Supplementary Material for "ReAGFormer: Reaggregation Transformer with Affine Group Features for 3D Object Detection"

Chenguang Lu[1], Kang Yue[2,1], and Yue Liu[1]✉

[1] Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Beijing Institute of Technology, Beijing, China
chenguang.lu@outlook.com, liuyue@bit.edu.cn
[2] Institute of Software, Chinese Academy of Sciences, Beijing, China
einhep@gmail.com

In this supplementary material, we provide additional details on our method and more evaluation results. First, we provide more details on our ReAGFormer in Sec. 1 and Sec. 2, including network architecture and training details. Then, more ablation studies are provided in Sec. 3. Moreover, we also report the per-category quantitative results of our method on the ScanNet V2 [1] and SUN RGB-D [2] datasets in Sec. 4. Finally, we show more qualitative results in Sec. 5.

1 Network Architecture

Our ReAGFormer consists of four downsampling stage. The first stage is a standard set abstraction layer [3], and all other stage consist of group embedding and reaggregation Transformer block with affine group features (ReAGF Transformer block). Note that we do not use the ReAGF Transformer block in stage 1, we argue that the point feature extraction is not complete in the shallower layers, and therefore the dependencies between points cannot be effectively modeled using Transformer [4] in the early stages. Another advantage of such an architecture is the reduced computational cost, since the shallow layer has more points and modeling dependencies on n points requires $O(n^2)$ computational complexity. For upsampling stage, we use 2 feature propagation layers with multi-scale connections. The model architecture details are shown in Table 1.

In the ReAGF Transformer block, we set the number of heads of ASA and RCA to 8, and use 2 MLP layers. The feature dimension of each MLP layer is set to $[C_{in}, 4C_{in}, C_{in}]$ where C_{in} is the dimension of input feature. All MLP layers and attention modules use dropout with the probability 0.1.

To eliminate the semantic gap between group embedding and ReAGF Transformer block, we introduce the feature connection bridge. From group embedding to ReAGF Transformer block, we adopt linear projection and layer normalization, without the activation function. For the output of the ReAGF Transformer block, we adopt the linear projection, batch normalization and ReLU activation function before being fed to the subsequent group embedding.

Table 1. Network architecture details. "SA", "ReAGF" and "M-FP" denote the set abstraction layer, the ReAGF Transformer block and the feature propagation layer with multi-scale connections, respectively. "Input" and "Output" are the number of input and output points. We follows the corresponding baseline [5–7] to set N_{in}. r indicates the ball query radius in the group embedding, and k is the number of points in each group. Shared MLP$[c_1, c_2, ..., c_i]$ indicate the feature dimensions of the MLP in the group embedding or feature propagation layer where c_i denotes the feature dimension of the i-th layer.

Stage	Type	Input	Output	r	k	Shared MLP
stage 1	SA	N_{in}	2048	0.2	64	[3+1,64,64,128]
stage 2	ReAGF	2048	1024	0.4	32	[128+3,128,128,288]
stage 3	ReAGF	1024	512	0.8	16	[288+3,128,128,288]
stage 4	ReAGF	512	256	1.2	16	[288+3,128,128,288]
upsampling	M-FP	256	512	-	-	[288+288,288,288]
upsampling	M-FP	512	1024	-	-	[288+288+288+288,288,288]

2 Training Details

In this section, we provide more training details on the three models, *i.e.* ReAGF-VoteNet, ReAGF-BRNet and ReAGF-Group-Free.

ReAGF-VoteNet. For the ScanNet V2 dataset, we follow VoteNet [5] to randomly sample 40K points as input, and each point contains coordinate information and height information but does not use color information. During data augmentation, we randomly flip the point cloud, rotate each point uniformly in the range $[-5°, 5°]$ and scale each point uniformly in the interval [0.9, 1.1]. The model is optimized using the AdamW optimizer [8] with the batch size 8 and initial learning rate 0.002 for 240 epochs. The learning rate is decayed at 160 epoch and 200 epoch with a decay ratio of 0.1. The learning rate of the Transformer block is always 1/20 of the learning rate of the other part. We apply a weight decay of 0.1 and a gradient normalized clipping with a maximum norm of 10.

