VISTA: Boosting 3D Object Detection via Dual Cross-VIew SpaTial Attention

Shengheng Deng[1,*], Zhihao Liang[1,3,*], Lin Sun[2] and Kui Jia[1,4,†]
[1]South China University of Technology, [2]Magic Leap, Sunnyvale, CA
[3]DexForce Technology Co., Ltd., [4]Peng Cheng Laboratory
{eedsh, eezhihaoliang}@mail.scut.edu.cn, kuijia@scut.edu.cn, lsun@magicleap.com

Abstract

Detecting objects from LiDAR point clouds is of tremendous significance in autonomous driving. In spite of good progress, accurate and reliable 3D detection is yet to be achieved due to the sparsity and irregularity of LiDAR point clouds. Among existing strategies, multi-view methods have shown great promise by leveraging the more comprehensive information from both bird's eye view (BEV) and range view (RV). These multi-view methods either refine the proposals predicted from single view via fused features, or fuse the features without considering the global spatial context; their performance is limited consequently. In this paper, we propose to adaptively fuse multi-view features in a global spatial context via Dual Cross-VIew SpaTial Attention (VISTA). The proposed VISTA is a novel plug-and-play fusion module, wherein the multi-layer perceptron widely adopted in standard attention modules is replaced with a convolutional one. Thanks to the learned attention mechanism, VISTA can produce fused features of high quality for prediction of proposals. We decouple the classification and regression tasks in VISTA, and an additional constraint of attention variance is applied that enables the attention module to focus on specific targets instead of generic points. We conduct thorough experiments on the benchmarks of nuScenes and Waymo; results confirm the efficacy of our designs. At the time of submission, our method achieves 63.0% in overall mAP and 69.8% in NDS on the nuScenes benchmark, outperforming all published methods by up to 24% in safety-crucial categories such as cyclist.

1. Introduction

LiDAR is one of the prominent sensors which is widely used in autonomous driving to provide precise 3D information of the objects. Therefore, LiDAR based 3D object detection has attracted a lot of attention. Many 3D object detection algorithms [19, 29, 32] apply the convolutional neural networks to 3D point clouds by voxelizing the unordered and irregular point clouds into volumetric grids. Nevertheless, the 3D convolutional operator is computationally inefficient and memory-consuming. To mitigate these issues, a line of works [6, 28] utilize sparse 3D convolutions [9–11] in the network backbones to extract features. As illustrated in the Figure 1 (a), these works project the 3D feature maps into the bird's eye view (BEV) or range view (RV), and the object proposals are produced from these 2D feature maps using proposal methods.

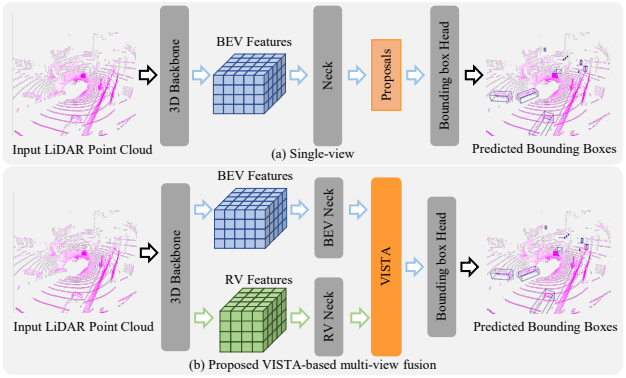

Figure 1. Comparison between the single-view detection and the proposed VISTA-based multi-view fusion. (a) shows the single-view detection pipeline. (b) illustrates the proposed VISTA-based multi-view fusion. The BEV and RV features which are extracted by a shared 3D backbone are passed into individual necks, and pass through the VISTA to output high quality fused features.

Different views have their own advantages and drawbacks to consider. In BEV, objects do not overlap with each other and the size of each object is invariant to the distance from the ego-vehicle. RV is the native representation of LiDAR point clouds, therefore, it can produce compact and dense features. However, projection would inevitably impair the integrity of spatial information conveyed in the 3D space no matter which of BEV or RV is chosen. For example, due to the self-occlusion and the

* indicates equal contribution.
†Correspondence to Kui Jia <kuijia@scut.edu.cn>.

characteristic of LiDAR data generation, the BEV representation is extremely sparse and it consolidates the height information of 3D point clouds, occlusion and variation in object size would be severer in RV since it loses depth information. Obviously, learning jointly from multiple views, a.k.a multi-view fusion provides us a solution to accurate 3D object detection. Some of the previous multi-view fusion algorithms [7, 17] produce the proposals from a single view and utilize the multi-view features to refine proposals. Performances of such algorithms highly depend on the quality of the produced proposals; however, proposals generated from a single view make no use of all the available information, possibly leading to suboptimal solutions. Other works [5, 31] fuse the multi-view features according to the coordinate projection between different views. Accuracy of such fusion methods relies on the complementary information provided in the corresponding region of the other view; yet the occlusion effect is inevitable, inducing low-quality multi-view feature fusion.

