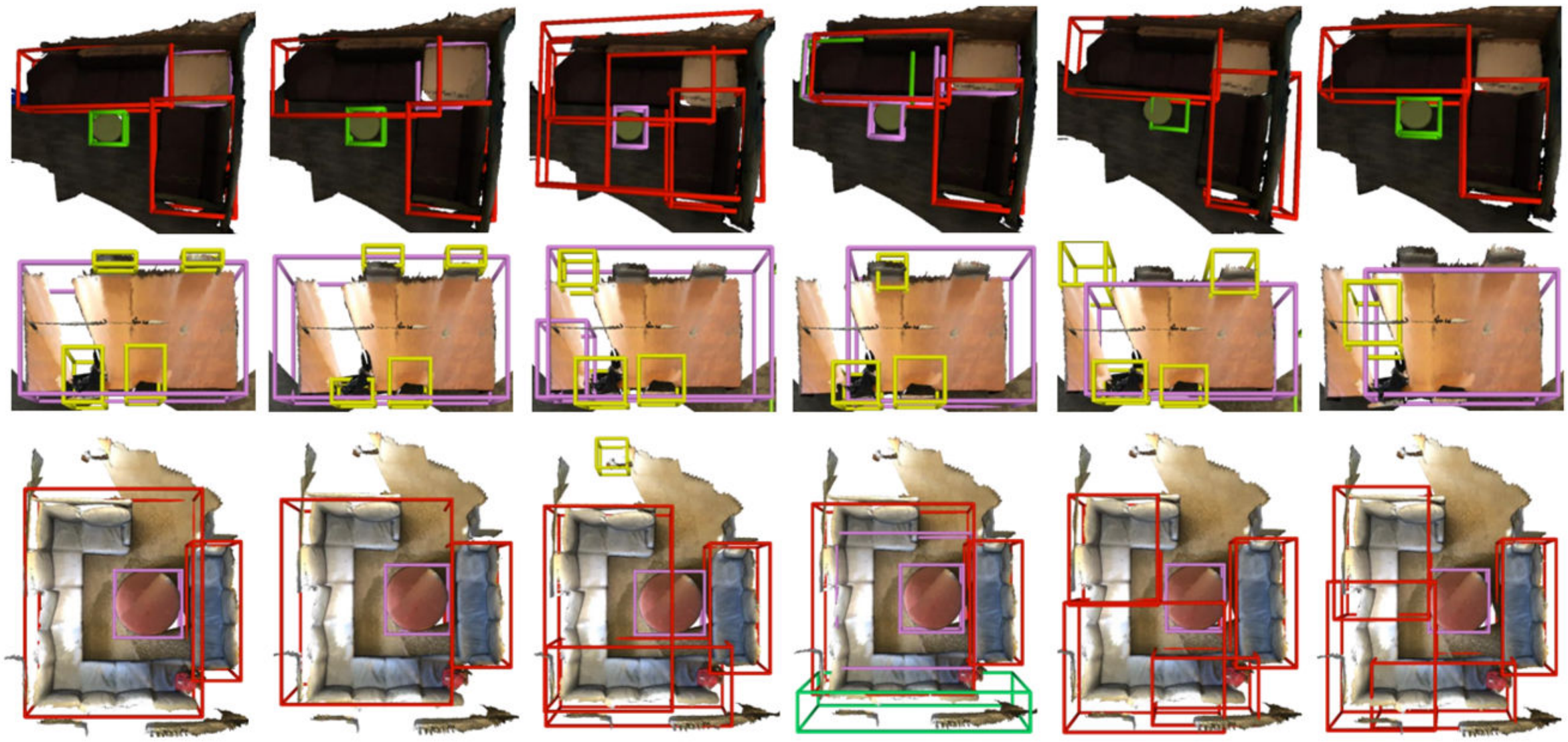
Method	cab	fridg	shlf	stove	bed	sink	wshr	tolt	bthtb	oven	dshwshr	frplce	stool	chr	tble	TV	sofa	mAP
ImVoxelNet [25]	20.7	33.3	13.5	4.3	57.7	25.8	53.8	65.2	66.0	25.5	2.2	0.2	2.5	26.7	24.5	0.0	41.6	27.3
NeRF-Det [36]	34.7	61.1	30.7	9.4	73.2	29.9	62.6	77.2	86.4	45.0	7.4	2.1	12.1	46.4	38.3	0.1	55.5	39.5
ImGeoNet [28]	55.8	**82.6**	**48.4**	20.4	89.3	52.8	**80.0**	92.5	94.7	66.0	18.1	**68.8**	30.6	72.3	70.3	2.2	79.0	60.2
Atlas+FCAF	56.0	62.7	20.9	19.5	88.4	60.7	53.1	89.7	**94.9**	42.4	3.4	48.7	23.2	63.0	62.2	2.2	80.3	51.3
Neucon+FCAF	51.2	79.5	19.6	14.8	62.5	46.0	39.6	41.2	56.5	29.0	7.1	46.9	10.4	37.1	33.2	2.8	40.6	36.3
Ours(CN-RMA)	**73.5**	82.3	47.7	**37.4**	**91.6**	**74.5**	78.0	**93.3**	93.3	**74.6**	**53.6**	67.2	**35.9**	**73.1**	**72.8**	**14.5**	**85.1**	**67.6**

