

PVTransformer: Point-to-Voxel Transformer for Scalable 3D Object Detection

Zhaoqi Leng¹, Pei Sun¹, Tong He¹, Dragomir Anguelov¹, and Mingxing Tan¹

Abstract—3D object detectors for point clouds often rely on a pooling-based PointNet [20] to encode sparse points into grid-like voxels or pillars. In this paper, we identify that the common PointNet design introduces an information bottleneck that limits 3D object detection accuracy and scalability. To address this limitation, we propose PVTransformer: a transformer-based point-to-voxel architecture for 3D detection. Our key idea is to replace the PointNet pooling operation with an attention module, leading to a better point-to-voxel aggregation function. Our design respects the permutation invariance of sparse 3D points while being more expressive than the pooling-based PointNet. Experimental results show our PVTransformer achieves much better performance compared to the latest 3D object detectors. On the widely used Waymo Open Dataset, our PVTransformer achieves state-of-the-art 76.5 mAPH L2, outperforming the prior art of SWFormer [27] by +1.7 mAPH L2.

I. INTRODUCTION

3D object detection for autonomous driving in urban environments requires processing a large amount of sparse and unordered points scattered across an open three-dimensional space. To manage the irregular distribution of points, existing methods aggregate points into two or three-dimensional voxel representation [35], utilizing a PointNet-type feature encoder [20] to aggregate point features to voxels, which is then followed by a backbone and detection heads. However, the existing point architecture is often overlooked and limited by its minimalist design, i.e., a few fully connected layers followed by a max pooling layer. The key to PointNet-type modules, as highlighted in the original paper [20], lies in the max pooling layer, which extracts information from unordered points and serves as an aggregation function. Despite the utilization of numerous fully connected layers for feature extraction, the features of all points within a voxel are combined through a simple pooling layer. We observe that the vanilla pooling operation in 3D object detection introduces an information bottleneck, hindering the performance of modern 3D object detectors. Unlike the standard 2D max-pooling in image recognition, which operates on a limited set of pixels, the point-voxel pooling layer in 3D detectors must aggregate a large number of unordered points. For instance, in the Waymo Open Dataset [26], it is common to find over 100 points in a single 0.32m × 0.32m voxel, pooled into a single voxel feature vector. This results in significant loss of information from the point features after the pooling layer.

To address the limitations of the pooling-based PointNet architecture, we introduce **PVTransformer**, a new attention-point-voxel architecture for 3D object detection based on the

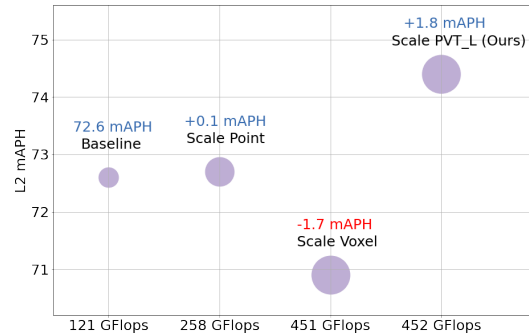


Fig. 1: **PVTransformer (PVT) as a scalable architecture.** PVTransformer addresses the pooling bottleneck in previous voxel-based 3D detectors and demonstrates better scalability compared to scaling PointNet (Scale Point) and voxel architectures (Scale Voxel). The size of points represents the model Flops. More details are shown in Figure 4, 5.

Transformer [29]. The aim of PVTransformer is to mitigate the information bottleneck introduced by the pooling operation in modern 3D object detectors, by end-to-end learning a point-to-voxel encoding function via an attention module. In PVTransformer, each point in a voxel is considered a token, and a single query vector is used to query all the point tokens, thereby aggregating and encoding all point features into a single voxel feature vector. The attention-based aggregation module in PVTransformer serves as a set operator, maintaining permutation invariance, but is more expressive than max pooling. Notably, unlike other Transformer-based point networks such as Point Transformer [32], which use pooling to aggregate points, PVTransformer is designed to *learn* a feature aggregation function without the need for a heuristic pooling.

We evaluate our PVTransformer on the Waymo Open Dataset, the largest public 3D point cloud dataset [26]. Experimental results show that our PVTransformer significantly outperforms previous PointNet-based 3D object detectors, by improving point-to-voxel aggregation. Moreover, PVTransformer enables us to scale up the model, achieving a new state-of-the-art 76.1 mAPH L2 and 85.0/84.7 AP L1 for vehicles and pedestrians. Notably, our voxel backbone and loss design are largely based on the prior art of SWFormer [27], but our newly proposed point-to-voxel Transformer enables us to achieve +1.7 mAPH L2 than the baseline SWFormer.

Our contributions can be summarized as:

- **New architecture:** Introduction of an attention-based point-voxel architecture, PVTransformer, aiming to ad-

¹Waymo Research

dress the pooling limitation of PointNet.