To boost the performances of 3D object detection, in this paper, given learned 3D feature maps from both BEV and RV, we propose to produce high quality fused multi-view features from the global spatial context via Dual Cross-VIew SpaTial Attention (VISTA) for proposal prediction, as demonstrated in Figure 1 (b). The proposed VISTA utilizes the attention mechanism originated in the transformer which is successfully applied to various research context (e.g. natural language processing, 2D computer vision). Compared with direct fusion via coordinate projections, the inbuilt attention mechanism in VISTA exploits the global information and adaptively models all the pairwise correlations across views by treating features of individual views as sequences of feature elements. To model the cross-view correlations comprehensively, the local context in both views must be taken into account, thus we replace the MLPs in the conventional attention module with the convolutional operators, of which we show the effectiveness in the Section 6. Nevertheless, learning the correlations across views is still challenging, as shown in Section 6. Directly adopting the attention mechanism for multi-view fusion brings little gains and thus, we argue that it is mainly due to the characteristic of the task 3D object detection itself.

Generally, the 3D object detection task could be divided into two sub-tasks: classification and regression. As elaborated in [5, 22], the 3D object detector faces many challenges when detecting objects in the whole 3D scenes, such as occlusion, background noise and the scarce texture information of point cloud. In consequence, the attention correlations are difficult to learn and the attention module tends to learn the mean of the whole scene, which is unexpected as the attention module is designed for paying attention to regions of interest. Therefore, we explicitly constrain the variance of the attention maps learned by the attention mechanism, which guides the attention module to be aware of the meaningful regions in the complex 3D outdoor scenes. Moreover, different learning targets for classification and regression determine the different expectations of the learned queries and keys in the attention module. The various regression targets (e.g. scale, translate) across different objects expect the queries and keys to be aware of the characteristic of the objects. The classification task instead, pushes the network to understand the common properties of the object classes. Inevitably, sharing the same attention modeling will bring conflicts into the training of these two tasks. Furthermore, on one hand, due to the loss of texture information, it is difficult for neural networks to extract semantic features from point clouds. On the other hand, the neural networks can easily learn the geometric property of objects from point clouds. As a result, during training, a dilemma that the classification being dominant by the regression is aroused. To tackle these challenges, we decouple these two tasks in the proposed VISTA to learn to aggregate different cues in terms of different tasks.

Our proposed VISTA is a plug-and-play module and can be adopted to the recent advanced target assign strategies. We test our proposed VISTA-based multi-view fusion on different target assign algorithms on the benchmark datasets of nuScenes [2] and Waymo [25]. Ablation studies on their validation sets confirm our conjecture. Thanks to the high quality fused features produced by the proposed VISTA, our proposed method outperforms all the published algorithms. At the time of submission, our final results achieve 63.0% in overall mAP and 69.8% in NDS on nuScenes leaderboard. On Waymo Open Dataset, we achieve 74.0%, 72.5%, and 71.6% level 2 mAPH on vehicle, pedestrian and cyclist. We summarize our main contributions as follows.

- We propose a novel plug-and-play fusion module Dual Cross-VIew SpaTial Attention (VISTA) to produce well-fused multi-view features to boost the performances of 3D object detector. Our proposed VISTA replaces the MLPs with convolutional operators, which is capable of better handling the local cues for attention modeling.
- We decouple the regression and classification tasks in the VISTA to leverage individual attention modeling to balance the learning of these two tasks. We apply the attention variance constraint to VISTA during training phase, which facilitate the attention learning and empower the network to attend to the regions of interest.
- We conduct thorough experiments on the benchmark datasets of nuScenes and Waymo. Our proposed VISTA-based multi-view fusion can be adopted in various advanced target assign strategies, easily boost the original algorithms and achieve state-of-the-art performances on the benchmark datasets. Specifically, our

proposed method outperforms the second best methods by 4.5% in overall performance, and up to 24% on the safety-crucial object categories like cyclist.

2. Related Works

2.1. Single-View 3D Detection

BEV-Based 3D Detection Most of the voxel-based 3D detection algorithms detect objects on BEV. [19] projects the point clouds into BEV pillars, and feeds the BEV pillars into 2D CNNs. Such a projection inevitably induces 3D spatial information loss. Recent works [6, 28, 30, 32] mitigate this issue by utilizing 3D CNNs to directly operate on 3D point clouds or processed 3D point clouds, e.g. voxels, and then project 3D feature maps into BEV, and finally detect objects on projected BEV features.

RV-Based 3D Detection Few works [22] detect objects from RV. RV representations provide compact features. However, as mentioned in [22], the RV detectors need more training data, and there exist great challenges posed by occlusion and variant object scales with range.

Nevertheless, both projections will damage the 3D spatial information integrity. We believe that a comprehensive 3D detection framework needs to learn from both views, and the performance is decided by the fused features with the complementary information from both views.

2.2. Multi-View 3D Detection

A line of works [4,18] realize multi-view fusion either by aggregating features to refine proposals or fusing features in the region constrained by the spatial projection. [7, 17] fuse the ROI features from point cloud and camera image for proposals refinement. Instead of fusing multi-view features at the ROI level, [31] fuses point-wise features from BEV and RV. Different from previous works, CVCNet [5] proposes a hybrid voxelization method to unify the benefits from both views, and utilizes hough transform to restrict the consistency between classification results from both views. However, the CVCNet does not utilize the multi-view features to produce proposals directly, thus fails to make full use of the fused features to do the 3D detection.

All of these works fuse features in limited regions or do not exploit the fused feature for the 3D detection. To leverage the multi-view features from the global spatial context, our proposed VISTA considers the interactions among features from different views in the entire scene.

