

HVNet: Hybrid Voxel Network for LiDAR Based 3D Object Detection

Maosheng Ye Shuangjie Xu Tongyi Cao
DEEPROUTE.AI

{maoshengye, shuangjiexu, tongyicao}@deeperoute.ai

Abstract

We present Hybrid Voxel Network (HVNet), a novel one-stage unified network for point cloud based 3D object detection for autonomous driving. Recent studies show that 2D voxelization with per voxel PointNet style feature extractor leads to accurate and efficient detector for large 3D scenes. Since the size of the feature map determines the computation and memory cost, the size of the voxel becomes a parameter that is hard to balance. A smaller voxel size gives a better performance, especially for small objects, but a longer inference time. A larger voxel can cover the same area with a smaller feature map, but fails to capture intricate features and accurate location for smaller objects. We present a Hybrid Voxel network that solves this problem by fusing voxel feature encoder (VFE) of different scales at point-wise level and project into multiple pseudo-image feature maps. We further propose an attentive voxel feature encoding that outperforms plain VFE and a feature fusion pyramid network to aggregate multi-scale information at feature map level. Experiments on the KITTI benchmark show that a single HVNet achieves the best mAP among all existing methods with a real time inference speed of 31Hz.

1. Introduction

3D object detection is the task to recognize and locate objects in 3D scene. It serves as a fundamental task for 3D scene understanding with wide applications in robotics and self-driving cars. Recent approaches utilize various types of data, including monocular images [23, 27], stereo images [11, 21] and point cloud [10, 35] from LiDAR.

Unlike 2D images, point cloud data has some unique properties. The location of each point reflects the surface of physical objects in the real world. In bird eye view, data is scale invariant inherently and objects are naturally separated, which facilitates the detection of occluded object. Due to its accuracy and robustness, LiDAR has become the prevalent sensor for autonomous. However, the density of point cloud varies greatly. The volume density is proportional to the inverse of square distance to the sensor. This makes it difficult to adopt point cloud segmentation net-

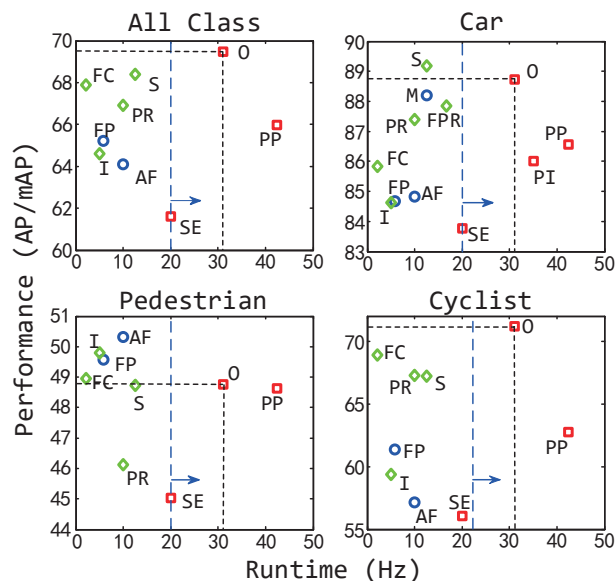


Figure 1. Bird's eye view performance vs speed in KITTI [5] test set. Multi-sensor methods drawn as blue circles; two-stage LiDAR based methods drawn as green diamonds and one-stage methods drawn as red squares. Methods on the right of the blue vertical bar are real-time. Methods should be at least 20Hz since onboard application should cover 360 degree rather than KITTI annotation at limited 90 degree. Drawn methods are FP: F-PointNet [20], AF: AVOD-FPN [9], M: MMF [13], I: IPOD [31], FC: F-ConvNet [26], S: STD [32], PR: PointRCNN [22], FPR: Fast Point R-CNN [2], SE: SECOND [28], PP: PointPillars [10], PI: PIXOR++ [29] and O: our HVNet. For PointPillars we use their runtime on PyTorch for a fair comparison. Our approach outperforms in most case, especially for mAP and Cyclist.

works or 2D image detection methods.

In order to handle these problems and utilize the advantages of LiDAR data, many approaches have been proposed recently (see Fig 1 caption). A noticeable line of methods combine 2D/3D voxelization with PointNet feature extractor in each voxel [10, 28, 35], which we call voxel feature encoding (VFE) methods. Using learnt point cloud features and 2D pseudo image feature map, these methods achieve both good accuracy and high inference speed. VFE methods generally contain three steps (Fig 2 VFE). **1. Voxelization:** the point cloud is assigned to a 2D voxel grid. **2. Voxel Feature Extraction:** a grid dependent point-wise feature is

