Distribution Aware VoteNet for 3D Object Detection

Junxiong Liang*, Pei An*, Jie Ma†

Huazhong University of Science and Technology
{liangjunxiong, anpei, majie}@hust.edu.cn

Abstract

Occlusion is common in the actual 3D scenes, causing the boundary ambiguity of the targeted object. This uncertainty brings difficulty for labeling and learning. Current 3D detectors predict the bounding box directly, regarding it as Dirac delta distribution. However, it does not fully consider such ambiguity. To deal with it, distribution learning is used to efficiently represent the boundary ambiguity. In this paper, we revise the common regression method by predicting the distribution of the 3D box and then present a distribution-aware regression (DAR) module for box refinement and localization quality estimation. It contains scale adaptive (SA) encoder and joint localization quality estimator (JLQE). With the adaptive receptive field, SA encoder refines discriminative features for precise distribution learning. JLQE provides a reliable location score by further leveraging the distribution statistics, correlating with the localization quality of the targeted object. Combining DAR module and the baseline VoteNet, we propose a novel 3D detector called DAVNet. Extensive experiments on both ScanNet V2 and SUN RGB-D datasets demonstrate that the proposed DAVNet achieves significant improvement and outperforms state-of-the-art 3D detectors.

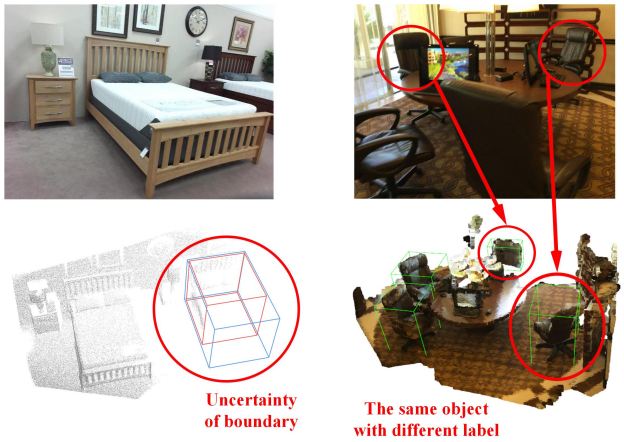

Figure 1: Occlusion is common in 3D scenes. (a) The boundary of a partly-seen object is ambiguous. (b)The ground truth size of the same object may be different due to the occlusion.

Introduction

3D object detection has a pivotal role in the field of robotic scene perception. Current 3D detectors (Qi et al. 2019; Shi, Wang, and Li 2019) have made progress these years. However, they regress 3D bounding box directly, limited in Dirac delta distribution. They do not fully consider the label ambiguity of the occluded object. Occlusion is common in the 3D scene. As shown in Figure 1 (a), we can hardly confirm the size of a partly seen object. The position and the size of an incomplete target tend to be ambiguous even though they were manually labeled. Dirac delta distribution fails to represent such uncertainty.

To solve the label ambiguity, some researchers made efforts in the 2D domain (He et al. 2019). GFocalLoss (Li et al. 2020b) has confirmed the effectiveness of distribution representation for box boundary. It predicts the general distribution of the distance between the anchor point and four boundaries. However, it is unsuitable for 3D object detection for the inconsistency with the common 3D box representation. Most 3D detectors regress the residual offset of the center, size, and orientation. Hence, we keep this representation method and learn the distribution of the offsets above (see Figure 2). We design distribution aware regression (DAR) module for precise boundary perception.

For accurate distribution prediction, we aim to improve the procedure of feature extraction. VoteNet (Qi et al. 2019) proposed an approach for proposal generation by predicting the corresponding center of each point. Nevertheless, it used a fixed region for set abstraction, which is unsuitable for objects of different sizes. The vote point remains a certain distance from the center when the voting step brings unsatisfactory prediction, especially for the large object (see Figure 3 (a)). A fixed region brings unbalanced pooling quality to different categories. Besides, the set abstraction performance heavily depends on the voting accuracy. To avoid such concerns, we adopt scale adaptive (SA) encoder to re-pooling the features with a rough 3D proposal. According to the pri-

*These authors contributed equally.
†Corresponding author

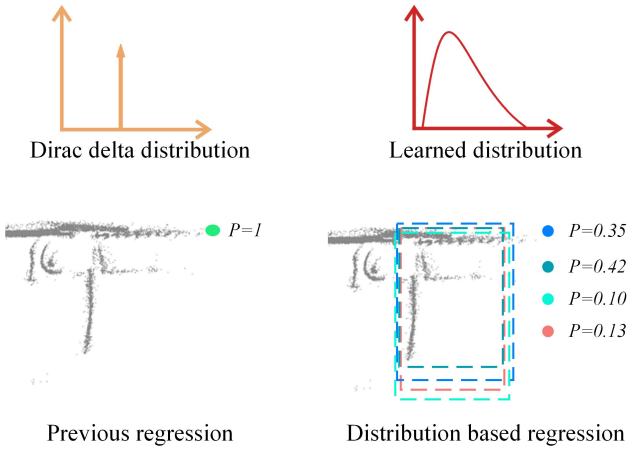

Figure 2: The comparison of two regression methods. P stands for the probability of each box. Utilizing the distribution is more in line with the data characteristics.

or result, SA encoder adjusts the receptive field adaptively to extract distribution aware representation.