2.3. Attention in Transformer

Thanks to the capability of effectively capturing long-range dependencies of features in the input feature sequences, transformer [26] have been widely transferred into computer vision task [3, 8, 13, 14, 21, 24, 34]. The core component in transformer is the self-attention module, which explicitly models the pair-wise correlations among the input feature sequences. ViT [8] divides the images into patches for attention construction to realize image classification. PCT [13] modifies the attention module into a discrete Laplacian operator for point cloud classification. CT3D [24] reweights the proposal features via channel-wise attention for bounding box refinement.

Unfortunately, due to the characteristic of 3D object detection task itself and the inherent property of outdoor 3D point clouds, the existing attention module fails to focus on the regions of the interest in the scenes. Moreover, the network training will be easily dominant by the regression task. The proposed VISTA, instead, addresses the above issues via the decoupled attention modeling and the designed training constraint, thus is able to produce high quality fused multi-view features for 3D object detection.

3. Overview

Given LiDAR point clouds from one scene, the task of 3D object detection is to accurately predict the categories of objects and output the oriented bounding boxes enclosed the objects in the scene. As illustrated in Figure 1 (a), in most 3D object detectors, the learned 3D feature grids will be collapsed into 2D feature maps of BEV or RV, followed by a 2D bounding box head. We term these methods as single-view detector. Inevitably, the 'collapse' operation will impair the spatial integrity, which will lead to inferior bounding box prediction. To compensate the loss, various multi-view fusion methods have been proposed. As elaborated in Section 1, hindered by the quality of proposals or the limited region of fusion, all of these methods cannot fuse multi-view features comprehensively. We argue that one should consider the global context during multi-view fusion to better leverage the complementary information conveyed in two views. Therefore, we propose Dual Cross-VIew SpaTial Attention (VISTA) module which decouples the classification and regression tasks in 3D object detection, and is trained under designed attention constraint.

We follow the general fusion pipeline that is widely used in existing fusion algorithms. The overall architecture is demonstrated in Figure 1 (b). We adopt the widely used sparse 3D ResNet [5] as our shared 3D backbone to produce 3D feature maps $F_{3d} \in \mathbb{R}^{B\times C\times H\times W\times D}$, where the B is the batch size, C is the feature dimension, H, W, and D are the size of F_{3d} corresponding to width, height, and depth axes, respectively. For BEV and RV, F_{3d} is collapsed into two 2D feature maps $F_{bev} \in \mathbb{R}^{B\times (C\times W)\times H\times D}$ and $F_{rv} \in \mathbb{R}^{B\times (C\times D)\times H\times W}$. F_{bev} and F_{rv} are fed into individual 2D feature extractor, namely 2D neck. As most recent state-of-the-art detectors [6, 30], we adopt UNet-like architecture as our 2D neck which contains several convolutional layers and each of which is followed by a normalization and activation function. After 2D necks, the VISTA

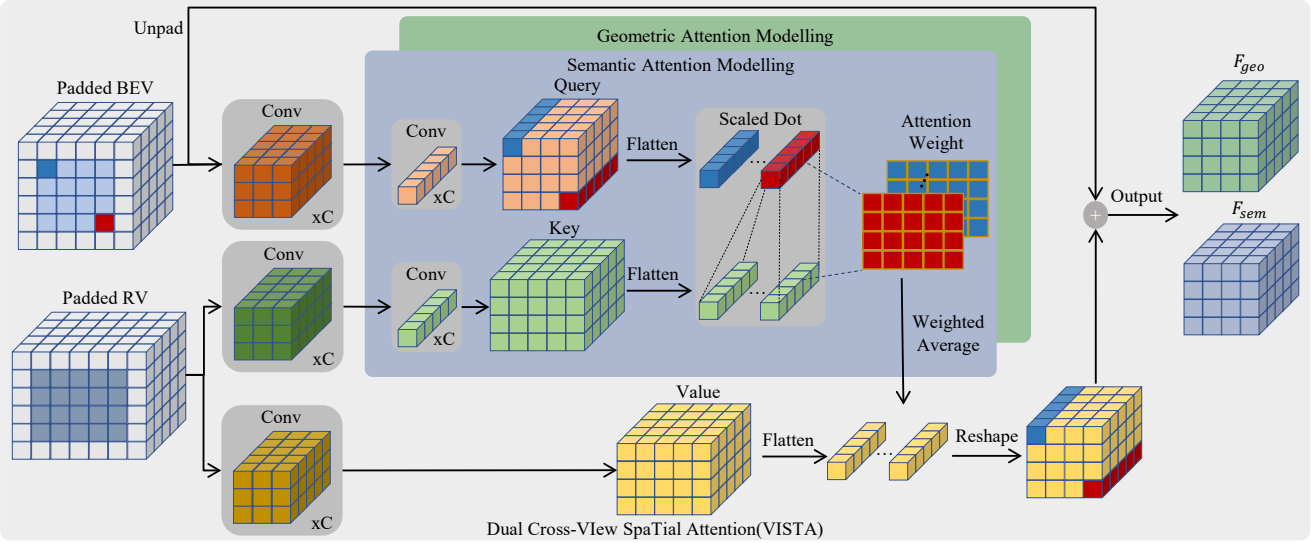

Figure 2. The architecture of the proposed VISTA.

takes F_{bev} and F_{rv} as inputs, and outputs the fused multi-view features for producing the detection results.