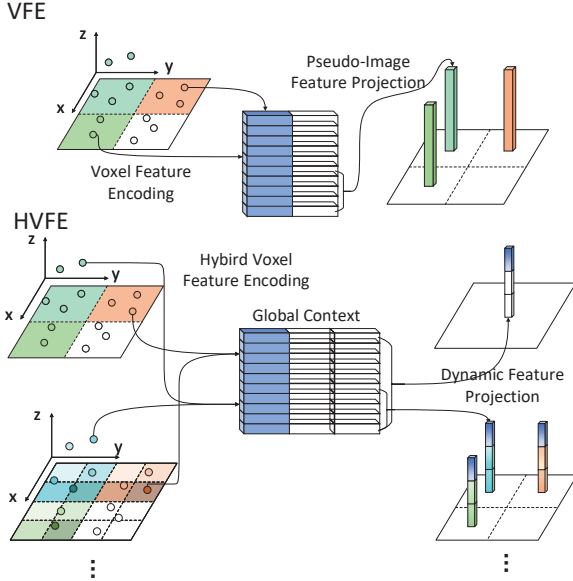


Figure 2. The voxel feature extraction strategy of VFE methods and our HVNet. Each point feature in VFE methods only contains one specified scale knowledge. As to HVNet, point features under hybrid scales are aggregated into a global context, then projected to features in dynamic target scales.

computed for each point and fed into a Pointnet style feature encoder. **3. Projection:** the point-wise feature is aggregated to the voxel-level feature and projected back to their original grid, forming a pseudo-image feature map.

A key parameter in VFE methods is the size of voxel. A smaller voxel captures finer geometry feature and better localizes the object but suffers from longer inference time. A coarser voxel leads to a smaller feature map and faster inference speed but has inferior performance, especially for small objects. In [10], a grid size of 0.12 gives the best result on KITTI dataset with a speed of 25 Hz. However, the inference space is limited to the 90 degree FOV in front with a 70 meter range. Other work uses multiple voxel scales to enhance the performance [24] of the network but the runtime is still dominated by the largest feature map.

In order to utilize fine-grained voxel feature as well as maintain runtime efficiency, we propose Hybrid Voxel Network (HVNet). HVNet decouples the feature extraction scale and the pseudo-image projection scale. Similar to [10], the HVNET contains three steps (Fig 2 HVFE): **1. Multi-Scale Voxelization:** We create a set of feature voxel scale and each point is assigned to multiple voxels. **2. Hybrid Voxel Feature Extraction:** for each scale, a voxel dependent feature is computed for each point and fed into the attentive voxel feature encoder (AVFE). Feature from each voxel scale is concatenated point-wise. **3. Dynamic Feature Projection** To project the feature back to a pseudo-image, we create another set of multi-scale projection voxels, which can be *independent* of the feature voxels. The multi-scale point-wise feature from the last step is aggregated according to the projection scale voxels, forming

multiple pseudo-image feature maps. By these strategies, the HVNet is able to be time-efficient. As an example, a set of feature voxel scale can be $\{0.1m, 0.2m, 0.4m\}$ and the projection voxel scale can be $\{0.4m\}$. For a $80m \times 80m$ area we end up with a feature map of size 200×200 . For a normal VFE structure, we need a feature map of size of 800×800 to capture small feature at $0.1m$ scale.

We evaluated our HVNet on the KITTI [5] dataset. HVNet achieves the state-of-the-art performance among all the LiDAR one-stage methods (Fig. 1). Besides, even when compared with methods of two stage and multi-sensor, our approach also outperforms in *mAP* and *Cyclist*. At the same time, HVNet runs at a real-time speed of 31Hz.

In summary, our contributions reside as follows:

- We propose an novel Hybrid Voxel Feature Encoder that fuses multi-scale VFE efficiently by decoupling the feature extraction voxels and the feature map projection voxels.
- We maintain all the voxels implicitly and perform the scale aggregation in the compact point-wise form to achieve high computation and memory efficiency.
- We propose an attentive VFE that achieves great improvement with little computation overhead.
- Extensive experiments are conducted on the KITTI to demonstrate the effectiveness of each component.

2. Related Work

2.1. 3D Object Detection

There are roughly two different lines for existing methods of 3D object detection with point cloud:

Multi-sensor based 3D object detection. MV3D [1] is a pioneering work which projects point cloud into bird eye view and front view. AVOD [9], inspired by MV3D, designs 3D object anchors and then aggregates corresponding feature maps by projection to generate a comprehensive representation for detection. F-PointNet [20] extends the 2D detections from image into corresponding frustums in the 3D space. MMF [13] is proposed to exploit multiple related tasks including depth completion and 2D object detection for accurate multi-sensor 3D object detection. However, although multiple sensors could provide extra information, the inference efficiency for these frameworks are relatively low. Besides, synchronization requirement between multiple sensors makes it hard to deploy.