- **Novel scaling studies:** Initiating the exploration of scalability in Transformer-based 3D detector architectures.
- **Extensive studies:** Through extensive architecture search, we show the effectiveness of the proposed PVTransformer architecture, which achieves a state-of-the-art 76.5 mAPH L2 on the Waymo Open Dataset.

II. RELATED WORKS

3D object detection has been well explored in previous works, and can be roughly divided into point-based and voxel-based approaches. Most point-based detectors [19], [18], [25], following seminal works such as PointNet [20] and PointNet++ [21], directly operate on the unordered points with fully connected layers to learn per-point features. These models are often not scalable because directly processing large amount of unordered points is expensive [14].

On the contrary, voxel-based models, such as VoxelNet [35] and PointPillars [9], are generally more scalable by discretizing the sparse point clouds into grids, and thus become increasingly popular in recent research. Voxel-based detectors first convert unordered points into ordered two or three-dimensional voxel grids, by aggregating nearby points based on their coordinates, and then operate on the voxel grid to produce box predictions. The aggregation step typically relies on a simple max pooling layer, which introduces an information bottleneck. Our PVTransformer aims to redesign this fundamental pooling-based point aggregation module with a transformer-based learned aggregation function.

Recent works [16], [3], [5], [1] attempt to directly predict 3D bounding boxes purely based on range images. Meanwhile, multi-view 3D detectors [33], [28] leverage multiple views of point clouds, such as points, voxels, and range images. Transformer architecture is also adopted as the backbone [32], [17], [13], [30], [10], [4], as the set operation is naturally suitable for processing sparse LiDAR points or voxels. Unlike those works, PVTransformer aims to use the attention module in Transformers to learn a feature aggregation function between the point and voxel representations.

III. PVTRANSFORMER

In this section, we provide an overview of our proposed PVTransformer architecture for voxel-based 3D detectors. First, we present the conventional network structure of voxel-based 3D detectors. Then, we introduce the details of our PVTransformer architecture and its key innovations. Finally, we discuss the scalability of the proposed PVTransformer.

A. Voxel-based 3D detectors

In this work, we rely on 2D voxel-based (pillar) representation, which is introduced in PointPillars [9]. A typical voxel-based 3D detector starts with raw 3D point clouds and uses a lightweight PointNet to extract per-point features. These point features are then aggregated into individual voxel grids using a max pooling layer, followed by a high-capacity backbone network for final box predictions.

As illustrated in Figure 3 (a), the pooling layer in PointNet aggregates all the point features, represented as $\mathbb{R}^{n \times d}$ within a single voxel, into a single voxel feature $\mathbb{R}^{1 \times d}$. Here, n represents the number of points per voxel and d represents the feature size. On the Waymo Open Dataset, the number of LiDAR points n in a pillar with size $0.32\text{m} \times 0.32\text{m}$ commonly exceeds 30, resulting in a $30 \times$ compression of information. This aggressive pooling operation causes significant information loss and creates an information bottleneck between point features and voxel features.

To overcome this limitation, we propose to use attention to learn the weighting of point features, instead of the sample max pooling operation. This idea leads to our proposed architecture PVTransformer.

B. Point-to-voxel encoding

A key contribution of our PVTransformer architecture is replacing the *non-trainable* pooling operation used in PointNet with a *trainable* Transformer-based aggregation module.

Treating points as tokens. In PVTransformer, each point within a voxel is treated as the value and key token of a Transformer layer. A query vector, learned from all the points within the voxel, is used to encode the point information into a single voxel representation, which we will discuss in details in the next paragraph. The advantage of this design is that a multi-head attention module can learn a much more expressive point-to-voxel aggregation function compared to a standard pooling operation. The attention layer can dynamically weight point features, which leads to more expressive aggregation mechanism. In fact, the attention mechanism supersedes pooling operation, i.e., when the learned attention weight for each feature is one-hot based on the largest feature value, the attention module is equivalent to max pooling. Meanwhile, when the learned attention weight is the same for all the points, the attention module becomes the same as average pooling.

Latent query. A simple approach is to use a randomly initialized latent vector, as Set Transformer and Preceiver [10], [8]. Here, a single randomly initialized latent vector $Q \in \mathbb{R}^{1 \times d}$ is used to query a set of LiDAR point features $L \in \mathbb{R}^{n \times d}$, where d represents the number of features and n represents the number of points inside a voxel. Point features are aggregated and encoded as a single voxel feature vector by $V = \text{Transformer}(q = Q, v = L, k = L) \in \mathbb{R}^{1 \times d}$, where the latent variable Q are learned end-to-end during training. However, we find that using latent query can cause training instability.