4. Dual Cross-VIew SpaTial Attention

For most voxel-based 3D detectors that densely produce pillar-wise proposals, generating nutrient feature maps guarantees the detection qualities empirically. In the case of multi-view 3D object detection where the proposal comes from the fused feature maps, an overall consideration of global spatial context during fusion is required. To this end, we seek to utilize the ability of capturing global dependencies of attention module for multi-view fusion, namely cross-view spatial attention. Before considering the global context, the cross-view spatial attention module needs to aggregate the local cues for constructing the correlations between different views, as we shown in Section 6. Therefore, we are motivated to propose VISTA, wherein the standard attention module based on multi-layer perceptron is replaced with a convolutional one. However, learning the attention in the complicated 3D scenes is difficult. To adopt the cross-view attention for multi-view fusion, we further decouple the classification and regression tasks in the VISTA, and apply the proposed attention constraint to boost the learning process of attention mechanism.

In this section, we will first introduce the overall architecture of the proposed Dual Cross-VIew SpaTial Attention (VISTA) in detail, then we elaborate the decoupling design and the attention constraints for the proposed VISTA.

4.1. Overall Architecture

As shown in Figure 2, VISTA takes feature sequences from two different views as inputs and models the cross-view correlations among the multi-view features. Unlike the vanilla attention module that uses linear projections to transform input feature sequences, VISTA projects the input feature sequences $\mathcal{X}_1 \in \mathbb{R}^{n \times d_f}$ and $\mathcal{X}_2 \in \mathbb{R}^{m \times d_f}$ into queries $\mathcal{Q} \in \mathbb{R}^{n \times d_q}$ and keys $\mathcal{K} \in \mathbb{R}^{m \times d_q}$ (values $\mathcal{V} \in \mathbb{R}^{m \times d_v}$) via convolutional operators of 3×3 kernels, where d_q and d_v are the feature dimensions of queries (keys) and values. To decouple the classification and regression tasks, $\mathcal{Q}$ and $\mathcal{K}$ are further projected into $\mathcal{Q}_i, \mathcal{K}_i, i \in \{sem, geo\}$ via individual MLP (implemented as 1D convolution). To compute the weighted sum of the values $\mathcal{V}$ as the cross-view output $\mathcal{F}_i \in \mathbb{R}^{n \times d_v}$, the scaled dot-product is applied to obtain the cross-view attention weight $\mathcal{A}_i \in [0, 1]^{n \times m}$:

$$\mathcal{A}_i = softmax(\frac{\mathcal{Q}_i \mathcal{K}_i^\top}{\sqrt{d_q}}), i \in \{sem, geo\} \quad (1)$$

and the output will be $\mathcal{F}_i = \mathcal{A}_i \mathcal{V}$. The output $\mathcal{F}_i$ will be fed into individual Feed Forward Network $\mathcal{FFN}_i$ (FFN) to obtain the final results. We adopt the architecture widely used in previous works [3,26] as our FFN to ensure the non-linearity and diversity. The proposed VISTA is an one-stage method that directly generates proposals based on the features fused across views; such a design can leverage more information for accurate and efficient 3D detection.

4.2. Decoupling Classification and Regression

The VISTA decouples the classification and regression tasks. After the shared convolutional operators, the queries and keys are further processed by individual linear projection to produce $\mathcal{Q}_i$ and $\mathcal{K}_i$, which will then participate in different attention modelling in terms of semantic information or geometric information. The motivation of such decoupling is the different impacts that the supervised signal of classification and regression will have on the training.

Given a query object in the scene, for classification, the attention module needs to aggregate the semantic cues from

the objects in the global context to enrich the semantic information conveyed in the fused features. Such targets require the learned queries and keys to be aware of the commonalities among different objects of the same category, for the sense that the objects of the same category should match each other in regard to the semantic meaning. However, regression task can not take the same set of queries and keys since different objects have their own geometric characteristic (e.g. translation, scale, velocity, etc), the regression features should be diverse across different objects. Therefore, sharing the same queries and keys will induce conflicts to the attention learning during the joint training of classification and regression.

Furthermore, no matter single view or multi-view, the classification and regression results are all predicted from the same feature maps in the conventional voxel-based 3D detectors. However, due to the inherent property of the 3D scene, there exists inevitable occlusion and loss of texture information in the 3D point cloud, thus the 3D detector is difficult to extract the semantic features, leading great challenges to the learning of classification. On the contrary, the rich geometric information conveyed by the 3D point cloud relaxes the burden of understanding the geometric property of objects, which is the basis of learning regression task. As a result, during network training, there comes the imbalanced learning between the classification and regression, where the learning of classification is dominant by the regression. Such an imbalanced learning is a common issue in the 3D object detection involving classification and regression based on 3D point cloud, which will have negative impacts on the detection performances. To be concrete, the 3D detector will not be robust across different object categories (e.g. truck and bus) that have similar geometric features, as we shown in the Section 6.4.

To mitigate the issues described above, we are motivated to individually set up attention modelling for semantic and geometric information respectively. The output of the attention module are the F_{sem} and F_{geo} based on the constructed semantic and geometric attention weight. The supervision of classification and regression are applied on the F_{sem} and F_{geo} respectively, which guarantees the effective learning of the corresponding tasks.