Except the location, the box confidence is also a crucial aspect for prediction. For the box scoring, localization quality estimation is one of the essential parts. VoteNet treats the distance between vote points and their corresponding centers as the location score. It is not a proper way to represent the localization quality. Recent studies (Zheng et al. 2021; Wang et al. 2020) utilize an individual branch for IoU prediction. Coincidentally, GFocalLoss (Li et al. 2020a) points out the relationship between the boundary distribution and localization quality in 2D object detection. A flat distribution probably leads to inaccurate boundaries while a sharp one may bring a precise result. We also found this phenomenon in the 3D scene. Hence, we propose joint localization quality estimator (JLQE), which combines the physical feature and distribution feature for location confidence prediction. Furthermore, we refine Distribution-Guided Quality Predictor (DGQP) (Li et al. 2020a) and provide a robust distribution encoder. By using the proposed JLQE, we can obtain a reliable box score.

In response to the shortcomings mentioned above, we develop a novel 3D object detector called DAVNet, based on VoteNet. It adopts a distribution-guided box regression to generate a precise bounding box with reliable confidence. The contribution of this paper can be summarized as below:

- The proposed DAR module conducts box refinement by regressing the distribution, which can indicate the boundary uncertainty. We revise the distribution encoder and propose JLQE, which provides a reliable location score by further leveraging the learned distribution. SA encoder extracts feature with more physical information by applying adaptive receptive field.
- The proposed DAVNet achieves significant improvement on SUN RGB-D (mAP@0.25: 60.32) and ScanNet V2 (mAP@0.25: 67.11). The proposed network outperforms state-of-the-art methods on these two datasets.

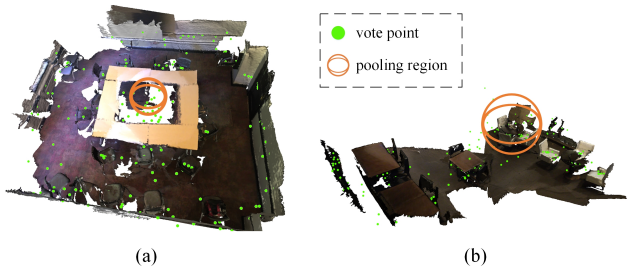

Figure 3: Using a unified pooling region is not suitable for all scenes. (a) A small region may lose some valid points when RPN brings poor effect. (b) A large one may contain more noise points in a complex scene.

Related Work

3D Object Detection from Point Cloud

With the rapid development of the 2D object detection method (Girshick 2015; Ren et al. 2015), more recent attention has focused on the 3D domain. There comes a lot of enlightening literature. At early stage, there were some template-based methods (Li et al. 2015; Nan, Xie, and Sharf 2012). After that, deep network are widely used such as (Song and Xiao 2016; Hou, Dai, and Nießner 2019; Yi et al. 2019). To deal with the sparsity and the disorder of point clouds, some researchers (Su et al. 2015; Qi et al. 2016; Wei, Yu, and Sun 2020) project them to the multiple views and extract view-wise features. Others (Maturana and Scherer 2015; Wu et al. 2015; Ben-Shabat, Lindenbaum, and Fischer 2017) transform the point cloud into a set of voxel for 3D convolution. PointNet (Qi et al. 2017a,b) provides new insight into feature extraction, by learning point-wise features directly. PointRCNN (Shi, Wang, and Li 2019) is a pioneering work, providing a two-stage 3D detector that is widely used. VoteNet (Qi et al. 2019) utilize the Hough voting for proposal generation and its variant, MLCVNet (Xie et al. 2020), captures the contextual information with self-attention. However, previous works do not pay enough attention to the label uncertainty on the dataset, limited in Dirac delta distribution. In this paper, we take advantage of the learned distribution to indicate such ambiguity, leading to a significant improvement.

Representation of Bounding Box

Most of the existing works describe the bounding box in the form of Dirac delta distribution. Gaussian YOLOv3 (Choi et al. 2019) introduces the boundary uncertainty to re-score the bounding box and KL Loss (He et al. 2019) confirms the effectiveness of distribution regression for box refinement. However, both of them adopt a Gaussian assumption and predict the variance. GFocalLoss (Li et al. 2020b,a) take a step forward. They relax the assumption and predict a more flexible distribution.

Prior studies mentioned above focus on the 2D domain. CaDNN (Reading et al. 2021) adopts pixel-wise depth distribution learning for monocular 3D object detection. However, most of these studies have been limited to image input,