Residual query. To address training instability, we propose using a residual query vector to stabilize the training. We use the max pooled features of the points within a voxel as a prior, which is added to the latent query. This residual query variable, is used to query and encode the point features, leading to the final voxel representation, as shown in Figure 3 (b).

The residual variable $RQ = \text{maxpool}(L) + Q$, where $RQ \in \mathbb{R}^{1 \times d}$, is used to query the point features, i.e., $V =$

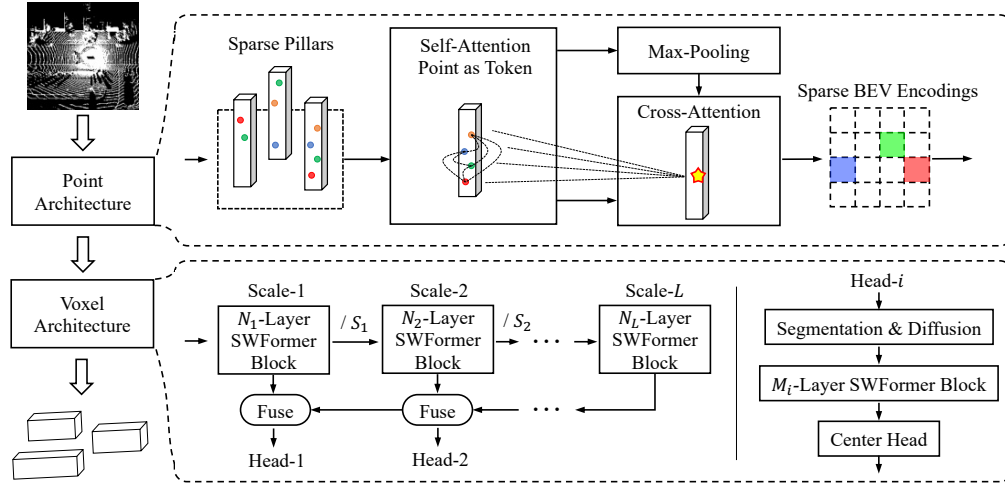


Fig. 2: **Overview of the PVTransformer architecture.** The PVTransformer architecture contains a point architecture and a voxel architecture. Its novelty lies in the point architecture, substituting PointNet with a novel Transformer design. In the point architecture, points are binned into pillars, each is considered as a token. Within a voxel, points undergo self-attention Transformer followed by a cross-attention Transformer to aggregate point features into voxel features, details further shown in Figure 3 (b). The sparse BEV voxel features proceed to the voxel architecture, employing a multi-scale sparse window Transformer (SWFormer Block) [27] for encoding and CenterNet heads for bounding boxes predictions [31].

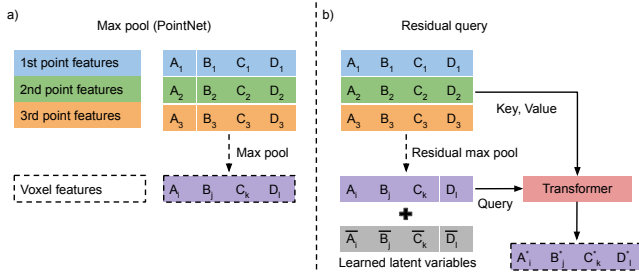


Fig. 3: **Point-to-voxel aggregation in PVTransformer.** This module replaces PointNet’s max pooling [20] with a Transformer layer. (a) The vanilla max pooling layer aggregates point features by selecting the element-wise max feature (A_i, B_j, C_k, D_l) from 3 point features to form the voxel features, where $i, j, k, l \in [1, 3]$. (b) Our proposed residual query uses the sum of max pooled feature and a learnable latent vector $(\bar{A}_i, \bar{B}_j, \bar{C}_k, \bar{D}_l)$, i.e. $(A_i + \bar{A}_i, B_j + \bar{B}_j, C_k + \bar{C}_k, D_l + \bar{D}_l)$, to query and aggregate point features.

$Transformer(q = RQ, v = L, k = L) \in \mathbb{R}^{1 \times d}$. The latent query Q allows a more flexible query vector representation than using max pooled feature along, while the max pooled feature vector $maxpool(L)$ helps stabilize the training.

C. Model scaling

PVTransformer mitigates the point-voxel information bottleneck, which leads to the second contribution of this work: *systematic study of the scaling properties of the Transformer-based 3D detectors*. To the best of our knowledge, this is the first time the scaling of Transformer-based 3D object detection has been explored in depth. Our results not only show that PVTransformer has better scaling properties by achieving state-of-the-art performance and providing valu-

able new insights on the best scaling practice.

Factorized scaling. In this study, we analyze the scalability of Transformer-based 3D detectors by factorizing PVTransformer architectures into two separate components: the point architecture and the voxel architecture, as depicted in Figure 2. Our PVTransformer exhibits significant better scaling property compared to PointNet and proves to be a more effective solution than simply scaling the voxel architecture, which is prone to overfitting.