4.3. Attention Constraint

The proposed VISTA faces many challenges when learns to model the cross-view correlation from the global context. The 3D scenes contain plenty of background points (approximately up to 95%), and only a small portion are points of interest that contributes to the detection results. During the training of the cross-view attention, the massive background points will bring unexpected noise to the attention module. Moreover, the occlusion effect in the complex 3D scenes brings inevitable distortion to the attention learning. Consequently, the attention module tends to attend to the irrelevant regions, as shown in the Section 6.4. The extreme case of the poor learning of attention is the global average pooling (GAP) operation, as we demonstrated in the Section 6, without any explicit supervisions, directly adopting the attention module for multi-view fusion yields performance similar to the GAP, which indicates that the attention module cannot model the cross-view correlations well.

To empower the attention module to focus at specific targets rather than generic points, we propose to apply a constraint on the variance of the learned attention weights. Thanks to the proposed constraint, we enable the network to have the ability to learn where to attend. By combining the attention variance constraint with the conventional classification and regression supervised signal, the attention module focuses at the meaningful targets in the scenes, as we shown in the Section 6.4, thus producing high quality fused features. We formulate the proposed constraint as an auxiliary loss during training. For simplicity, we ignore the batch dimension, given a learned attention weight $\mathcal{A} \in \mathbb{R}^{N_{bev} \times N_{rv}}$ where the N_{bev} and N_{rv} are the number of pillars in BEV and RV respectively, the set of the scale and center location of ground-truth bounding boxes in x-y planes $\mathbb{B} = \{b_q | b_q = (w_q, h_q, x_q, y_q), q = 1, ..., N_{box}\}$, where the N_{box} is the number of boxes in the scene. For each pillar in BEV, we calculate the real-world coordinates of its center based on the voxel size and obtain the set $\mathbb{C} = \{c_j | c_j = (x_j, y_j), j = 1, ..., N_{bev}\}$. The attention weights of each ground-truth bounding box are obtained by:

$$\mathcal{A}_q = \mathcal{A}[p, :], \text{s.t.} \begin{cases} x_q - w_q/2 \leq x_p \leq x_q + w_q/2 \\ y_q - h_q/2 \leq y_p \leq y_q + h_q/2 \end{cases} \tag{2}$$

Then we formulate the variance constraint for all ground-truth bounding boxes as follows:

$$\mathcal{L}_{var} = -\frac{1}{N_{box}} \sum_{q}^{N_{box}} \frac{1}{N_q} \sum_{i}^{N_q} Var(A_q[i]) \tag{3}$$

where N_q is the number of pillars that enclosed by the b_q, Var($\cdot$) calculates the variance of the given vector.

5. Implementation

Voxelization We voxelize the point clouds according to the x,y,z axes. For nuScenes dataset, the range for voxelization are $[-51.2, 51.2]$m, $[-51.2, 51.2]$m, and $[-5.0, 3]$m in terms of x,y,z. For Waymo dataset, the range are $[-75.2, 75.2]$m, $[-75.2, 75.2]$m, and $[-2, 4]$m. Unless specifically mentioned, all of our experiments are conducted in low voxelization resolution of $[0.1, 0.1, 0.1]$m of the x,y,z axes.

Augmentation The point clouds are randomly flipped according to the x,y axes, rotated around z axis with a range

Methods	NDS	mAP	runtime	car	truck	cons.	bus	trailer	barrier	motorcycle	bicycle	pedestrian	traffic cone
PointPillars [19]	45.3	30.5	17ms	68.4	23.0	4.1	28.2	23.4	38.9	27.4	1.1	59.7	30.8
WYSIWYG [15]	41.9	35.0	-	79.1	30.4	7.1	46.6	40.1	34.7	18.2	0.1	65.0	28.8
PointPainting [27]	59.2	46.4	-	77.9	35.8	15.8	36.2	37.3	60.2	41.5	24.1	73.3	62.4
CBGS [33]	63.3	52.8	55ms	81.1	48.5	10.5	54.9	42.9	65.7	51.5	22.3	80.1	70.9
CVCNet [5]	64.2	55.8	91ms	82.7	46.1	20.7	45.8	46.7	69.9	61.3	34.3	81.0	69.7
OHS [6]	66.0	59.3	60ms	83.1	50.9	23.0	56.4	53.3	71.6	63.5	36.6	81.3	73.0
CenterPoint [30]	67.3	60.3	70ms	**85.2**	53.5	20.0	63.6	**56.0**	71.1	59.5	30.7	**84.6**	78.4
VISTA-OHS (Ours)	**69.8**	**63.0**	69ms	84.4	**55.1**	**25.1**	**63.7**	54.2	**71.4**	**70.0**	**45.4**	82.8	**78.5**

Table 1. 3D detection results on the nuScenes test server. "cons." refers to construction vehicle.

of $[-0.3925, 0.3925]$ rad, scaled with a factor ranging from 0.95 to 1.05, and translated with range $[0.2, 0.2, 0.2]$ m in x,y,z axes. The class-balanced grouping and sampling [33], and the database sampling [28] are adopted to increase the ratio of positive samples during training.

Joint Training We train the VISTA on various target assignment [6, 30, 33]. To train the network, the original loss for different target assignments is calculated, we recommend readers to refer to their original papers for more details of the loss. Briefly, we take classification and regression into account:

$$\mathcal{L}_{target} = \lambda_1 \mathcal{F}_{cls}(\hat{y}, y) + \lambda_2 \mathcal{F}_{reg}(\hat{b}, b) \tag{4}$$

where λ_1 and λ_2 are the loss weights, $\mathcal{F}_{cls}(\cdot, \cdot)$ is the classification loss function between ground-truth labels $\hat{y}$ and predictions y, $\mathcal{F}_{reg}(\cdot, \cdot)$ is the regression loss function between ground-truth bounding boxes $\hat{b}$ and predicted ones b.