Per-category mAP@0.25 results on ARKitScenes

Method	cab	fridg	shlf	stove	bed	sink	wshr	tolt	bthtb	oven	dshwshr	frplce	stool	chr	tble	TV	sofa	mAP
ImVoxelNet [25]	3.3	14.3	0.7	0.0	20.5	5.2	27.5	36.3	20.4	4.9	2.2	0.0	0.4	4.8	4.1	0.0	5.1	8.8
NeRF-Det [36]	10.8	48.0	5.7	0.6	36.1	7.9	46.3	60.8	64.9	21.0	5.6	0.0	2.9	18.8	14.1	0.0	28.2	21.9
ImGeoNet [28]	31.8	72.5	21.7	3.9	83.3	19.9	71.2	84.8	91.0	44.4	15.9	23.1	13.3	49.3	45.1	0.1	67.2	43.4
Atlas+FCAF	37.1	62.2	8.4	7.6	83.3	30.9	47.6	81.6	**94.1**	29.2	3.4	19.2	17.2	44.9	50.5	0.0	75.0	40.7
Neucon+FCAF	30.3	74.2	8.4	6.0	50.4	15.5	30.1	28.8	49.5	17.9	5.4	16.4	6.6	24.2	26.9	0.0	32.1	24.9
Ours(CN-RMA)	**56.0**	**79.8**	**27.8**	**21.3**	**87.4**	**51.6**	**75.9**	**89.1**	92.2	**60.8**	**53.3**	**40.6**	**25.1**	**60.5**	**60.1**	**1.2**	**77.3**	**56.5**

