Supplementary Materials for "VISTA: Boosting 3D Object Detection via Dual Cross-VIew SpaTial Attention"

A. Implementation Details

Attention Module Directly sequentializing the original feature maps into feature sequences will construct the intermediate attention maps with huge memory occupancy, which leads to unaffordable GPU memory costs. Considering efficiency and simplicity, the original feature maps are downsampled by average pooling before being passed into attention module, then the outputs of attention module are mapped back to original sizes via inverse mapping according to the pooling field. In practice, the kernel size of the average pooling is set to [4,4] and [4,1] for BEV and RV, respectively.

Voxelization We voxelize the point clouds according to the x,y,z axes. All ablation studies are conducted in low voxelization resolution of [0.1,0.1,0.1]m according to the x,y,z axes. To benchmark the results of our proposed VISTA-OHS on the nuScenes dataset, we follow the OHS [1] to tune up the voxelization resolution to [0.08, 0.08, 0.08]m. In terms of the Waymo Open Dataset, following the official configurations provided by CenterPoint, we keep the low resolution unchanged.

Training We follow the CBGS [5] to train the proposed VISTA using Adam [3] optimizer scheduled by one-cycle learning rate policy [2]. For Adam optimizer, the weight decay is set to 1e-2. And for the one-cycle learning rate policy, we set the max learning rate as 1e-3 for nuScenes and 3e-3 for Waymo Open Dataset, and the momentum is ranging from 0.95 to 0.85. We train the proposed VISTA on 4 RTX3090 GPUs for 20 epochs with batch size 16 on the nuScenes dataset, and for 36 epochs with batch size 16 on the Waymo Open Dataset. During training, the proposed attention variance constrain is applied on both the regression and classification branches of the decoupling architecture.

Code Our implementations are based on the open-sourced code released by CBGS [5] [1] and CenterPoint [4] [2]. Code and experimental configurations will be released upon the acceptance of the paper.

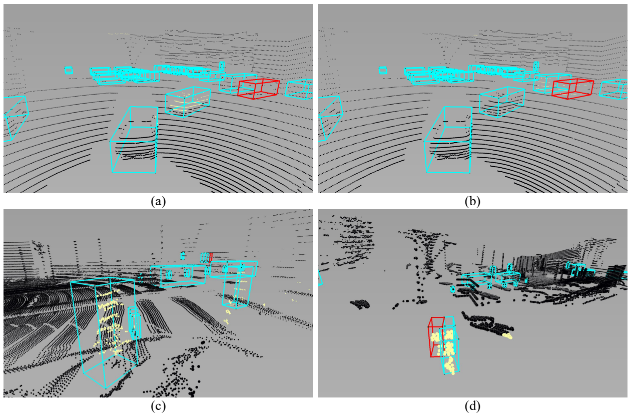

Figure 1. The attention weights learned by the classification and regression branches of the decoupled attention modeling. We choose two samples to present, each of which is demonstrated in one row. The left column illustrates the attention maps for classification tasks, the right column shows the regression one. The query boxes are painted in red, the brighter the points, the larger the attention weights.

B. Extra Analysis of Decoupling Design

In this section, we present the attention weights in terms of regression and classification tasks in the Figure 1, which are learned in the decoupled attention modeling. Given the area containing a query bounding box from target view (BEV) to query the source view (RV), we get the corresponding cross-view attention weights for each pillar in the above area, and map the weight back to the origin point set for visualization. We observe that, different supervised signals lead to different attention weights. For classification task, the attention module tends to focus on the other objects in the whole scenes to enrich the semantic information contained in the fused features, as shown in the (a) and (c). To understand the geometric properties (e.g. scale, translation) of the query objects, the attention modeling for regression task instead, paying its attention to the local regions in which the query objects are, as we demonstrated in the (b)

[1] https://github.com/poodarchu/Det3D
[2] https://github.com/tianweiy/CenterPoint