Data Augmentation. Strong regularization is important when scaling up the model capacity. We use LidarAugment, including 10 data augmentations, during training [11].

Optimization for TPU accelerators. Despite the recent advancements in optimizing 3D detectors for GPU training and inference, the use of TPU accelerators, which are widely adopted for training *large-scale* Transformer models, has not been extensively explored for 3D detection models.

TPU accelerators are efficient for dense, static shape computations, but less so for dynamic operations. In this study, we identify that dynamic voxelization proposed in [34] is 2X slower on TPU, despite being FLOPS efficient as shown in Table I, thus for the following experiments, we use fixed voxelization.

IV. EXPERIMENTS

Our primary evaluation of the PVTransformer architecture is conducted on the Waymo Open Dataset [26], which comprises 798 training, 202 validation and 150 test run segments, each run containing around 200 frames, including 360 degree of LiDAR around the ego-vehicle. The Waymo Open Dataset offers a unique advantage over other 3D outdoor detection datasets [6], [2], as it contains more than $10 \times$ annotated

Lidar frames, providing a valuable opportunity to study the scalability of transformer-based 3D detectors.

A. Implementation details

Overall architecture of PVTransformer. The PVTransformer architecture is composed of a point architecture and a voxel architecture, as shown in Figure 2. The key novelty of PVTransformer lies in its point architecture, which utilizes a transformer module to aggregate point features. This is different from other PointNet-based transformer 3D detectors [27], [4]. The point architecture extracts point features and transforms them into $0.32\text{m} \times 0.32\text{m}$ bird’s eye view pillars, as shown in Figure 3. These features are then fed into the voxel architecture to predict bounding boxes. The box detection is through CenterNet heads [31]. We follow the loss implementations in [27], which use a weighted average of foreground segmentation loss, box regression loss, and heatmap loss. We experimented both single-scale and multi-scale backbone designs and selected the multi-scale design for its superior performance, as shown in Figure 4.

Efficient static-shape point architecture. Dynamic voxelization [34] is a widely used in 3D detection models due to its FLOPs efficiency. However, we have found that dense operations can be optimized through TPU XLA, leading to much more efficient computation, as demonstrated in Table I.

We benchmark the latency of a depth-two PointNet with a channel size of 128. Here, we use a 3-frame training setting, as in Table II, and define $f = 4$, $m = 199600$, and $N = 52000$ as the original number of point features, total number of points per voxel, and total number of voxels, respectively.

For the dynamic voxelization method, the initial point tensors $\tilde{P} \in \mathbb{R}^{m \times f}$ are transformed into $P \in \mathbb{R}^{m \times 128}$ via two fully connected layers. The transformed point tensors are then aggregated into a voxel tensor $V \in \mathbb{R}^{N \times 128}$ through dynamic gather in the forward pass and scatter in the backward pass.

In contrast, the PointNet fixed voxelization method simplifies the implementation by bucketing the point tensors into $\tilde{P} \in \mathbb{R}^{N \times 32 \times f}$ and transforming them into $P \in \mathbb{R}^{N \times 32 \times 128}$ through two fully connected layers. The voxel features $V \in \mathbb{R}^{N \times 128}$ are then aggregated via a simple max pooling operation along the second axis of P , with all the operations being static without any dynamic gather or scatter.

In our proposed PVTransformer architecture, we take advantage of the simplification provided by the fixed voxelization approach. First, the point features are transformed into $P \in \mathbb{R}^{N \times 32 \times 128}$ via a single fully connected layer. Then, points within each pillar bucket are treated as tokens with total sequence length 32. We apply a point self-attention Transformer among point tokens within each voxel to further featurized points. Finally, point-to-voxel aggregation is implemented as a cross-attention Transformer between the point tensor $P \in \mathbb{R}^{N \times 32 \times 128}$ and a querying tensor $Q \in \mathbb{R}^{N \times 1 \times 128}$, along the second axis of P and Q , resulting in a voxel tensor with shape $V \in \mathbb{R}^{N \times 1 \times 128}$. The implementation of the querying vector is shown in Figure 3 (b).

Training setup. We employ the AdamW [15] optimizer and a cosine learning rate with warm up. The initial warm-up

Latency (ms)	batch size 1		batch size 2		batch size 4	
	forward	backward	forward	backward	forward	backward
Dynamic voxelization	13.7	16.6	24.3	33.1	48.3	66.0
Fixed voxelization	6.5	7.8	12.1	15.7	23.4	31.3
Speed up	2.1×	2.1×	2.0×	2.1×	2.1×	2.1×

TABLE I: $2 \times$ training speed with fixed voxelization.