For SUN RGB-D, we follow VoteNet to use 20K points as input and train model for 180 epochs with the batch size 8 and initial learning rate 0.001. The learning rate is decayed by 0.1 at 120 epoch and 160 epoch. Other training settings are the same as ReAGF-VoteNet on ScanNet V2. For both datasets, the model is implemented with MMDetection3D [9] and trained on a single GPU.

ReAGF-BRNet. For the ScanNet V2 dataset, we follow the input and data augmentation of BRNet [6], using 40K points as input and applying random flips, random rotations and random scaling. Each point uses both coordinate information and height information. We train model with the AdamW optimizer [8]. The batch size, initial learning rate and weight decay are set as 8, 0.002 and 0.01, respectively. We train the model for 220 epochs and decay the learning

rate by 10× after 140 epochs and 180 epochs. The learning rate of the Transformer block is 1/20 of the other parts. We follow BRNet [6] to apply a gradient normalized clipping, and the maximum norm is set to 10.

For the SUN RGB-D dataset, we also follow BRNet [6] to randomly sample 20K points from raw point clouds as input. We train the model with the initial learning rate 0.001 and batch size 8. The other settings follow ReAGF-BRNet on ScanNet V2 dataset. For the SUN RGB-D and ScanNet V2 datasets, we implement our model based on the officially released code of BRNet and train it on a single GPU.

ReAGF-Group-Free. For the ScanNet V2 dataset, we follow Group-Free [7] to use 50K points as input and use random flip, random rotation in the interval $[-5°, 5°]$ as well as random scaling in the range [0.9, 1.1]. We train the network with the batch size 8, initial learning rate 0.0015 for 400 epochs, using the AdamW optimizer [8]. Following Group-Free, the learning rate is decayed by 0.1 at 280 epoch and 340 epoch. In the training stage, the weight decay is set to 5e-4 and a gradient normalized clipping with a maximum norm of 0.1 is applied. The learning rate of the Transformer block is set to 1/20 of the rest part of the network. Our experiments are conducted on the MMDetection3D toolbox. We train the model on a single GPU.

For the SUN RGB-D dataset, we follow the input of Group-Free, *i.e.* using 20K points as input for each point clouds. The batch size and initial learning rate are set to 8 per-GPU and 0.002, respectively. The rest of the implementation is consistent with ReAGF-Group-Free on ScanNet V2 dataset. Our model is implemented based on the officially released code of Group-Free and trained on two GPUs

3 More Ablation Studies

In this section, we conduct more ablation studies to analyze our model. Our experiments are trained on ReAGF-VoteNet, and evaluated on the ScanNet V2 val set.

Different Feature Reaggregation Methods. To select the appropriate symmetric function for feature reaggregation in our model, we investigate the performance using different symmetric function (*i.e.* max pooling and average pooling) while keeping the rest part of network fixed, and the results are shown in Table 2. We observe that max pooling achieves better detection results, thus we select max pooling in our ReAGFormer.

Multi-scale Connection on Skip Connection. Except for introducing multi-scale connection on the feature propagation layer, we also use multi-scale connection on the skip connection between the downsampling and upsampling stage. Table 3 ablates the effectiveness of multiscale connection on skip connection. We can observe that multi-scale form of skip connection is beneficial to improve

Table 2. Ablation study on the performance of different feature reaggregation methods.

Method	mAP@0.25	mAP@0.5
Average pooling	62.8	40.7
Max pooling	**66.1**	**45.4**

Table 3. Ablation study on the effectiveness of multi-scale connection on skip connection.

Multi-scale connection on skip connection	mAP@0.25	mAP@0.5
-	65.2	44.4
✓	**66.1**	**45.4**

the detection performance, which indicates that it can further boost the feature fusion efficiency.