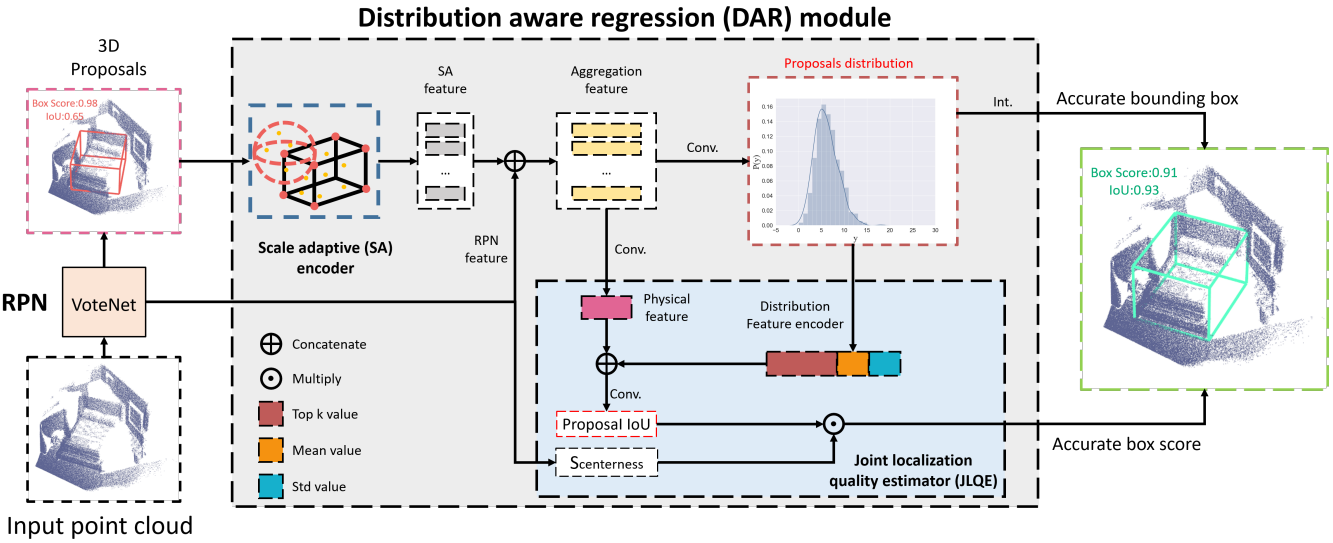

Figure 4: Overview of the proposed DAVNet. DAR module aims at box refinement. It regresses the distribution of the box location and provides precise prediction by integration. Inside, SA encoder generates discriminative features for distribution learning, which is utilized by JLQE to provide a reliable location score.

unsuitable for point cloud data. Concerning such inconsistency, we provide new insights into point cloud 3D object detection by introducing distribution representation.

Localization Quality Estimation

Localization quality estimation plays a vital role in box scoring. Some existing works (Tychsen-Smith and Petersson 2018; Jiang et al. 2018; Wu, Li, and Wang 2020) set a separated branch to predict the IoU or center score in the 2D domain. In addition, CIA-SSD (Zheng et al. 2021) demonstrates the advantages of using IoU estimation for 3D object detection. By the way, how to further improve the accuracy remains consideration. We utilize the distribution feature efficiently and provide a powerful way for it.

Method

As shown in Figure 4, the proposed DAVNet consists of two parts: region proposal network (RPN) for proposal generation by using the voting strategy on the point cloud, and the proposed DAR module for bounding box refinement by learning the box distribution.

Region Proposal Network

RPN of the proposed 3D detector is VoteNet. It is an anchor-free detector, saving a large amount of memory. The backbone of RPN is PointNet++ (Qi et al. 2017b). 2048 points are sampled from the inputs with furthest point sampling (FPS). They are sent into RPN as input. The backbone in it encodes the point cloud feature for proposal voting. After that, RPN conducts set abstraction to aggregate vote point feature and predicts 3D proposals for DAR module. Due to the limitation in the set abstraction and regression method in VoteNet (mentioned below), the 3D proposals may be inaccurate when facing a complex 3D scene.

Distribution Aware Regression

The 3D proposals in RPN are predicted directly, without consideration to the severe label ambiguity. There is a certain limitation associated with the most common regression method. As shown in Figure 2, it takes the boundary of the target as a certain value, following Dirac delta assumption. However, the target boundary is probably uncertain due to the occlusion or noise in the complex 3D scene. We are not able to confirm the exact boundary or center when the target is occluded (see Figure1 (a)). The same objects may have different ground truth sizes due to the occlusion (see Figure1 (b)). Dirac delta distribution fails to represent such ambiguity. Learning the distribution of it can better handle such a situation and improve the box precision.

Scale adaptive encoder. The set abstraction in VoteNet utilizes a fixed region to aggregate the neighbor point. However, the vote points generated by RPN are probably not convincing for the large object, since its points are far from the center. The fixed area fails to cover all the valid points when their location is inaccurate (see Figure 3 (a)). Simply enlarging the region may contain more noise points in the crowded scene (see Figure 3 (b)). Hence, a unified region brings unbalanced pooling quality among different categories.

To address such concerns, we decouple the pooling quality and the voting quality by introducing SA encoder. We first generate 4 * 4 * 4 grid points by uniformly sampling within each 3D proposal. We generate the grid point feature by interpolation from its 3 neighbour points. The proposal feature is obtained from the concatenated grid point feature through several MLP layers. SA encoder can deal with the

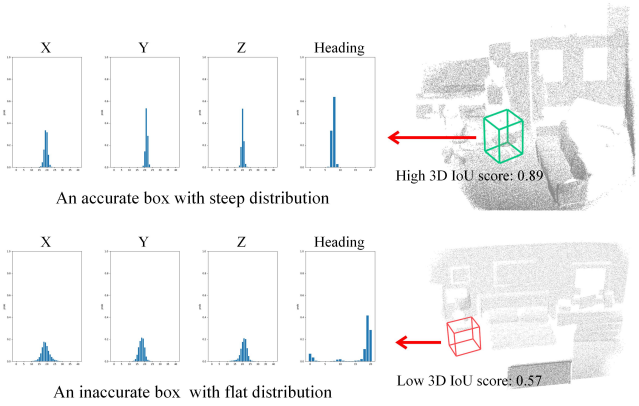

Figure 5: The relationship between the localization quality and the distribution.