Avg	Linear Atten	Conv Atten	Var Cons	Decouple	mAP	car	truck	cons.	bus	trailer	barrier	motorcycle	bicycle	pedestrian	traffic cone
					59.5	84.2	56.5	19.7	65.6	36.0	67.2	63.7	47.1	83.5	68.4
✓					59.2	84.0	53.8	19.7	65.5	36.6	64.5	67.4	48.2	83.1	67.1
	✓				58.7	84.6	52.4	18.4	65.5	36.0	64.6	66.6	44.0	83.7	68.4
		✓			60.0	84.4	54.1	20.4	67.2	36.7	64.4	66.6	50.6	83.5	69.6
		✓	✓		60.4	84.7	55.8	19.8	67.1	36.2	68.6	67.3	50.6	83.9	69.1
		✓	✓	✓	**60.8**	**84.8**	**57.2**	**20.5**	**67.6**	**36.8**	**69.0**	**67.7**	**50.7**	**84.1**	**69.7**

Table 1. The detailed ablation studies on the validation set of the nuScenes dataset. “cons.” refers to construction vehicle

Method	mAP	car	truck	cons.	bus	trailer	barrier	motorcycle	bicycle	pedestrian	traffic cone
Decouple+Cls Var	60.6	84.8	57.0	20.4	66.9	36.2	68.3	65.7	49.1	83.5	69.5
Decouple+Reg Var	60.5	84.7	55.4	19.3	66.4	36.0	68.6	67.1	50.2	84.0	69.3
Decouple+Both Var	**60.8**	**84.8**	**57.2**	**20.5**	**67.6**	**36.8**	**69.0**	**67.7**	**50.7**	**84.1**	**69.7**

Table 2. Ablation studies of the attention variance constrain being applied on different tasks. “Cls” and “Reg” stand for the classification and regression, respectively. “cons.” refers to construction vehicle. The ablation studies are conducted on the validation set of the nuScenes dataset.

and (d). The different preferences of the individual attention modeling on the regions of interest further demonstrate the efficacy of our decoupling design.

During training, we apply the proposed attention variance constrain on both the classification and regression attention weights of the decoupled attention modeling. To further verify the different impacts that the classification and regression task will have on the network, we apply the attention variance constrain on different attention weights. The performances are demonstrated in the Table 2. We observe that, when apply the attention variance constrain on the attention weights of the classification task, the network yields better performances on the large objects (e.g. truck, construction vehicle). We argue that such performance gains are mainly due to the enriched semantic features. Since the most parts of the large objects in point cloud representation are empty, aggregating the corresponding dense cross-view features from the other objects is beneficial for the network to infer the categories of the objects. When it comes to the small objects (e.g. barrier, motorcycle, pedestrian), the small sizes of the objects make the network easier to consider the local context to understand the geometric properties, therefore, the regression task is better at handling the small objects when being applied the attention constrain. After adopting the proposed attention variance constrain in both classification and regression, the network benefits from the advantages on the large also the small objects, and yields the best performances, as shown in the last row of Table 2.

Nevertheless, the decoupling design definitely brings extra parameters. To further clarify that the performance gains come from the proposed decoupling structure, we conduct an experiment that uses a single attention and adds several convolutional layers to the detection head (Deeper Head); the setting keeps the number of parameters roughly the same. Validation results in Table 3 show that the alternative setting of replacing the decoupling does not bring benefits, which further verifies the efficacy of our proposed decoupling design.

	Deeper Head	Decoupling
mAP Gains	60.40→60.45 (+0.05)	60.40→60.81 (**+0.41**)

Table 3. The mAP gains on nuScenes Validation Set

C. Extra Analysis of Variance Constraint

We apply the variance constraint on the positive samples during the training phase, which may form an “upweight” of the positive samples. Hence, the stated performance gains benefit from the proposed variance constraint could be attributed to such an “upweight” training. To clarify, we disable the variance loss and run a set of experiments that either scale up or scale down the background predictions, as shown in Table 4. Table 4 shows that our variance loss is not equivalent to scaling of labels, thus verifying its efficacy.

	Scaling down (x0.5)	Scaling up (x2)	Ours
mAP	59.6	60.4	**60.8**

Table 4. The mAP on nuScenes Validation Set

D. Detailed Ablations

In this section we extend the ablation studies to each category in Table 1 to present the category-wise performances. The ablation studies are conducted on the validation set of nuScenes dataset.

References

[1] Qi Chen, Lin Sun, Zhixin Wang, Kui Jia, and Alan Yuille. Object as hotspots: An anchor-free 3d object detection approach via firing of hotspots. In *European Conference on Computer Vision*, 2020. 1

[2] Sylvain Gugger. The 1cycle policy. *https://sgugger.github.io/the-1cycle-policy.html*, 2018. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1

[5] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 1