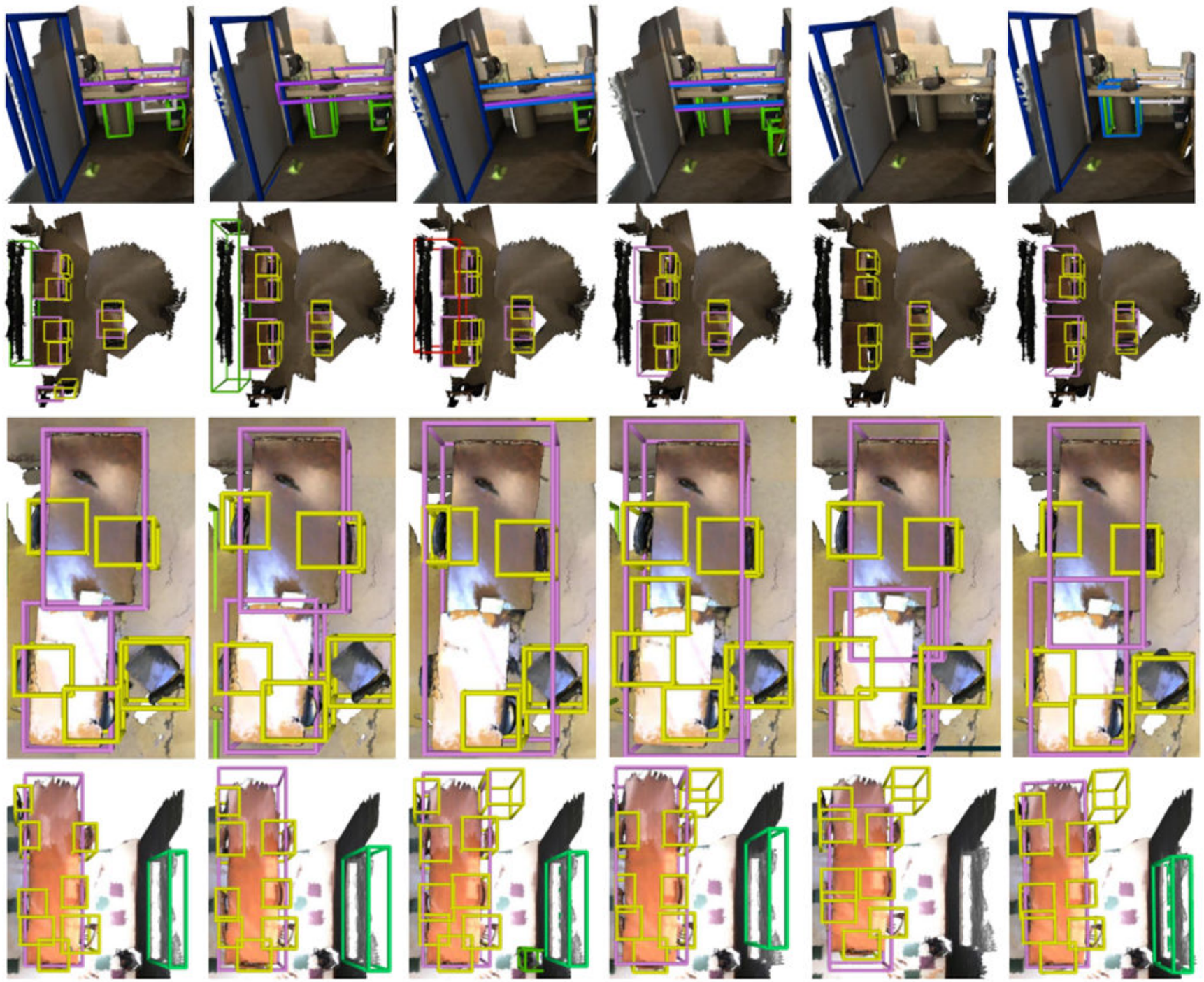
Per-category mAP@0.5 results on ARKitScenes