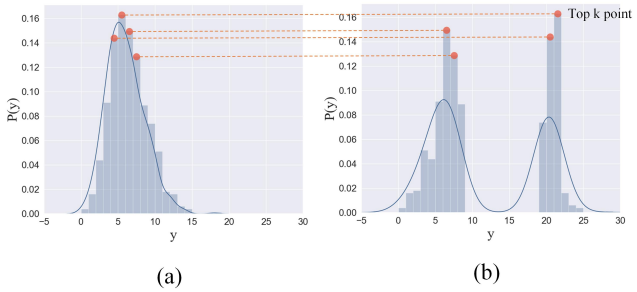

Figure 6: The modification of the distribution encoder. Two distributions result in the same feature by using DGQP in GFocalLoss because they have the same top k $P(y)$. Adding a standard deviation of y_i can avoid such confusion.

targets of different sizes and extract SA features by employing adaptive receptive fields according to the prior proposal from RPN. The RPN feature in Figure 4 is obtained from the set abstraction in RPN. We further combine the SA feature and RPN feature to create distribution-aware representation.

Distribution calculation. Probability distribution $P(x)$ of the regression value y satisfies the formula below

$$y = \int_{-\infty}^{+\infty} P(x)\, x dx \tag{1}$$

$$\tilde{y} = \int_{y_{\min}}^{y_{\max}} P(y)\, y dy \approx \sum_{i=0}^{N_{\text{bins}}} P(y_{\text{i}})\, y_{\text{i}} \tag{2}$$

$$\sum_{i=0}^{N_{\text{bins}}} P(y_{\text{i}}) = 1 \tag{3}$$

where $\tilde{y}$ is an approximation of y, whose range is denoted as $[y_{\min}, y_{\max}]$. We convert the Eq. 1 into a discrete form (Eq. 2) by approximating the integration interval with a set of bins $\{y_i \,|y_{\min} \leq y_i \leq y_{\max}\}, i \in [1, N_{\text{bins}}]$. y represents the offset $\Delta\tilde{\varphi}$ we predicted in Eq. 4. Thus we can calculate the offset by learning its distribution $P(x)$. After generating the aggregation feature, DAR module adopts a stacked convolution layer to predict the probability of each bin mentioned above. We can calculate the output y through the integration in Eq. 2.

For box representation, GFocalLoss predicts the offsets between the four boundaries and the anchor point in 2D object detection, which is not feasible for the 3D task. For a 3D box, similarly regressing the six offsets of the boundary is unfriendly to the network convergence, since all the variables are related to the vote point position. Their regression distribution will change during the training process. Most of 3D detectors commonly predict the residual offset of the center $(\Delta x, \Delta y, \Delta z)$, size$(\Delta w, \Delta h, \Delta l)$ and orientation$(\Delta\theta)$. To a certain extent, the data distribution of $\Delta w, \Delta h, \Delta l$ is fixed during training. We regress the normalized offset $\Delta\tilde{\varphi}$ in Eq. 4.

$$\Delta\tilde{\varphi} = \frac{\Delta\varphi}{\delta_\varphi}, \varphi = x, y, z, w, h, l, \theta \tag{4}$$

where $\Delta\tilde{w}$, $\Delta\tilde{h}$, $\Delta\tilde{l}$ are normalized by $w_{\text{mean}}, h_{\text{mean}}, l_{\text{mean}}$, the mean value of the corresponding category in the whole dataset. δ_z is set to h_{mean} and δ_θ is set to π. Due to the uncertain orientation of the objects, δ_x and δ_y should keep the same. Hence, we normalize $\Delta\tilde{x}, \Delta\tilde{y}$ with the maximum between the w_{mean} and l_{mean}. The normalization can bring a uniform distribution for each regression variable which is beneficial for the network convergence.

The settings of hyperparameters will be discussed in Section 4.4. By taking advantage of the distribution-based regression, we improve the performance of the box localization significantly.

Joint localization quality estimator. The confidence of a target contains two aspects, classification score and location score. VoteNet takes $S_{\text{centerness}}$ as the location score which is supervised by the distance between the vote point and the target center. $S_{\text{centerness}}$ has only focused on the voting quality and ignored the performance on box prediction. 3D IoU, as the final evaluation metric, can represent the location confidence comprehensively. Hence, we carry out IoU prediction to provide accurate box evaluation.

There is a strong correlation between the steepness of the distribution and the IoU score in 3D scene. As illustrated in Figure 5, an accurate bounding box accompanies a sharp distribution. DGQP (Li et al. 2020a) utilizes the top k and mean value of $P(y)$ to encode the steepness of it, which is restricted to unimodal distribution.

The two distributions in Figure 6 may result in the same feature since their top k probabilities are the same. However, the distribution in Figure 6 (a) is sharper than those in Figure 6 (b). DGQP has limitations in representing the multimodal distribution, because it has only used the probability information $(P(y))$. It brings great challenge to the network convergence and makes the training procedure unstable. To avoid such confusion, we add the standard deviation of y among the selected bins. By using the appended features, we

	table	sofa	booksh	chair	desk	dresser	nightst	bed	bathtub	toilet	mAP
AP@0.25											
VoteNet	47.3	64.0	28.8	75.3	22.0	29.8	62.2	83.0	74.4	**90.1**	57.7
MLCVNet	50.4	66.3	**31.9**	75.8	26.5	31.3	61.5	**85.8**	**79.2**	89.1	59.8
H3DNet	-	-	-	-	-	-	-	-	-	-	60.1
our	**52.1**	**66.4**	29.4	**77.1**	**27.5**	**32.2**	**65.0**	84.6	78.9	90.0	**60.3**
AP@0.5											
VoteNet	14.1	40.6	5.7	51.1	4.3	10.4	33.7	44.9	**52.9**	59.4	31.7
H3DNet	-	-	-	-	-	-	-	-	-	-	39.0
our	**28.4**	**52.9**	**8.7**	**60.8**	**9.3**	**20.2**	**49.1**	**57.9**	39.5	**67.3**	**39.4**

Table 1: Performance comparison with the state-of-the-art 3D object detectors on SUN RGB-D validation set. The result of VoteNet on mAP@0.5 is obtained from the model it released. Those of MLCVNet are not released.