Point cloud based 3D object detection. Voxel based methods [3, 9, 30, 25] share a main idea to project sparse point cloud into compact representation. VoxelNet [35] employs VFE layers based on PointNet for 3D space points characterization. SECOND [28] utilizes Sparse Convolution [6] to accelerate VoxelNet and improve performance. Based on SECOND, PointPillars [10] applies voxelization only in XY direction to form pseudo-image featurization for

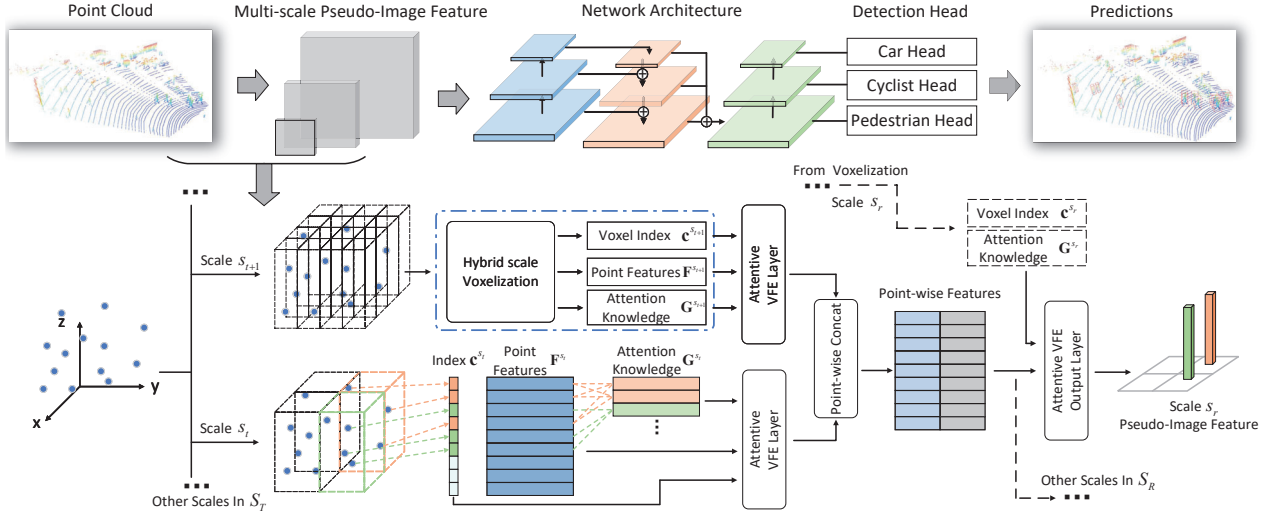


Figure 3. The first line is the whole structure of HVNet. The second line is the Hybrid Voxel Feature Extractor architecture, consisting of four parts: a) multi-scale pre-voxelization in x - y plane; b) multi-scale point cloud data encoding by multi-stream AVFE layers parallelly; c) encoded features aggregation; d) dynamic feature projection by AVFEO layers with aggregated features and target scale knowledge.

point cloud. However, point number in each voxel is limited, leading to information loss. In this work, we propose a novel encoder for lossless 3D representation.

There have been many two-stage works [31, 26, 32, 22, 2] recently. Fast Point R-CNN [2] applies two-stage framework exploiting volumetric representation for initial predictions and raw point cloud for refinement. STD [32] first generates proposals and compact proposal features with PointsPool in the first stage, and then predicts 3D bounding box in the second stage. These methods have high accuracy due to better recall rate, but long inference time (lower than $20Hz$). Therefore, We focus on one-stage unified framework with both time efficiency and accuracy in this paper.

2.2. Multi-scale feature aggregation

Multi-scale strategy [14, 13, 2] has been proved to be effective to 3D object detection. ContFuse [14] uses continuous convolution to aggregate multi-scale feature maps from different ResNet Block [7]. Voxel-FPN [24] employs a RPN-FPN module [15] to aggregate different voxel features generated by voxelization from various voxel sizes. However, the voxel feature maps have strict correspondence to each voxel size, making the multi-scale learning inadequacy and inflexible. Instead, our HVNet, a novel multi-scale zoomable aggregation strategy, is able to zoom scales into a many-to-many relationship.

3. Approach

In this paper we describe our one-stage 3D object detector HVNet. The overall network is presented in the first line of Figure 3, which consists of three main stages: 1) hybrid voxel feature extraction that projects the raw point cloud into pseudo-image feature; 2) 2D convolutional middle layers for multi-scale feature fusion; and 3) a detection head that predicts 3D boxes for different classes separately.

