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

Supplementary Material

6. Implementation Details

6.1. Training Details

In the pretraining stage of the Atlas [19] 2D backbone and reconstruction network (Stage 1), we directly utilize the checkpoint provided by Atlas to load our 2D backbone and 3D reconstruction network for ScanNet [8]. For ARKitScenes [1], we randomly select 50 images as input for each scene and set the voxel grid size to $(160 \times 160 \times 64)$. The network is trained for 80 epochs using the ADAM optimizer with a learning rate of 0.0005. In this stage, we perform data augmentation by applying random translations and rotations to the scenes following Atlas.

In the pretraining stage of the FCAF3D [24] detection network (Stage 2), we generate point clouds with features for each scene using our RMA method and use these point clouds to train the detection network. For ScanNet, we use 50 images and set the voxel grid size to $(256 \times 256 \times 96)$ for generating the point clouds. For ARKitScenes, we use 40 images and set the voxel grid size to $(192 \times 192 \times 80)$. We train the network for 12 epochs on both datasets, repeating the dataset 10 times for ScanNet and 3 times for ARKitScenes. The ADAM optimizer is employed with an initial learning rate of 0.001 and weight decay of 0.0001. The learning rate decreases on the 8th and 11th epochs. In this stage, we random sample $N_{pt} = 500000$ points for each scene, and perform data augmentation by randomly translating, rotating, flipping, and scaling the input point clouds, following FCAF3D.

In the joint fine-tuning stage (Stage 3), we use 40 images and set the voxel grid size to $(192 \times 192 \times 80)$. The ADAM optimizer is used with an initial learning rate of 0.001 and weight decay of 0.0001. We fine-tune the ScanNet network for 100 epochs, with the learning rate decreasing on the 80th epoch. For ARKitScenes, we fine-tune the network for 40 epochs, with the learning rate decreasing on the 27th and 36th epochs. In this stage, we do not conduct random data augmentation on input camera parameters and ground truth TSDF. Instead, after obtaining point clouds with features, we random sample $N_{pt} = 500000$ points for each scene, and perform data augmentation by randomly translating, rotating, flipping, and scaling the input point clouds, following FCAF3D.

For validation, we employ 50 images and set the grid size to $(256 \times 256 \times 96)$ for ScanNet, and we use 40 images and set the grid size to $(192 \times 192 \times 80)$ for ARKitScenes.

6.2. Experiment Details

Since the code has not been made publicly available, we directly reference the experimental results from the ImGeoNet [28] paper.

For experiments on ScanNet, we utilize the provided pretrained networks of ImVoxelNet [25], NeRF-Det [36], Atlas [19], and NeuralRecon [26]. For baselines with retraining, we retrain the FCAF3D [24] network for 12 epochs using the reconstructed scene point clouds, repeating the dataset 10 times.

For experiments on ARKitScenes, we train the Atlas network for 80 epochs and the NeuralRecon network for 15 epochs. We train the ImVoxelNet network for 12 epochs, repeating the dataset 9 times. We train the NeRF-Det network for 12 epochs without repeating the dataset. For baselines with retraining, we retrain the FCAF3D network for 12 epochs using the reconstructed scene point clouds, repeating the dataset 3 times.

7. Additional Experiment Results

The results of Table 5 illustrate that retraining the reconstruction network of the two-stage baseline with reconstructed scene point clouds is better than directly using the network trained by ground truth point clouds. The per-category mAP@0.25 and mAP@0.5 scores of experiments on the ScanNet [8] and ARKitScenes [1] datasets are shown in Tables 6, 7, 8, and 9. Additional visualization results of ScanNet and ARKitScenes are presented in Figures 6 and 7.

Method	Dataset	Retrain	mAP@0.25↑	mAP@0.5↑
Atlas [19]+FCAF3D [24]	ScanNet[8]	✓	55.4	33.8
Atlas+FCAF3D	ScanNet		39.4	22.1
NeuralRecon [26]+FCAF3D	ScanNet	✓	51.5	31.6
NeuralRecon+FCAF3D	ScanNet		29.6	13.1
Atlas+FCAF3D	ARKitScenes [1]	✓	51.3	40.7
Atlas+FCAF3D	ARKitScenes		34.1	25.8
NeuralRecon+FCAF3D	ARKitScenes	✓	36.3	24.9
NeuralRecon+FCAF3D	ARKitScenes		13.6	9.1

Table 5. Comparison of retraining for two-stage baselines.

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**

Table 6. Per-category AP@0.25 scores for 18 categories from the ScanNet [8] dataset. Atlas+FCAF denotes the two-stage baseline combining Atlas [19] and FCAF3D [24], and Neucon+FCAF denotes the two-stage baseline combining NeuralRecon [26] and FCAF3D. We directly cite the experimental results from the ImGeoNet [28] paper.

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**

Table 7. Per-category AP@0.5 scores for 18 categories from the ScanNet [8] dataset. Atlas+FCAF denotes the two-stage baseline combining Atlas [19] and FCAF3D [24], and Neucon+FCAF denotes the two-stage baseline combining NeuralRecon [26] and FCAF3D. We directly cite the experimental results from the ImGeoNet [28] paper.