	Input	mAP 0.25	mAP 0.5
DSS	Geo+RGB	15.2	6.8
MRCNN 2D-3D	Geo+RGB	17.3	10.5
F-PointNet	Geo+RGB	19.8	10.8
GSPN	Geo+RGB	30.6	17.7
3D-SIS	Geo+RGB	40.2	22.5
3D-SIS	Geo only	25.4	14.6
VoteNet	Geo only	58.6	33.5
MLCVNet	Geo only	64.5	41.4
3D-MPA	Geo only	64.2	49.2
H3DNet	Geo only	**67.2**	48.1
DAVNet(ours)	Geo only	67.1	**50.2**

Table 2: Performance comparison with state-of-the-art methods on ScanNet V2 validation set

can distinguish the two distributions in Figure 6. It improves its robustness while keeping the scale invariance of DGQP.

We transform the aggregation feature into the same shape as the distribution feature and concatenate them for IoU prediction. Benefiting from the distribution statistics, we obtain an IoU with higher precision, denoted as S_{IoU}. Figure 7 shows the result of $S_{\text{centerness}}$ and S_{IoU}, which strongly verifies the accuracy of S_{IoU}. $S_{\text{centerness}}$ is supervised by a hard label, resulting in unsatisfactory performance. The final location scores S_{Box} is obtained by multiplying these two scores (Eq. 5), which benefits in the Non-Maximum Suppression (NMS) stage.

$$S_{\text{Box}} = S_{\text{centerness}} * S_{\text{IoU}} \tag{5}$$

Loss Function

The loss calculation in RPN (l_{RPN}) follows those in VoteNet. The loss in DAR module (l_{DAR}) is defined in Eq. 7.

$$l_{\text{RPN}} = l_{\text{vote}} + 0.1 * l_{\text{cls}} + l_{\text{box}} \tag{6}$$

$$l_{\text{DAR}} = \tilde{l}_{\text{box}} + l_{\text{iou}} + l_{\text{dfl}} \tag{7}$$

where l_{box} and $\tilde{l}_{\text{box}}$ represent the regression loss of the bounding box in RPN and DAR respectively. Since a value can obtained from unlimited distribution, we use Distribution Focal Loss(DFL) (Li et al. 2020b) to supervise the

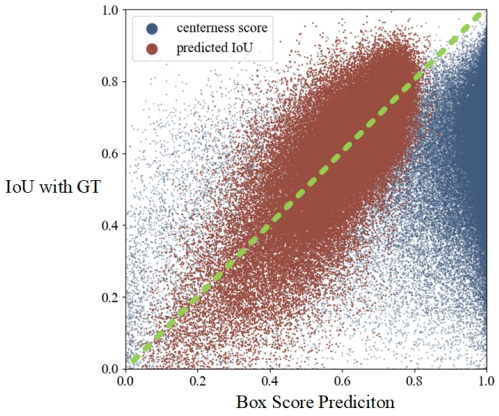

Figure 7: Performance comparison between the $S_{\text{centerness}}$ and S_{IoU}. The points closed to the green line have higher accuracy.

distribution.

$$l_{\text{dfl}} = \sum_{i=0}^{N_{\text{bins}}-1} f_{y_i}^{y_{i+1}} \tag{8}$$

$$f_{y_i}^{y_{i+1}} = -(y_{i+1} - y) * lnP(y_i) - (y - y_i) * lnP(y_{i+1}) \tag{9}$$

where $f_{y_i}^{y_{i+1}}$ is set to 0 when $y \notin [y_i, y_{i+1})$. l_{dfl} guarantees the sharpness and accuracy of the distribution. We use the cross entropy loss for the classification (l_{cls}) and the smooth $L1$ loss for all the regression ($l_{\text{vote}}, l_{\text{box}}, \tilde{l}_{\text{box}}, l_{\text{iou}}$). The total loss is $l_{\text{loss}} = l_{\text{RPN}} + l_{\text{DAR}}$.

Experiment

Datasets

We evaluate our network on ScanNet V2 (Dai et al. 2017) and SUN RGB-D (Song, Lichtenberg, and Xiao 2015). ScanNet has 1513 indoor 3D scenes and 1201 of them are used for training. It provides complete scenes, fused by several RGB-D images. Different from ScanNet, frames on SUN RGB-D are captured from a single view, causing severe incompleteness in point clouds. There are 10335 frames with depth images and over 64000 labeled boxes with orientation in SUN RGB-D.