The total loss $\mathcal{L}$ is the weighted sum of $\mathcal{L}_{target}$ and $\mathcal{L}_{var}$: $\mathcal{L} = \mathcal{L}_{target} + \lambda_3 \mathcal{L}_{var}$. We set λ_1, λ_2, and λ_3 to 1.0, 0.25, 1.0. We apply Focal loss [20] as $\mathcal{F}_{cls}$, and L1 loss for $\mathcal{F}_{reg}$.

6. Experiments

	Avg	Linear Atten	Conv Atten	Var Cons	Decouple	mAP	NDS	Runtime
(a)						59.5	66.0	60ms
(b)	✓					59.2	65.8	61ms
(c)		✓				58.7	65.9	63ms
(d)			✓			60.0	66.8	64ms
(e)			✓	✓		60.4	67.5	64ms
(f)			✓	✓	✓	**60.8**	**68.1**	69ms

Table 2. Ablation studies of VISTA on multi-view fusion. The performance is evaluated on the nuScenes validation set.

Method	mAP	NDS	Method	mAP	NDS	Method	mAP	NDS
CBGS	51.9	61.5	CenterPoint	56.4	64.8	OHS	59.5	66.0
V-CBGS	**53.2** (+1.3)	**62.8** (+1.3)	V-CenterPoint	**57.6** (+1.2)	**65.6** (+0.8)	V-OHS	**60.8** (+1.3)	**68.1** (+2.1)

Table 3. 3D detection results of VISTA-based state-of-the-art methods (V-method) on nuScenes validation set. For efficiency, all methods are experimented based on the low voxelization resolution configuration provided in their official codebase.

We evaluate VISTA on nuScenes dataset and Waymo Open Dataset. We test the efficacy of VISTA on three state-of-the-art methods with different target assignment: CBGS [33], OHS [6], and CenterPoint [30].

6.1. Dataset and Technical Details

nuScenes Dataset contains 700 scenes for training, 150 scenes for validation, and 150 scenes for testing. The dataset is annotated at 2Hz, in total 40000 key-frames are annotated with 10 object categories. Following [33], we combine 10 sweeps for each annotated key-frame to increase the number of points. Average precision (mAP) and nuScenes detection score (NDS) are applied in our performance evaluation. NDS is a weighted average of mAP and other attributes metrics, including translation, scale, orientation, velocity, and other box attributes. During training, we follow CBGS [33] to optimize the model via Adam [16] optimizer with one-cycle learning rate policy [12].

Waymo Open Dataset contains 798 sequences for training, 202 sequences for validation. Each sequence is of 20s duration and sampled at 10Hz with a 64 channels LiDAR, containing 6.1M vehicle, 2.8M pedestrian, and 67k cyclist boxes. We evaluate our networks on the metric of standard mAP and mAP weighted by heading accuracy (mAPH), which are based on the IoU threshold of 0.7 for vehicles, 0.5 for pedestrians and cyclist. The official evaluation protocol evaluates the methods in two difficulty levels: LEVEL_1 for boxes with more than five LiDAR points, and LEVEL_2 for boxes with at least one LiDAR point.

6.2. Comparison with other methods

We submitted the test results of the proposed VISTA-based OHS to the nuScenes test server. To benchmark the results, we follow [30] to tune up the resolution for training and utilize the double flip for testing augmentation. Since our results are based on single model, methods which use ensemble models and extra data are not included in our comparisons. The test performance are shown in Table 1. The proposed VISTA achieves state-of-the-art performance on nuScenes test set, outperforming all published methods in both overall mAP and NDS by large margins. Particularly, the performances on the motorcycle and the bicycle surpass the second best method CenterPoint [30] by up to 48% in mAP. Specifically, the performance gains in

Method	LEVEL_1						LEVEL_2					
	Vehicle		Pedestrian		Cyclist		Vehicle		Pedestrian		Cyclist	
	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH
StarNet [23]	61.5	61.0	67.8	59.9	-	-	54.9	54.5	61.1	54.0	-	-
PointPillars [19]	63.3	62.8	62.1	50.2	34.7	25.3	55.6	55.1	55.9	45.1	33.3	24.3
PPBA [23]	67.5	67.0	69.7	61.7	-	-	59.6	59.1	63.0	55.8	-	-
RCD [1]	72.0	71.6	-	-	-	-	65.1	64.7	-	-	-	-
CenterPoint [30]	81.0	80.6	80.5	77.3	74.6	73.6	73.4	73.0	74.5	71.5	72.1	71.2
VISTA-CenterPoint	**81.7**	**81.3**	**81.4**	**78.3**	**74.9**	**73.9**	**74.4**	**74.0**	**75.5**	**72.5**	**72.5**	**71.6**

Table 4. 3D detection results on the Waymo test server

geometric-similar categories (e.g. truck, construction vehicle) confirm the efficacy of our proposed decoupling design.

To further validate the effectiveness of our proposed VISTA, we adopt the proposed VISTA to the CenterPoint [30], and submitted test results to the Waymo test server. During training and testing, we follow exactly the same rules as CenterPoint. The test performance is shown in Table 4. VISTA brings significant improvements to CenterPoint on all categories of all levels, outperforming all the published results.