Comparison with affine transformation in PointMLP. Both our proposed method and PointMLP [10] use affine transformation module. However, the methodology is also different. We use the MLP in the affine transformation (AT) module to align different group features. Let us disregard the activation function and BN for a moment, at this point the MLP can be represented as $Mul(F, \Phi)$, where $F \in \mathbb{R}^{k \times C}$ is the group features, $\Phi \in \mathbb{R}^{C \times C}$ is the parameter matrix, and $Mul(\cdot)$ is the matrix multiplication. When Φ is a diagonal matrix, the diagonal parameters can be transformed into $\phi \in \mathbb{R}^{C}$, and $Mul(F, \Phi)$ can be equated to $F \odot \phi$, where $\odot$ is the Hadamard product, *i.e.*, it is transformed into the form of the affine transformation of PointMLP. This means that the affine transformation in PointMLP is actually a special case of our AT module and our method is more generalized. In addition, we also apply PointMLP to the 3D object detection task. Specifically, we replaced the original backbone network in VoteNet [5] with PointMLP without changing the other network architectures. We compare its result with our ReAGF-VoteNet, as shown in Table 4.

4 More Quantitative Results

We report per-category results on the ScanNet V2 and SUN RGB-D datasets. Table 5 and Table 6 show the per-category results for different 3D IoU thresholds on ScanNet V2, respectively. Table 7 and Table 8 show the per-category results for different 3D IoU thresholds on SUN RGB-D dataset, respectively. We observe that our method can significantly improve detection performance, especially for cluttered, thin or similarly shaped objects. For example, for the ScanNet V2 dataset, our ReAGF-BRNet obtains 3.9%, 6.4% and 9.0% improvement compared to BRNet [6] for bookshelf, window and shower curtrain, respectively, on the more challenging mAP@0.5. Similar results can be observed with our method on the SUN RGB-D dataset.

Table 4. Comparison of our ReAGF-VoteNet and PointMLP for 3D object detection.

Method	SUN RGB-D		ScanNet V2	
	mAP@0.25	mAP@0.5	mAP@0.25	mAP@0.5
PointMLP + VoteNet	60.9	35.9	62.2	39.5
ReAGF-VoteNet	**62.3**	**40.7**	**66.1**	**45.4**

Table 5. 3D object detection results of mAP@0.25 for per-category on ScanNet V2 dataset. VoteNet* indicates that the implementation is based on MMDetection3D [9], which has better results than the original paper [5]. "+Ours" denotes replacing the original backbone with our ReAGFormer without changing the other network architectures. For Group-Free [7], we report the results for 6-layer decoder and 256 object candidates.

Method	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	mAP@0.25
VoteNet* [5]	47.7	**88.7**	89.5	**89.3**	62.1	54.1	40.8	54.3	12.0	63.9	69.4	52.0	52.5	73.3	95.9	52.0	**92.5**	42.4	62.9
+Ours (ReAGF-VoteNet)	**53.9**	88.4	89.5	86.9	**68.3**	**57.7**	**51.5**	**58.3**	16.5	**66.4**	**71.3**	**61.2**	**53.1**	**74.4**	98.4	51.0	91.6	**50.6**	**66.1**
BRNet [6]	49.3	**88.3**	91.9	86.9	**69.3**	59.2	45.9	52.1	15.3	**72.0**	**76.8**	57.1	60.4	73.6	93.8	58.8	**92.2**	47.1	66.1
+Ours (ReAGF-BRNet)	**51.7**	86.8	**92.5**	**90.6**	67.0	**60.1**	**52.4**	**58.4**	**18.5**	67.3	72.7	**64.1**	**60.7**	**75.2**	**96.5**	**61.9**	88.8	**48.7**	**67.4**
Group-Free [7]	**54.1**	86.2	92.0	84.8	**67.8**	55.8	46.9	**48.5**	15.0	59.4	80.4	64.2	**57.2**	76.3	97.6	**76.8**	92.5	55.0	**67.3**
+Ours (ReAGF-Group-Free)	50.7	**86.9**	**92.3**	**85.9**	67.6	**59.0**	**47.2**	39.5	**17.6**	**61.1**	**80.6**	**65.3**	55.8	**79.8**	**99.2**	67.1	**92.9**	**59.4**	67.1

Table 6. 3D object detection results of mAP@0.5 for per-category on ScanNet V2 dataset. VoteNet* indicates that the implementation of the result is based on the MMDetection3D [9] toolbox, which has better results than the original paper [5]. "+Ours" denotes replacing the original backbone with our ReAGFormer without changing other network architectures. For Group-Free [7], we report the results for 6-layer decoder and 256 object candidates.