ScanNet Visualization Results



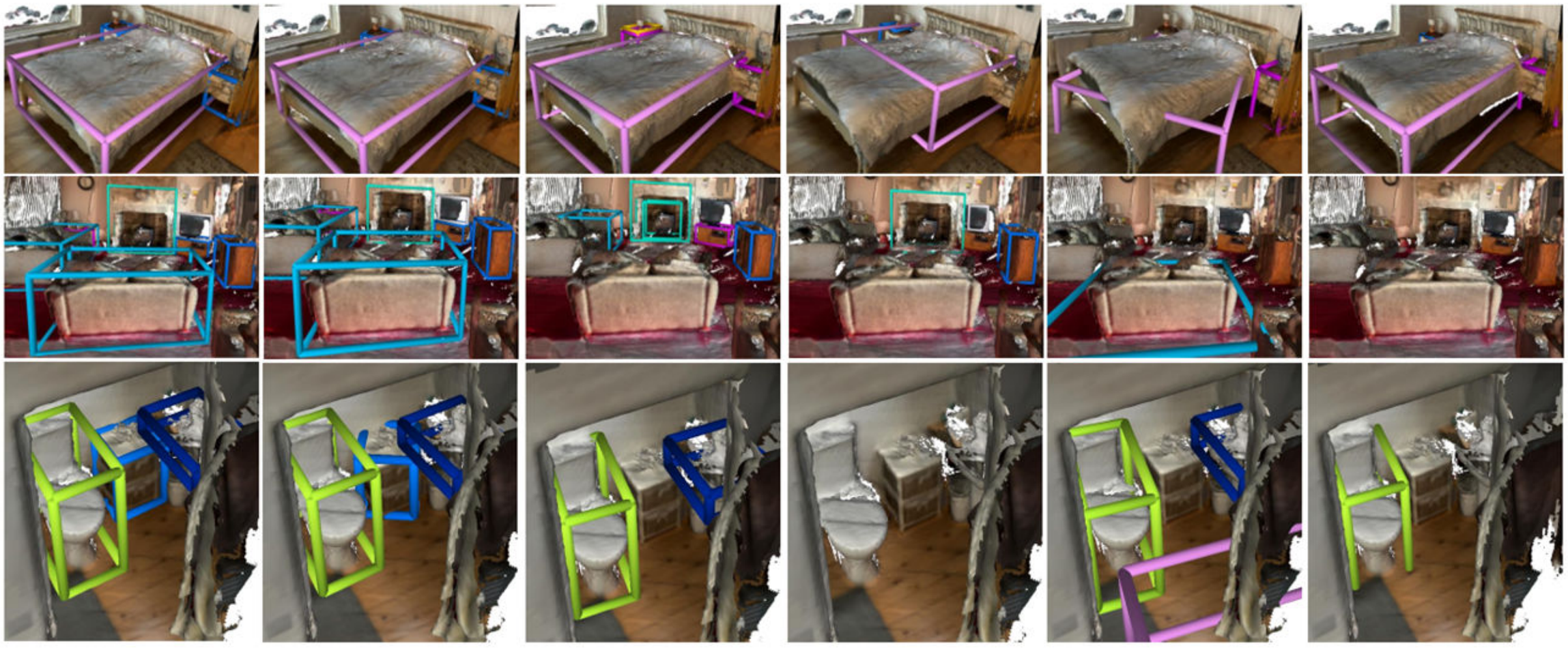
Ground Truth | CN-RMA | Atlas+FCAF | Neucon+FCAF | ImVoxelNet[25] | NeRF-Det[36]

ScanNet Visualization Results



Ground Truth | CN-RMA | Atlas+FCAF | Neucon+FCAF | ImVoxelNet[25] | NeRF-Det[36]

ARKitScenes Visualization Results



Ground Truth | CN-RMA | Atlas+FCAF | Neucon+FCAF | ImVoxelNet[25] | NeRF-Det[36]

ARKitScenes Visualization Results

Ground Truth | CN-RMA | Atlas+FCAF | Neucon+FCAF | ImVoxelNet[25] | NeRF-Det[36]

Conclusion

- Our method connects the reconstruction and detection networks seamlessly, enabling better exploitation of image features in 3D space.
- We propose an innovative occlusion-aware aggregation method, RMA, which uses the reconstructed Truncated Signed Distance Function (TSDF) to address the occlusion issues.
- Future work can focus on investigating alternative aggregation schemes or incorporating additional contextual information to further improve our RMA method.
- Our RMA aggregation method holds potential for integration into other 3D scene understanding tasks from multi-view images.

References

[1] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2397–2406, 2022.

[2] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689, 2021.

[3] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16, pages 414–431. Springer, 2020.

[4] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. In European Conference on Computer Vision, pages 477–493. Springer, 2022.

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2017.

[6] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021.

[7] Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In ICCV, 2023.

[8] Tao Tu, Shun-Po Chuang, Yu-Lun Liu, Cheng Sun, Ke Zhang, Donna Roy, Cheng-Hao Kuo, and Min Sun. Imgeonet: Image-induced geometry-aware voxel representation for multi-view 3d object detection. In Proceedings of the IEEE international conference on computer vision, 2023.

Thanks for watching!