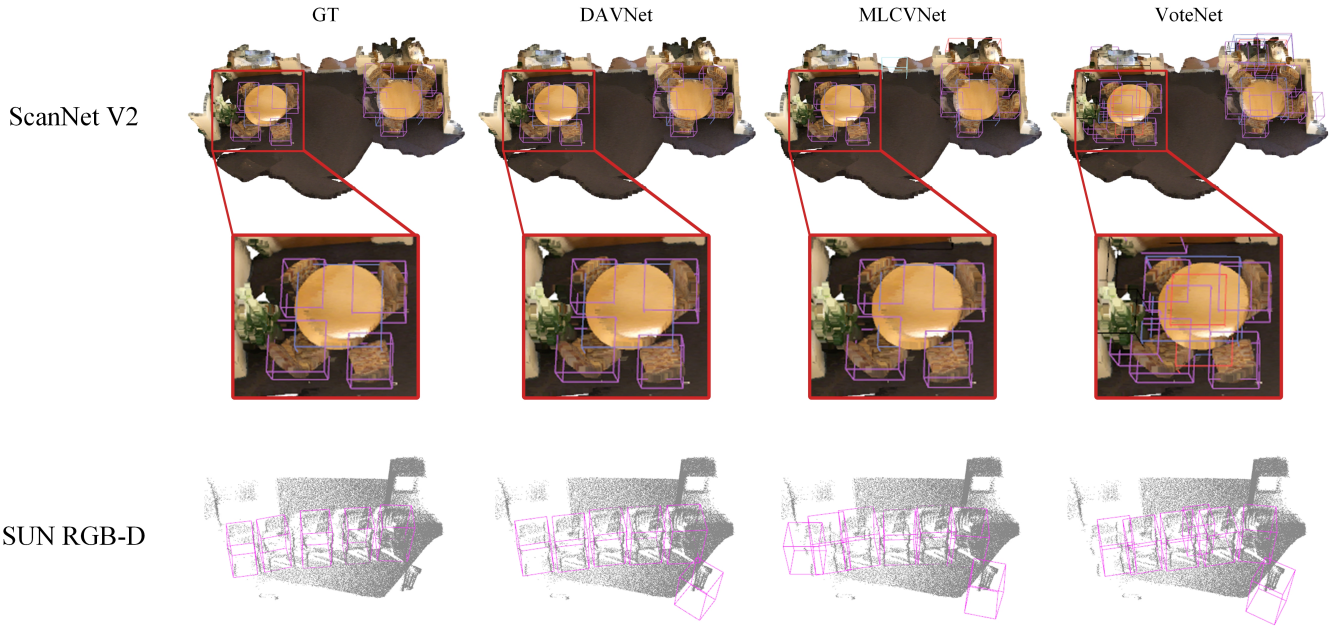

Figure 8: Qualitative comparison results of 3D object detection. DAVNet has better performance on box regression, benefiting from the effectiveness of the distribution representation. It can also suppress false detection with the help of the reliable score predicted by JLQE.

Implementation Details

We use an Adam optimizer to train our model in batch size 8 for both datasets. For ScanNet V2, the network is trained for 180 epochs. The learning rate is initialized as 0.01 and decreased by 10× after 120 and 160 epochs. For SUN RGB-D, we train for 200 epochs with a learning rate initialized as 0.001. It is decreased by 10× after 120, 160, and 180 epochs. We conduct all our training on one GTX1080Ti GPU.

Comparisons with State-of-the-Art Methods

Tabel 1 shows the performance on the validation set of SUN RGB-D. DAVNet reaches 60.3% and 39.4% on mAP@0.25 and mAP@0.5 respectively. The scenes in SUN RGB-D are challenging due to the serious occlusion. The DAR module can conquer these complex scenes. The proposed DAVNet surpasses VoteNet and MLCVNet on several labels especially the large ones such as dresser, sofa, and table. The score on bathtub drops from AP@0.25 to AP@0.5 since the number of this label is not enough for accurate distribution learning. The proposed DAVNet benefits from SA encoder which is adaptive to the object scale. The improvement on SUN RGB-D can strongly verify our effectiveness.

Table 2 shows the comparison on ScanNet V2. Deep sliding shape (Song and Xiao 2016), Mask RCNN (He et al. 2020), F-PointNet (Qi et al. 2018), GSPN (Yi et al. 2019) and 3D SIS (Hou, Dai, and Nießner 2019) utilize the geometry and RGB information, but they still need a proper way to fuse the multi modal information effectively. We also compare DAVNet with 3D-MPA (Engelmann et al. 2020) and H3DNet (Zhang et al. 2020). We achieve 67.1% and 50.2% on mAP@0.25 and mAP@0.5 separately, making 8.5% and 16.7% improvement from VoteNet. We outperform H3DNet by 2.1% on mAP@0.5, remaining a small gap on mAP@0.25. As illustrated in Figure 8, DAVNet achieves better performance in box regression, which confirms the effectiveness of the distribution representation. Meanwhile, with the help of the reliable score provided by JLQE, the number of false-positive results decreases. It brings a higher score in mAP@0.5.