Method	mAPH	Vehicle AP/APH 3D		Pedestrian AP/APH 3D	
	L2	L1	L2	L1	L2
PPillars [9]	51.9	63.3/62.7	55.2/54.7	68.9/56.6	60.4/49.1
LiDAR R-CNN [12]	58.0	73.5/73.0	64.7/64.2	71.2/58.8	63.1/51.7
CenterPoint [31]	67.1	76.6/76.1	68.9/68.4	79.0/73.4	71.0/65.8
PVRCNN [23]	67.3	77.5/76.9	68.7/68.2	78.9/75.1	69.8/66.4
AFDetV2[7]	68.1	77.6/77.1	69.7/69.2	80.2/74.6	72.2/67.0
RSN_3f [28]	68.1	78.4/78.1	69.5/69.1	79.4/76.2	69.9/67.0
PVRCNN++ [24]	69.1	79.3/78.8	70.6/70.2	81.8/76.3	73.2/68.0
SST_3f [4]	69.5	77.0/76.6	68.5/68.1	82.4/78.0	75.1/70.9
UPillars-L_3f [11]	71.0	79.5/79.0	71.9/71.5	81.5/77.3	74.5/70.5
PillarNet-34_2f [22]	71.8	80.0/79.5	72.0/71.5	82.5/79.3	75.0/72.0
CenterFormer_2f [36]	72.5	77.0/76.5	72.1/71.6	81.4/78.0	76.7/73.4
SWFormer_3f† [27]	72.8	80.9/80.4	72.8/72.4	84.4/80.7	76.8/73.2
PVTransformer_3f	74.0	81.9/81.3	73.8/73.4	85.3/81.8	78.0/74.6
PVTransformer_3f.L	74.4	82.2/81.7	74.3/73.9	85.3/81.9	78.2/74.8

TABLE II: PVTransformer outperforms other state-of-the-art architectures. PVTransformer increases all detection metrics compared to recent state-of-the-art SWFormer, CenterFormer, and PillarNet. PVT only scales the point architecture (see Table IV), while PVT_L scales both point and voxel architectures (see Table V). Waymo Open Dataset validation set 3D detection metrics are reported. † value reported in [11]. Our reproduced model has 72.6 mAPH L2.

learning rate $3e-4$, with a peak learning rate $1.2e-3$, 10000 warm-up steps, and a total of 320000 training steps. The model is trained with global batch size 64 for about 128 epochs. For most of the experiments, we use a 3-frames setting, stacking point clouds from the current frame, the previous frame, and the frame before the previous one.

B. PVTransformer for 3D object detection

Table II and Table III compares the performance of our PVTransformer to other recent 3D detectors. The main comparisons for PVTransformer are SWFormer and SST [27], [4], both are transformer-based detectors and trained with the 3 frames as our PVTransformer. Note, both SWFormer and SST are using a simple PointNet as the point architecture as opposed to our new attention design, as shown in Figure 2.

As shown in Table II, our PVTransformer improves upon the state-of-the-art 3D detectors across all metrics, including L1 and L2 difficulties, the Vehicle and Pedestrian classes, and AP and APH. These performance gains are consistent in the test set as shown in Table III, with an improvement of 1.7 and 1.9 mAPH L2 than recent state-of-the-art transformer detector SWFormer and convolution detector PillarNet [22] and achieving a new state-of-the-art 76.5 mAPH L2 on the test set, demonstrating the effectiveness of our PVTransformer architecture.

C. Scaling Transformer-based 3D detector

In this section, we present the first systematic architecture scaling study for Transformer-based 3D detector and

Method	mAPH L2	Vehicle AP/APH 3D		Pedestrian AP/APH 3D	
		L1	L2	L1	L2
Pillars [9]	55.1	68.6/68.1	60.5/60.1	68.0/55.5	61.4/50.1
CenterPoint [31]	69.1	80.2/79.7	72.2/71.8	78.3/72.1	72.2/66.4
RSN_3f [28]	69.7	80.7/80.3	71.9/71.6	78.9/75.6	70.7/67.8
PVRCNN++ [24]	71.2	81.6/81.2	73.9/73.5	80.4/75.0	74.1/69.0
SST_TS_3f [4]	72.9	81.0/80.6	73.1/72.7	83.1/79.4	76.7/73.1
PillarNet-34_2f [22]	74.6	83.2/82.8	76.1/75.7	82.4/79.0	76.7/73.5
SWFormer_3f [27]	74.8	84.0/83.6	76.3/76.0	83.1/79.3	77.2/73.5
PVTransformer_3f	76.1	84.8/84.5	77.3/77.0	84.2/80.6	78.5/75.1
PVTransformer_3f.L	76.5	85.0/84.6	77.6/77.3	84.7/81.2	79.0/75.7

TABLE III: **New state-of-the-art test-set results on Waymo Open Dataset.** PVTransformer achieves a new state-of-the-art mAPH L2, with +1.7 mAPH gain compared to the prior art of SWFormer.