Method	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	mAP@0.5
VoteNet* [5]	14.6	77.8	73.1	**80.5**	46.5	25.1	16.0	41.8	2.5	22.3	33.3	25.0	31.0	17.6	87.8	23.0	81.6	18.7	39.9
+Ours (ReAGF-VoteNet)	**20.4**	**80.6**	**77.4**	72.0	**51.4**	**27.8**	**23.3**	**48.3**	**6.4**	**27.7**	43.5	**32.3**	**43.8**	**30.1**	**91.0**	28.8	80.5	**32.3**	**45.4**
BRNet [6]	28.7	80.6	81.9	80.6	**60.8**	35.5	22.2	48.0	**7.5**	**43.7**	**54.8**	39.1	**51.8**	35.9	88.9	38.7	84.4	33.0	50.9
+Ours (ReAGF-BRNet)	**30.5**	**80.7**	**84.4**	**86.6**	60.2	**38.5**	**28.6**	**51.9**	5.1	33.9	50.3	**40.9**	47.1	**44.9**	**94.0**	**38.8**	**88.4**	**35.0**	**52.2**
Group-Free [7]	23.0	78.4	78.9	68.7	55.1	**35.3**	**23.6**	**39.4**	7.5	27.2	**66.4**	43.3	**43.0**	**41.2**	89.7	38.0	83.4	37.3	48.9
+Ours (ReAGF-Group-Free)	**25.7**	**80.3**	**80.6**	**74.2**	**57.6**	34.5	21.5	34.8	**8.5**	**35.6**	63.4	**48.2**	41.8	31.6	**91.3**	**41.4**	**86.8**	**43.1**	**50.0**

Table 7. 3D object detection results of mAP@0.25 for per-category on SUN RGB-D dataset. VoteNet* indicates that the implementation of the result is based on the MMDetection3D [9] toolbox, which has better results than the original paper [5]. "+Ours" denotes replacing the original backbone with our ReAGFormer without changing other network architectures. For Group-Free [7], we report the results for 6-layer decoder and 256 object candidates.

Method	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP@0.25
VoteNet* [5]	75.5	**85.6**	31.9	**77.4**	24.8	27.9	58.6	**67.4**	51.1	90.5	59.1
+Ours (ReAGF-VoteNet)	76.8	85.4	**35.2**	76.4	**30.7**	**38.8**	**69.0**	67.0	**53.4**	90.6	**62.3**
BRNet [6]	**76.2**	86.9	29.7	77.4	29.6	**35.9**	65.9	66.4	51.8	91.3	61.1
+Ours (ReAGF-BRNet)	74.3	**87.7**	**32.2**	**78.3**	**30.2**	33.2	**66.5**	**68.4**	**52.5**	**91.4**	**61.5**
Group-Free [7]	**80.0**	**87.8**	**32.5**	**79.4**	**32.6**	36.0	**66.7**	**70.0**	53.8	91.1	**63.0**
+Ours (ReAGF-Group-Free)	79.2	87.1	31.6	78.6	32.0	**38.1**	65.4	69.7	**55.1**	**91.7**	62.9

5 More Qualitative Results

We provide additional visualization results on the ScanNet V2 and SUN RGB-D datasets. Fig. 1 and Fig. 2 show the results on the ScanNet V2 dataset. Fig. 3 and

Table 8. 3D object detection results of mAP@0.5 for per-category on SUN RGB-D dataset. VoteNet* indicates that the implementation of the result is based on the MMDetection3D [9] toolbox, which has better results than the original paper [5]. "+Ours" denotes replacing the original backbone with our ReAGFormer without changing other network architectures. For Group-Free [7], we report the results for 6-layer decoder and 256 object candidates.

Method	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP@0.5
VoteNet* [5]	45.4	53.4	6.8	56.5	5.9	12.0	38.6	49.1	21.3	**68.5**	35.8
+Ours (ReAGF-VoteNet)	**48.4**	**59.6**	**12.2**	**57.8**	**8.2**	**24.8**	**52.8**	**53.4**	**24.3**	66.1	**40.7**
BRNet [6]	55.5	63.8	9.3	61.6	10.0	**27.3**	**53.2**	56.7	28.6	**70.9**	43.7
+Ours (ReAGF-BRNet)	**61.5**	**65.7**	**13.1**	**63.7**	**12.7**	23.6	49.8	**60.0**	**31.5**	66.7	**44.8**
Group-Free [7]	**64.0**	67.1	12.4	62.6	14.5	21.9	49.8	**58.2**	29.2	**72.2**	45.2
+Ours (ReAGF-Group-Free)	58.0	**67.9**	**13.5**	62.6	**15.0**	**23.5**	**56.5**	56.8	**31.4**	71.9	**45.7**