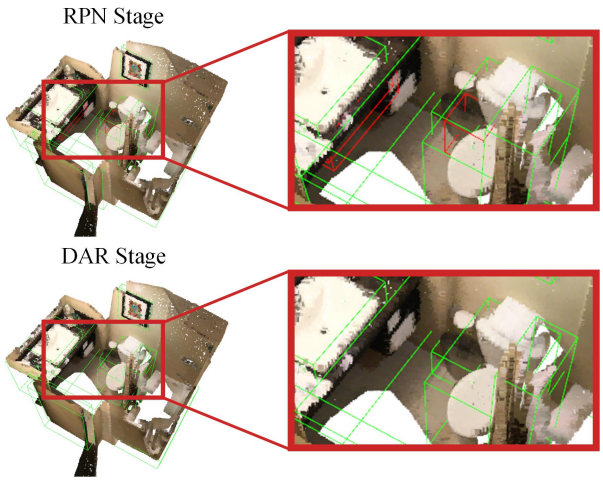

Figure 9: Comparison with the RPN results on mAP@0.25. The green box stands for positive detection while the red one represents false detection.

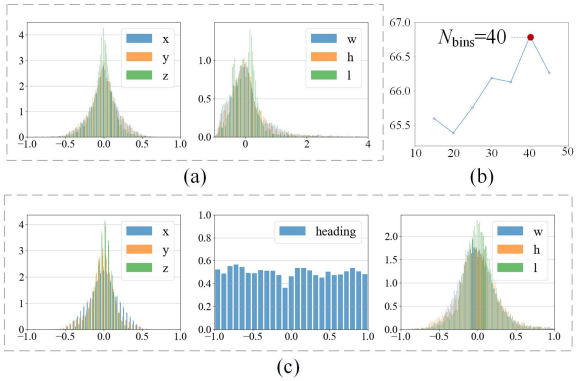

Figure 10: (a) Label distribution on ScanNet V2. (b) Performance of different N_{bins} on mAP@0.25. (c) Label distribution on SUN RGB-D. The distribution of $\Delta x, \Delta y, \Delta z$ is collected by a pre-trained network as they are associated with the voting quality. The others four are constant on the dataset.

SAE	DR	JLQE	Head	ScanNet	SUN RGB-D
			-	64.13	57.79
✓			RPN	64.97	58.50
✓			DAR	65.55	58.77
✓		✓	DAR	66.49	59.02
✓	✓		RPN	65.27	59.68
✓	✓		DAR	66.37	59.75
✓	✓	✓	DAR	67.11	60.32

Table 3: Ablation study for DAR module on validation set(mAP 0.25). SAE stands for SA encoder. DR means using the distribution regression. Head stands for the prediction result in two stages.

Ablation Study

To further validate the proposed method and analyze each individual component, we conduct extensive ablation experiments on these datasets. In this session, we will investigate the effectiveness of the three modules mentioned above. Table 3 displays the results of different combinations among the three modules. The baseline (1st row) is the VoteNet we trained and the 2nd, 3rd, and 4th rows stand for SA encoder module using the original regression method. The 5th, 6th, 7th rows apply distribution regression on the basis of 2nd, 3rd,4th rows respectively. Table 3 demonstrates that each module has made a certain improvement. We further investigate their effectiveness below.

Effect of scale adaptive encoder. The 3rd row shows that the SA encoder made 1.42% and 0.98% improvement from the baseline. Due to its robustness, it can also help fine-turn the RPN, bringing 0.84% and 0.71% improvement (see 2nd row). Hence, we can even take SA encoder as an auxiliary network and drop it during the inference, which can also achieve a certain improvement.

Effect of distribution aware regression. We validate the effect of the DAR module by processing distribution regression after the SA encoder. The 6th row in Table 3 shows that utilizing the distribution helps improve the regression quality, making remarkable improvement. As can be seen from Figure 9, DAR module also limits the false prediction, which validates its robustness in complex scenes.

standard deviation	Scannet mAP0.25
×	65.8
✓	66.2

Table 4: Effect of the standard deviation feature.

location score	features	mAP0.25
$S_{\text{centerness}}$	-	64.13
$S_{\text{centerness}} * S_{IoU}$	RPN feature	64.82
	Aggregation feature	65.4

Table 5: Effect of the S_{IoU} and the aggregation feature.

Effect of joint localization quality estimator. We investigate the effects of JLQE. The 4th, 7th rows in Table 3 show that replacing the location score by S_{Box} in Eq. 5 can bring significant improvement. Figure 7 illustrates the gap between the $S_{\text{centerness}}$ and S_{IoU}, which highlights its accuracy. In addition, Table 4 validates the effectiveness of the appended standard deviation feature in Figure 6.

To eliminate the impact from the other factor, we drop the DAR module and carry out IoU prediction in RPN. The performance are shown in Table 5. The 2nd row demonstrates that S_{IoU} is a powerful guidance for box scoring. Hence, making use of the predicted IoU is more effective.

We further investigate the effect of the distribution features. The 3rd row in Table 5 illustrates that, by using the aggregation feature, the accuracy can be further improved.

Analysis of the hyperparameters. To identify the correlation between the N_{bins} and the precision of the DAR module, we conduct an experiment about different N_{bins} (see Figure 10 (b)). Increasing it can help raise the accuracy, but it may also cost more memory. At last, it is set to 40. The settings of the distribution region $[y_{\min}, y_{\max}]$ follows the collected label distribution in Figure 10. On ScanNet V2, it is set to $[-1, 1]$ for $\Delta x, \Delta y, \Delta z$ and $[-1, 2]$ for $\Delta w, \Delta h, \Delta l$. On SUN RGB-D, they are all set to $[-1, 1]$.