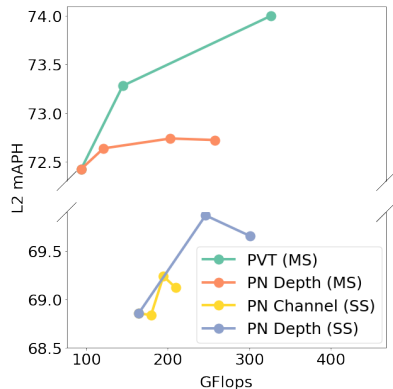


Fig. 4: **PVTransformer: better scalability.** Increasing PointNet’s (PN) depth (red, purple) and channel (yellow) yields modest performance improvements, while scaling PVTransformer PVT (green) shows significant improvements. Previous works, both single-scale (SS) [4] and multi-scale (MS) [27] architectures, use PointNet for point feature aggregation, yet it underperforms when scaled beyond certain thresholds, leading to overfitting. PVTransformer (green) overcomes these limitations by incorporating a Transformer-based point-to-voxel encoder, enabling effective scaling beyond 300 GFlops and achieving 74.0 mAPH L2 for vehicle and pedestrian detection on the Waymo Open Dataset Validation set.

demonstrate that our proposed PVTransformer scales better compared to scaling *point architectures* (PointNet) and *voxel architectures* (window Transformer) proposed in prior works. These results emphasize the significance of our novel Transformer-based point encoding module.

1) *Scaling point architecture:* The point architecture in PVTransformer encodes sparse point features into grid-like voxel features via Transformer architecture, shown in Figure 2. The key difference between PVTransformer and prior works lies in learning to aggregate point features through an attention mechanism rather than a heuristic pooling operation. This approach in PVTransformer alleviates the point-to-voxel information bottleneck introduced by PointNet, which has to aggregate information from over 30+ points through a simple max pooling.

Point architecture scaling space. We explore the scalability of the point architecture by scaling up our proposed

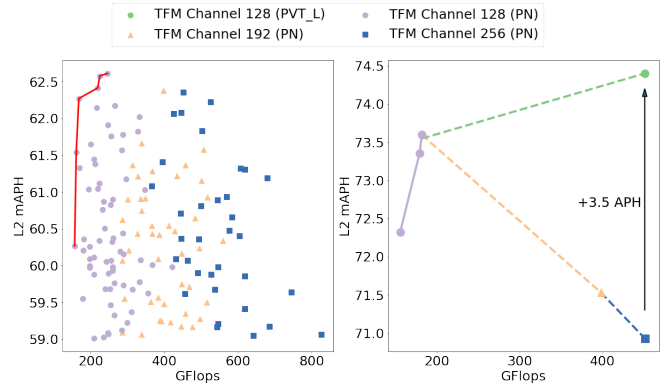


Fig. 5: **The voxel architecture has limited scalability when using PointNet (PN) to aggregate point features.** Right: Using Transformer to aggregate point features (PVT.L) is significantly better (green) compared to using PointNet and scaling only the channels in the voxel architecture to 256 (blue), a 3.5 mAPH L2 increase at a similar Flops. Left: Performance of random sampled voxel architectures from the search space (in Table V) after trained for 12.8 epochs. We observe that scaling the voxel architecture while using PointNet can lead to suboptimal performance. The Pareto curve (red curve) shows scaling voxel architecture channels from 128 to 192 and 256 channels leads to overfitting. Waymo Open Dataset validation set mAPH L2 on Vehicle and Pedestrian are reported.

FC Channel	[128, 192, 256, 320]
PointNet FC Depth	[2, 5, 7]
PVTransformer Depth	[FC, FC-PV, FC-PP-PV]

TABLE IV: **Point architecture and scaling space.** The point architecture for both PVTransformer_3f and PVTransformer_3f.L are colored in orange. **PointNet.** Two fully connected (FC) layers with channel size 128 are used as the seed architecture. PointNet is scaled either by increasing the first layer FC channels or adding more FC layers. **PVTransformer.** A single FC layer is used as the seed point architecture. The PVTransformer is scaled by adding a point-to-voxel aggregation layer (PV), as shown in Figure 3, and an additional point-to-point self-attention Transformer (PP) within the same voxel, as shown in Figure 2, 3.

PVTransformer architecture and the conventional PointNet architecture. The scaling space for point architecture is summarized in Table IV. The seed point architecture for PointNet is composed of two fully connected layers with hidden channel of [128, 128], as described in [27]. We explore two scaling strategies for PointNet, namely increasing the hidden channel size of the first layer or stacking more fully connection layers, as shown in Table IV.