Fig. 4 show the results on the SUN RGB-D dataset. The figures demonstrate that our method can generate more reasonable and accurate results. Fig. 5 visualizes the weights learned by our reaggregation cross-attention (RCA). We can observe that the reference point (green dot) focuses more on the points that belong to the same object as itself, which shows that our RCA enables the symmetric function to focus more on the object-level information.

References

1. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR. pp. 5828–5839 (2017)
2. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: CVPR. pp. 567–576 (2015)
3. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. NeurIPS **30**, 5099–5108 (2017)
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)
5. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: ICCV. pp. 9277–9286 (2019)
6. Cheng, B., Sheng, L., Shi, S., Yang, M., Xu, D.: Back-tracing representative points for voting-based 3d object detection in point clouds. In: CVPR. pp. 8963–8972 (2021)
7. Liu, Z., Zhang, Z., Cao, Y., Hu, H., Tong, X.: Group-free 3d object detection via transformers. In: ICCV (2021)
8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
9. Contributors, M.: MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d (2020)
10. Ma, X., Qin, C., You, H., Ran, H., Fu, Y.: Rethinking network design and local geometry in point cloud: A simple residual mlp framework. In: International Conference on Learning Representations (2021)

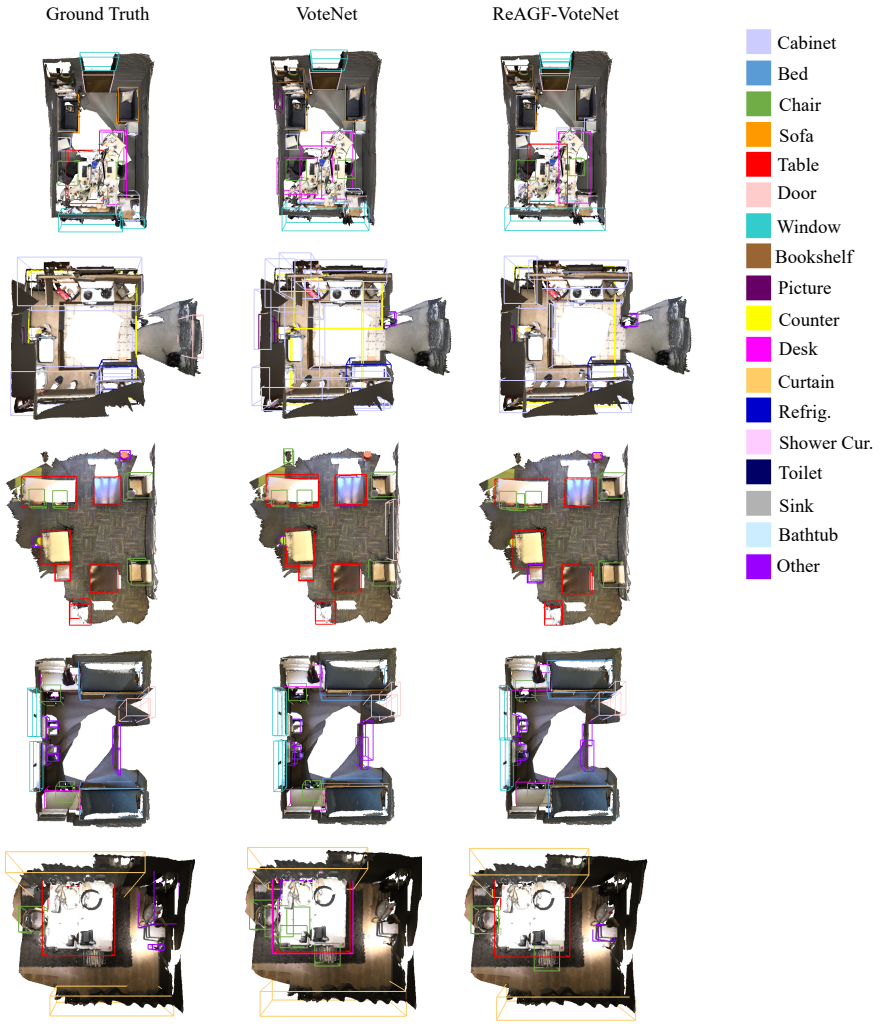

Fig. 1. Qualitative detection results on ScanNet V2 dataset. ReAGF-VoteNet denotes the replacement of the baseline original backbone with our ReAGFormer. Color is used for better illustration purpose, and it is not used in the experiment. (*Best viewed in color.*)