Conclusion

In this paper, we develop a novel 3D object detector named Distribution Aware VoteNet. Our main contribution is that we design DAR module, providing a distribution-based regression method for point cloud 3D object detection. Inside DAR module, we design JLQE to provide a reliable location score. It predicts the IoU of the bounding box by further utilizing the distribution statistics, which is correlative with the localization quality. We also propose SA encoder which adopts adaptive receptive fields to increase the robustness in the pooling stage. The experiment results on ScanNet V2 and SUN RGB-D demonstrate that the proposed DAVNet achieves significant improvement in 3D object detection compared with the previous state-of-the-art methods.

Acknowledgments

This work is supported by National Natural Science Foundation of China (U1913602, 61991412).

References

Ben-Shabat, Y.; Lindenbaum, M.; and Fischer, A. 2017. 3D Point Cloud Classification and Segmentation using 3D Modified Fisher Vector Representation for Convolutional Neural Networks. *CoRR*, abs/1711.08241.

Choi, J.; Chun, D.; Kim, H.; and Lee, H. 2019. Gaussian YOLOv3: An Accurate and Fast Object Detector Using Localization Uncertainty for Autonomous Driving. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 502–511.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T. A.; and Nießner, M. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2432–2443. IEEE Computer Society.

Engelmann, F.; Bokeloh, M.; Fathi, A.; Leibe, B.; and Nießner, M. 2020. 3D-MPA: Multi-Proposal Aggregation for 3D Semantic Instance Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 9028–9037. Computer Vision Foundation / IEEE.

Girshick, R. 2015. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 1440–1448.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. B. 2020. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2): 386–397.

He, Y.; Zhu, C.; Wang, J.; Savvides, M.; and Zhang, X. 2019. Bounding Box Regression With Uncertainty for Accurate Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2888–2897.

Hou, J.; Dai, A.; and Nießner, M. 2019. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 4421–4430.

Jiang, B.; Luo, R.; Mao, J.; Xiao, T.; and Jiang, Y. 2018. Acquisition of Localization Confidence for Accurate Object Detection. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, 816–832.

Li, X.; Wang, W.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2020a. Generalized Focal Loss V2: Learning Reliable Localization Quality Estimation for Dense Object Detection. *CoRR*, abs/2011.12885.

Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2020b. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Li, Y.; Dai, A.; Guibas, L. J.; and Nießner, M. 2015. Database-Assisted Object Retrieval for Real-Time 3D Reconstruction. *Comput. Graph. Forum*, 34(2): 435–446.

Maturana, D.; and Scherer, S. A. 2015. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015, Hamburg, Germany, September 28 - October 2, 2015*, 922–928.

Nan, L.; Xie, K.; and Sharf, A. 2012. A *search-classify* approach for cluttered indoor scene understanding. *ACM Trans. Graph.*, 31(6): 137:1–137:10.

Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep Hough Voting for 3D Object Detection in Point Clouds. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 9276–9285.

Qi, C. R.; Liu, W.; Wu, C.; Su, H.; and Guibas, L. J. 2018. Frustum PointNets for 3D Object Detection From RGB-D Data. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 918–927.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 77–85.

Qi, C. R.; Su, H.; Nießner, M.; Dai, A.; Yan, M.; and Guibas, L. J. 2016. Volumetric and Multi-view CNNs for Object Classification on 3D Data. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 5648–5656.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5099–5108.

Reading, C.; Harakeh, A.; Chae, J.; and Waslander, S. L. 2021. Categorical Depth Distribution Network for Monocular 3D Object Detection. *CoRR*, abs/2103.01100.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 91–99.

Shi, S.; Wang, X.; and Li, H. 2019. PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 770–779.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 567–576. IEEE Computer Society.

Song, S.; and Xiao, J. 2016. Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 808–816.

Su, H.; Maji, S.; Kalogerakis, E.; and Learned-Miller, E. G. 2015. Multi-view Convolutional Neural Networks for 3D Shape Recognition. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 945–953.

Tychsen-Smith, L.; and Petersson, L. 2018. Improving Object Localization With Fitness NMS and Bounded IoU Loss. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 6877–6885.

Wang, H.; Cong, Y.; Litany, O.; Gao, Y.; and Guibas, L. J. 2020. 3DIoUMatch: Leveraging IoU Prediction for Semi-Supervised 3D Object Detection. *CoRR*, abs/2012.04355.

Wei, X.; Yu, R.; and Sun, J. 2020. View-GCN: View-Based Graph Convolutional Network for 3D Shape Analysis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 1847–1856.

Wu, S.; Li, X.; and Wang, X. 2020. IoU-aware single-stage object detector for accurate localization. *Image Vis. Comput.*, 97: 103911.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3D ShapeNets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 1912–1920.

Xie, Q.; Lai, Y.; Wu, J.; Wang, Z.; Zhang, Y.; Xu, K.; and Wang, J. 2020. MLCVNet: Multi-Level Context VoteNet for 3D Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 10444–10453.

Yi, L.; Zhao, W.; Wang, H.; Sung, M.; and Guibas, L. J. 2019. GSPN: Generative Shape Proposal Network for 3D Instance Segmentation in Point Cloud. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 3947–3956.

Zhang, Z.; Sun, B.; Yang, H.; and Huang, Q. 2020. H3DNet: 3D Object Detection Using Hybrid Geometric Primitives. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII*, volume 12357 of *Lecture Notes in Computer Science*, 311–329. Springer.

Zheng, W.; Tang, W.; Chen, S.; Jiang, L.; and Fu, C. 2021. CIA-SSD: Confident IoU-Aware Single-Stage Object Detector From Point Cloud. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 3555–3562.