For PVTransformer, the seed architecture consists of a fully connected layer to featurized the point feature to 128 dimensions, followed by the proposed point-aggregation module, as shown in Figure 3. To scale up the point archi-

Backbone		Head	
TFM Channel	[128, 192, 256]	TFM Channel	[128, 192, 256]
TFM Heads	[4, 8, 16]	TFM Heads	[4, 8, 16]
TFM MLP Expansion	[2, 4]	TFM MLP Expansion	[2, 4]
0.32m Num of Blocks	[2, 3, 4, 6]	Vehicle Num of Blocks	[1, 2, 4, 6]
0.64m Num of Blocks	[2, 3, 4, 6]	Pedestrian Num of Blocks	[1, 2, 4, 6]
1.28m Num of Blocks	[2, 3, 4, 6]		
5.12m Num of Blocks	[2, 3, 4, 6]		
10.24m Num of Blocks	[2, 3, 4, 6]		

TABLE V: **Voxel architecture and search space.** The parameters for PVTransformer_3f and PVTransformer_3f_L backbones are highlighted in blue and red respectively. The value is colored in orange when both backbones share the same parameter. We search the hyperparameters for each Transformer layer, including the channels size, number of heads and fully connected layer expansion ratio for both the backbone and detection heads. In addition, we search the number of repeated blocks for each Transformer layer at each resolution and for each detection head.

ture of PVTransformer, we introduce an additional point-point self-attention layer between the fully connected layer and the point-aggregation module, as shown in Figure 3.

In addition, we investigate the point architecture scaling using two types of Transformer backbones, a single-scale backbone and a multi-scale backbone. For the single-scale backbone, we follow [4] and use 6 layers of window Transformer at the same resolution 0.32m. For multi-scale backbone, we follow [27] and use 12 layers of window Transformer blocks with [2, 3, 2, 3, 2] layers at resolution [0.32m, 0.64m, 1.28m, 5.12m, 10.24m]. For both backbones, we use the same hidden channel of 128, 8 heads, and expansion ratio of 2 for all Transformer layers.

Superior scalability of PVTransformer. As shown in Figure 4, our PVTransformer demonstrates better scaling compared to PointNet. Our multi-scale PVTransformer improves upon single-layer PointNet from 72.4 mAPH L2 to 74.0 mAPH L2, a 1.6 mAPH increase. Notably, the scaling performance is not achievable with commonly used PointNet, as increasing the depth or hidden channel size leads to performance plateau and overfitting, as evidenced by the saturation and decline in performance beyond a certain scale.

2) *Scaling voxel architecture:* The voxel architecture predicts bounding boxes based on voxelized features, which includes both the backbone and detection heads architectures. To capture the intertwined nature of different components in the voxel architecture, we parameterize the voxel architectures, as shown in Table V, and random sample 150 architectures from the search space to evaluate the performance. To save cost, we train each architecture on the full training set but reduce the training steps to 1/10, a total of 12.8 epochs as a proxy setting. Here, we use two-layers PointNet as the point architecture. Our results show without the proposed PVTransformer, scaling voxel architecture is suboptimal.

Voxel architecture scaling is challenging. Our architecture studies, as shown in Figure 5, reveal that the model performance plateaus after scaling up the voxel architecture beyond 250 GFlops. Interestingly, we find that scaling Transformer channel size does not lead to better model

Method	mAPH L2	Vehicle AP/APH 3D		Pedestrian AP/APH 3D	
		L1	L2	L1	L2
PointNet-1f	68.1	79.0/78.5	70.5/70.1	81.5/73.5	73.5/66.1
PVTransformer_1f	71.1	80.9/80.4	72.5/72.1	83.7/77.6	75.9/70.1

TABLE VI: **PVTransformer generalizes to single frame settings.** WOD val. set 3D detection metrics are reported.

Method	mAPH L2	Vehicle AP/APH 3D		Pedestrian AP/APH 3D	
		L1	L2	L1	L2
PN averaged pool_3f	72.1	80.1/79.7	72.2/71.8	83.7/79.8	76.1/72.4
PN max pool_3f	72.8	80.9/80.4	72.8/72.4	84.4/80.7	76.8/73.2
PVTransformer_3f	74.0	81.9/81.3	73.8/73.4	85.3/81.8	78.0/74.6

TABLE VII: **PVTransformer outperforms other point aggregation methods.** WOD val. set 3D detection metrics are reported.

performance. The performance-computation Pareto frontier strictly lies in the channel size 128 architecture regime.

Upon retraining the selected voxel architectures on the Pareto frontier as well as two additional voxel architectures with Transformer channel size 192 and 256 for 128 epochs, we find wider voxel architectures are prone to overfitting, while a smaller 128 channel model has the best performance with 73.5 mAPH L2.

We further show our PVTransformer is more scalable by replacing PointNet with our attention aggregation module and scaling the architecture to 452 GFlops. Our PVT_L achieves a much higher 74.4 mAPH L2, a 3.5 mAPH increase, compared to scaling the Transformer channel size to 256 at a similar 451 GFlops, as shown in Figure 1, 5.