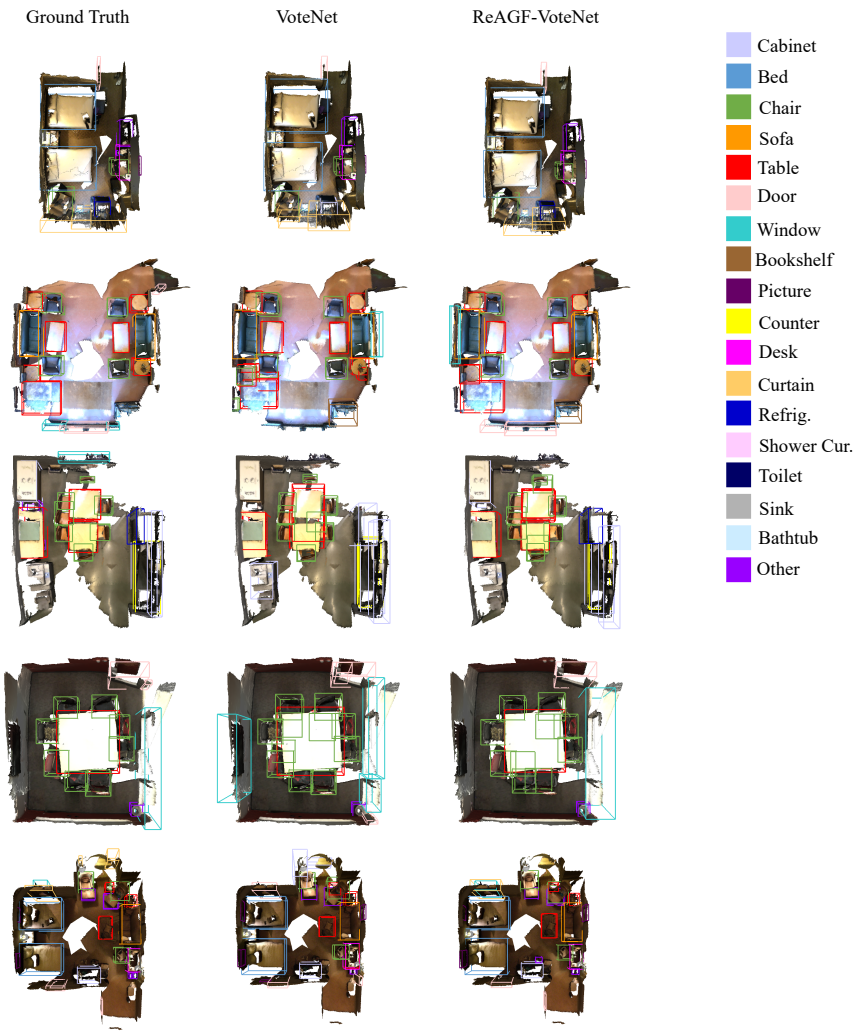

Fig. 2. Qualitative detection results on ScanNet V2 dataset. ReAGF-VoteNet denotes the replacement of the baseline original backbone with our ReAGFormer. Color is used for better illustration purpose, and it is not used in the experiment. (*Best viewed in color.*)