6.3. Ablation Studies

VISTA in Multi-View Fusion As shown in Table 2, to demonstrate the superiority of the proposed VISTA, we conduct the ablation studies with OHS [6] as our baseline (a) on the validation set of nuScenes dataset. As elaborated in the Section 4.3, without attention constraint, the extreme case for learned attention weights will be global average pooling (GAP). To clarify, we manually obtain the RV features via GAP, and add them back to all BEV features as fusion. Such a GAP-based fusion method (b) drops the performance of baseline to 59.2% in overall mAP, indicating the necessary of adaptively fusing multi-view features from global spatial context. Directly adopt the VISTA for multi-view fusion (d) results in 60.0% in mAP. When replace the convolutional attention module to the conventional linear one (c), the overall mAP drops to 58.7%, which reflects the significance of aggregating local cues for constructing cross-view attention. After adding the proposed attention variance constraint, as demonstrated (e), the performances raise to 60.4% in overall mAP. The performance gains from (d) to (e) rows indicate that the attention mechanism can be well guided via the attention constraint, and as we will analyze in the Section 6.4, the attention module is able to attend to regions of interest of whole scenes. Nevertheless, the shared attention modeling will bring the conflicts between the learning of classification and regression tasks, where the former task will further be dominant by the latter one in 3D object detection. As shown in (f), after decoupling the attention modeling, the performances raise from 60.4% to 60.8% in overall mAP, further verifying our assumption.

VISTA in Different Target Assignments The proposed VISTA is a plug-and-play multi-view fusion method and can be adopted in various recent advanced target assign strategies with slight modifications. To demonstrate the effectiveness and generalization ability of the proposed VISTA, we implement the VISTA on CenterPoint [30], OHS [6], and CBGS [33], which are recent state-of-the-art methods. These methods stand for different main stream target assignments in terms of anchor-based or anchor-free manners. We evaluate the results on the validation set of the nuScenes dataset, for verification, all the methods are implemented based on the low voxelization resolution (i.e. $[0.1, 0.1, 0.1]$m of x,y,z axes) configuration provided by their official codebase. As demonstrated in Table 3, all the three target assignments achieve large performance gains in both overall mAP and NDS scores (approximately 1.3% and 1.4% in mAP and NDS), indicating that the proposed VISTA can fuse multi-view features of universally high quality via dual cross-view attention mechanism.

VISTA in Real-World Application We demonstrate the runtime of the proposed VISTA on one RTX3090 GPU in Table 2. Without any modifications, the baseline (a) runs at 60ms per frame. After adopting the convolutional attention module (d) in the baseline, the runtime increases to 64ms. We can observe from (e) and (f) that, while applying the proposed attention variance constraint does not influence the inference speed, the decoupling design costs 5ms, yet the extra decay is still negligible. Being running with such efficiency, we argue that the proposed VISTA completely meets the requirements of real-world application.

6.4. Analysis of VISTA

We argue that the VISTA trained by the proposed attention constraint can capture the global and local correlations between BEV and RV, thus can effectively perform multi-view fusion for accurate bounding boxes prediction. To vividly present the effectiveness of attention variance constraint in training VISTA, we visualize the constructed cross-view correlations with and without attention variance constraint in Figure 3. Given the area containing a bounding box from target view (BEV) to query the source view (RV),

https://github.com/poodarchu/Det3D
https://github.com/tianweiy/CenterPoint

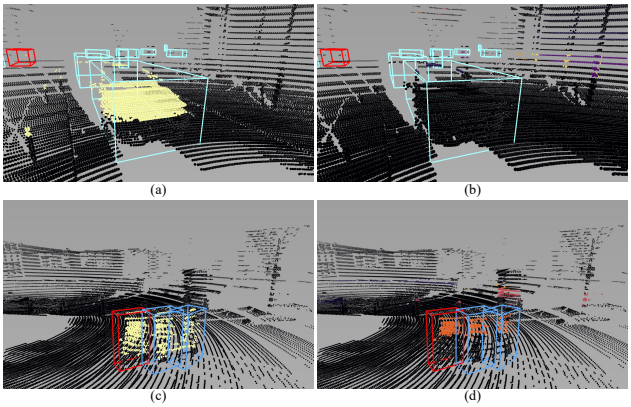

Figure 3. The visualization of learned attention map of VISTA with ((a) and (c)) and without ((b) and (d)) the attention variance constraint. Each row presents one sampled scene. The query bounding boxes are illustrated in red color. The brighter the color of the points, the higher the attention weight of the points.

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, without the proposed attention variance constraint, the learned attention weights hold small values for almost every pillars in the RV, resulting in an approximate global average pooling operation. In Figure 3 (b) and (d), the attention module attends to the background points far from the query cars and pedestrians, and the attention weights for each focusing region are relatively low. The attention module trained with attention variance constraint instead, highlights the objects with the same categories of the queries, as presented in Figure 3 (a) and (c). Especially, for the query cars, the attention module trained via attention variance constraint successfully attends on the other cars in the scenes.