D. Generalize to the single frame setting

In addition to 3-frame temporal training, Table VI shows our PVTransformer significantly improves the single-frame setting, leading to a 3 APH improvement.

E. Ablation on different aggregation methods

As shown in Table VII, our PVTransformer outperforms averaged pooling based feature aggregation, demonstrating that naive averaging is suboptimal and learning to aggregate, as proposed in this work, is necessary.

V. CONCLUSION

In this work, we aim to enable better scalability for large-scale 3D object detectors and identify that the pooling-based PointNet introduces an information bottleneck for modern 3D object detectors. To address this limitation, we propose a new PVTransformer architecture, which uses an attention-based Transformer to aggregate point features into voxel features. We show that such a point-to-voxel Transformer is more expressive than a simple PointNet’s pooling layer, and thus achieve much better performance than previous 3D object detectors. Our PVTransformer significantly outperforms the prior art, such as SWFormer, and achieves new state-of-the-art results on the challenging Waymo Open Dataset.

REFERENCES

- [1] Alex Bewley, Pei Sun, Thomas Mensink, Dragomir Anguelov, and Cristian Sminchisescu. Range conditioned dilated convolutions for scale invariant 3d object detection. In *Conference on Robot Learning*, 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/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.
- [3] Yuning Chai, Pei Sun, Jiquan Ngiam, Weiyue Wang, Benjamin Caine, Vijay Vasudevan, Xiao Zhang, and Dragomir Anguelov. To the point: Efficient 3d object detection in the range image with graph convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2021.
- [4] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Embracing single stride 3d object detector with sparse transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8458–8468, 2022.
- [5] Lue Fan, Xuan Xiong, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Rangedet: In defense of range view for lidar-based 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2918–2927, 2021.
- [6] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.
- [7] Yihan Hu, Zhuangzhuang Ding, Runzhou Ge, Wenxin Shao, Li Huang, Kun Li, and Qiang Liu. Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 969–979, 2022.
- [8] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.
- [9] 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/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019.
- [10] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019.
- [11] Zhaoqi Leng, Guowang Li, Chenxi Liu, Ekin Dogus Cubuk, Pei Sun, Tong He, Dragomir Anguelov, and Mingxing Tan. Lidaraugment: Searching for scalable 3d lidar data augmentations. *arXiv preprint arXiv:2210.13488*, 2022.
- [12] Zhichao Li, Feng Wang, and Naiyan Wang. Lidar r-cnn: An efficient and universal 3d object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7546–7555, 2021.
- [13] Xinhai Liu, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. Point2sequence: Learning the shape representation of 3d point clouds with an attention-based sequence to sequence network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8778–8785, 2019.
- [14] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.
- [15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [16] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [17] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021.
- [18] 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.
- [19] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018.
- [20] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.
- [21] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.
- [22] Guangsheng Shi, Ruifeng Li, and Chao Ma. Pillarnet: High-performance pillar-based 3d object detection. *arXiv preprint arXiv:2205.07403*, 2022.
- [23] Shaoshuai Shi, Chaoxu Guo, Jihan Yang, and Hongsheng Li. Pvr-cnn: The top-performing lidar-only solutions for 3d detection/3d tracking/domain adaptation of waymo open dataset challenges. *arXiv preprint arXiv:2008.12599*, 2020.
- [24] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *arXiv preprint arXiv:2102.00463*, 2021.
- [25] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019.
- [26] 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/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.
- [27] Pei Sun, Mingxing Tan, Weiyue Wang, Chenxi Liu, Fei Xia, Zhaoqi Leng, and Dragomir Anguelov. Swformer: Sparse window transformer for 3d object detection in point clouds. *arXiv preprint arXiv:2210.07372*, 2022.
- [28] Pei Sun, Weiyue Wang, Yuning Chai, Gamaleldin Elsayed, Alex Bewley, Xiao Zhang, Cristian Sminchisescu, and Dragomir Anguelov. Rsn: Range sparse net for efficient, accurate lidar 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5725–5734, 2021.
- [29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.
- [30] Jiancheng Yang, Qiang Zhang, Bingbing Ni, Linguo Li, Jinxian Liu, Mengdie Zhou, and Qi Tian. Modeling point clouds with self-attention and gumbel subset sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3323–3332, 2019.
- [31] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.
- [32] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021.
- [33] 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 *CORL*, 2019.
- [34] 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*, pages 923–932. PMLR, 2020.
- [35] 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*, pages 4490–4499, 2018.
- [36] Zixiang Zhou, Xiangchen Zhao, Yu Wang, Panqu Wang, and Hassan Foroosh. Centerformer: Center-based transformer for 3d object detection. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 496–513. Springer, 2022.