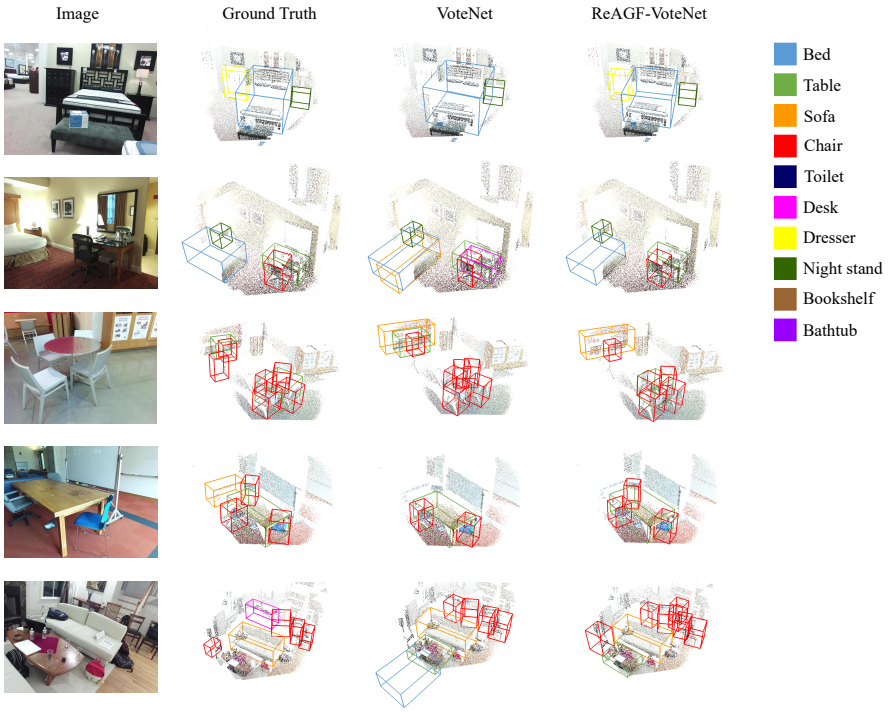

Fig. 3. Qualitative detection results on SUN RGB-D dataset. ReAGF-VoteNet denotes the replacement of the baseline original backbone with our ReAGFormer. Images and colors are only used for better illustration, and they are not used in our network. (*Best viewed in color.*)

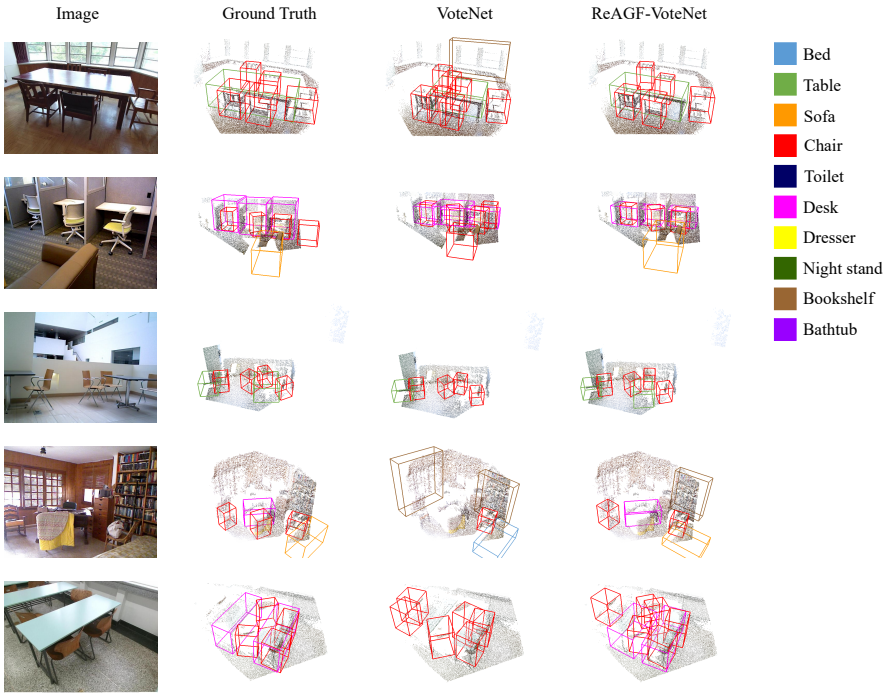

Fig. 4. Qualitative detection results on SUN RGB-D dataset. ReAGF-VoteNet denotes the replacement of the baseline original backbone with our ReAGFormer. Images and colors are only used for better illustration, and they are not used in our network. (*Best viewed in color.*)

Fig. 5. Visualization of attention weights for reaggregation cross-attention (RCA). Green dot indicates reference point, and redder color indicates greater weight.