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**

Table 8. Per-category AP@0.25 scores for 17 categories from the ARKitScenes [1] dataset. Atlas+FCAF denotes the two-stage baseline combining Atlas [19] and FCAF3D [24], and Neucon+FCAF denotes the two-stage baseline combining NeuralRecon [26] and FCAF3D. We directly cite the experimental results from the ImGeoNet [28] paper.

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**

Table 9. Per-category AP@0.5 scores for 17 categories from the ARKitScenes [1] dataset. Atlas+FCAF denotes the two-stage baseline combining Atlas [19] and FCAF3D [24], and Neucon+FCAF denotes the two-stage baseline combining NeuralRecon [26] and FCAF3D. We directly cite the experimental results from the ImGeoNet [28] paper.

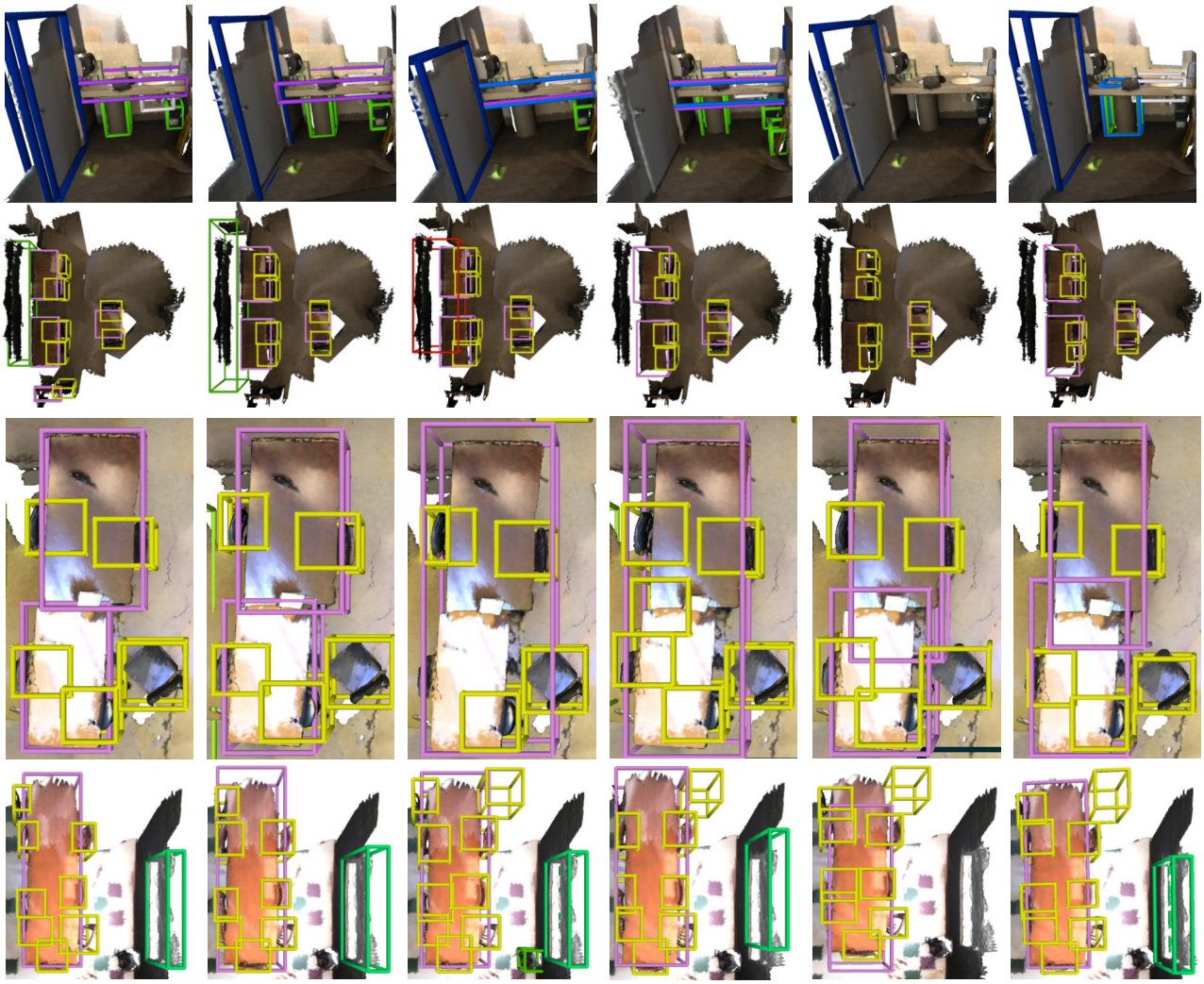

Figure 6. **Visualization of 3D object detection results from ScanNet [8].** From above to below are scene0084_00, scene0609_00, scene0030_00, and scene0599_02 from ScanNet. Atlas+FCAF denotes the two-stage baseline combining Atlas [19] and FCAF3D [24], and Neucon+FCAF denotes the two-stage baseline combining NeuralRecon [26] and FCAF3D.

Figure 7. **Visualization of 3D object detection results from ARKitScenes [1].** From above to below are scenes 47331266, 41069021, 42897538, 42445028, 45663114, 44358513, and 47333904 from ARKitScenes. Atlas+FCAF denotes the two-stage baseline combining Atlas [19] and FCAF3D [24], and Neucon+FCAF denotes the two-stage baseline combining NeuralRecon [26] and FCAF3D.