The another key design of our proposed VISTA is the decoupling of the classification and regression tasks. The individual attention modeling for these two tasks mitigates the imbalanced learning issues, therefore the detection results are more accurate and reliable. To present the significance of our design, we present the detection results before and after the decoupling in the Figure 4. Each row represent one scene, and the left column presents the results with decoupling, the other column shows the results without decoupling. As illustrated in Figure 4 (b) and (d), the 3D detector without decoupling design easily mistakes the objects A for the other B of similar geometric property, we term such phenomenon as A-to-B, such as bus (purple)-to-truck (yellow), bus (purple)-to-trailer (red), and bicycle (white)-to-motorcycle (orange), proving the imbalanced training of classification and regression tasks. Moreover, the confused predictions are not accurate when compare the right column to the left one. On the contrary, the VISTA with proposed decoupling design successfully distinguishes the categories of the objects, and predicts the tight bounding boxes, as

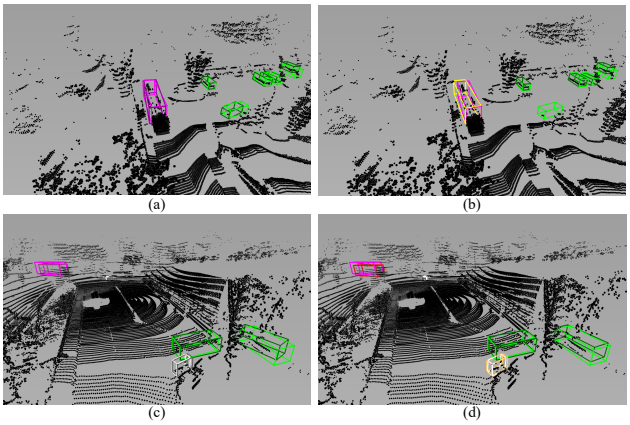

Figure 4. The visualization of the detection results learned with and without decoupling design. Each row represents one sample scene. The bounding boxes illustrated in light color refer to the ground truth bounding boxes, the boxes painted in dark color represent the correct prediction results, and the boxes illustrated in different accent color indicate the wrong predictions.

shown in the Figure 4 (a) and (c), demonstrating the efficacy of the proposed decoupling design.

7. Discussion

Broader Impact 3D object detection is vital to autonomous driving. Our proposed VISTA can identify the objects of interest precisely and comprehensively to ensure the safety in autonomous driving. However, misuse of technology will allow some malicious people or teams to invade and attack this perception part. Therefore, we encourage future research to mitigate these risks and make LiDAR sensor more robust and better.

Limitations We realize the multi-view fusion via the cross-view attention mechanism in the proposed VISTA. The attention mechanism requires large amount of data to train, and might perform poorly given insufficient data.

Conclusion In this paper, we propose VISTA, a novel plug-and-play multi-view fusion strategy for accurate 3D object detection. To empower the VISTA to have the ability to attend to the specific targets rather than generic points, we propose to constraint the variance of the learned attention weights. We decouple the classification and regression tasks to handle the issue of imbalanced training. Our proposed plug-and-play VISTA is able to produce high quality fused features for the prediction of proposals, and can be applied with various target assignment methods. Benchmarking on the nuScenes and Waymo datasets demonstrate the efficacy and generalization ability of our proposed method.

Acknowledgement This work is supported in part by the Guangdong R&D key project of China (Grant No.: 2019B010155001) and the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (Grant No.: 2017ZT07X183).

References

[1] Alex Bewley, Pei Sun, Thomas Mensink, Dragomir Anguelov, and Cristian Sminchisescu. Range conditioned dilated convolutions for scale invariant 3d object detection. *arXiv preprint arXiv:2005.09927*, 2020.

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

[4] Ke Chen, Ryan Oldja, Nikolai Smolyanskiy, Stan Birchfield, Alexander Popov, David Wehr, Ibrahim Eden, and Joachim Pehserl. Mvlidarnet: Real-time multi-class scene understanding for autonomous driving using multiple views. In *International Conference on Intelligent Robots and Systems*, 2020.

[5] Qi Chen, Lin Sun, Ernest Cheung, and Alan L Yuille. Every view counts: Cross-view consistency in 3d object detection with hybrid-cylindrical-spherical voxelization. *Advances in Neural Information Processing Systems*, 2020.

[6] 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.

[7] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Benjamin Graham. Spatially-sparse convolutional neural networks. *arXiv preprint arXiv:1409.6070*, 2014.

[10] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[11] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017.

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

[13] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 2021.

[14] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021.

[15] Peiyun Hu, Jason Ziglar, David Held, and Deva Ramanan. What you see is what you get: Exploiting visibility for 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

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

[17] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *IEEE International Conference on Intelligent Robots and Systems*, 2018.

[18] Ankit Laddha, Shivam Gautam, Stefan Palombo, Shreyash Pandey, and Carlos Vallespi-Gonzalez. Mvfusenet: Improving end-to-end object detection and motion forecasting through multi-view fusion of lidar data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[19] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[22] Gregory P Meyer, Ankit Laddha, Eric Kee, Carlos Vallespi-Gonzalez, and Carl K Wellington. Lasernet: An efficient probabilistic 3d object detector for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[23] Jiquan Ngiam, Benjamin Caine, Wei Han, Brandon Yang, Yuning Chai, Pei Sun, Yin Zhou, Xi Yi, Ouais Alsharif, Patrick Nguyen, et al. Starnet: Targeted computation for object detection in point clouds. *arXiv preprint arXiv:1908.11069*, 2019.

[24] Hualian Sheng, Sijia Cai, Yuan Liu, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, and Min-Jian Zhao. Improving 3d object detection with channel-wise transformer. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[25] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[27] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detec-

tion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[28] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 2018.

[29] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[30] 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.

[31] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In *Conference on Robot Learning*, 2020.

[32] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[33] 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.

[34] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.