3.1. Hybrid Voxel Feature Extractor

The three main steps included in the hybrid voxel feature extractor (HVFE), i.e., hybrid scale voxelization, hybrid voxel feature encoding and dynamic feature projection. Furthermore, to decouple the hybrid voxel scale and pseudo-image feature projection scale for a more flexible balance between speed and accuracy, the global context for hybrid voxel representation is introduced under the voxel-wise attention guidance, whose detail will be described in Sec. 3.2. Overall structure of this stage is shown in Fig. 3.

Point cloud representation. A point cloud would be represented by an unordered point set $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N\}$ with $\mathbf{p}_i \in \mathbb{R}^d$, which includes $\mathbf{x}_i \in \mathbb{R}^3$ for point coordinate (x, y, z) and $\mathbf{f}_i \in \mathbb{R}^{d-3}$ for associated feature vector. The physical dimension of the scene that we detect objects on is defined as $L \times W \times H$, from the minimum $(x_{min}, y_{min}, z_{min})$ to the maximum $(x_{max}, y_{max}, z_{max})$.

Hybrid scales voxelization. HSV is a way of voxelization that assigns points to evenly spaced grid of voxels. The assignment phase in prior methods, like VoxelNet [35] and PointPillar [10], is accompanied by buffer allocation with a fixed size. Points will be dropped when the buffer capacity for a voxel is exceeded, causing randomness and information loss. Instead, we maintain only the voxel index of each point. Assume that the point cloud is discretized into numerous voxels $\mathbf{V} = \{\mathbf{v}_j | j = 1, \dots, N_v\}$ with resolution of $v_L \times v_W \times v_H$ per voxel. v_H is set to H , followed by [10]. The point-wise index \mathbf{c}_i for target voxel scale size s is recorded as a cursor to which voxel it lies in, given as:

$$\mathbf{c}_i^{(s)} = \left\lfloor \frac{(x_i - x_{min})}{v_L s} \right\rfloor \left\lfloor \frac{W}{v_W s} \right\rfloor + \left\lfloor \frac{(y_i - y_{min})}{v_W s} \right\rfloor. \quad (1)$$

Only the point to voxel mapping relation is needed in our approach, unlike MVF [34] which need bi-directional relationships. In this way, only tensor \mathbf{c} are allocated during

voxelization stage. Besides, we keep all points retained inside a voxel other than methods, which also becomes the precondition for voxel attention.

Hybrid voxel feature extraction. The HVFE module provides a novel way to create fine-grained point cloud feature with little computation cost. Take PointPillar [10] as an example (Fig 2), previous voxel based methods are limited to the fixed voxel scale s , where each encoded point feature only contains the scale knowledge of s . Besides, the size of final projected features is the same as the corresponding voxel size, which means that the computation complexity will rise rapidly with the addition of fine-grained voxel scale. Assume that the projected pseudo-image feature $\mathbf{I}^s \in \mathbb{R}^{N_W \times N_L \times N_H}$, where $N_L = \lfloor L/(v_L s) \rfloor$, $N_W = \lfloor W/(v_W s) \rfloor$ and N_H is the projection output dimension. When s reduces n times, $N_L \times N_W$ will increase n^2 times.

Shown in Fig 3, the HVFE module takes point features of a set of voxel scales $S_T = \{s_t | 1 < t < N_T\}$ as input, encoding them as unified point-wise feature \mathbf{H} with information from different voxel scales. Drawn in Fig. 3, firstly the raw point cloud is hybrid voxelized to get cursor set $\mathbf{c}^{S_T} = \{\mathbf{c}_i^{(s_t)} | s_t \in S_T, i \in [1, N]\}$. Secondly, attention feature set \mathbf{G}^{S_T} whose tensor dimensionality is same as the pointnet features $\mathbf{F} \in \mathbb{R}^{N \times q}$ are obtained. Thirdly, the AVFE layer takes \mathbf{F} , \mathbf{G}^{S_T} and \mathbf{c}^{S_T} together to encode raw data into voxel scale feature set \mathbf{H}^{S_T} specified to each voxel scale s with $\mathbf{H}^{s_t} \in \mathbb{R}^{N \times 2q}$. All the AVFE layers for different scales share the same parameter. After that, $\mathbf{H} \in \mathbb{R}^{N \times e}$ is aggregated by concatenation of the \mathbf{H}^{S_T} , where $e = 2qN_T$. It's validated by experiments that only limited N_T of input multi-scale features is able to bring enough information to retrieve features of unseen scales out of S_T , utilizing the dynamic feature projection.

Dynamic feature projection. The global context \mathbf{H} is a high level feature containing information of hybrid voxel scales. Therefore, the feature projection scales $S_R = \{s_r | 1 < r < N_R\}$ of pseudo-image features that are out of S_T can be projected dynamically and flexibly with \mathbf{H} . In order to restore information at scales s_r from \mathbf{H} , some scale-related information needs to be used as guidance in the projection, containing point-wise index \mathbf{c}^{s_r} and the attention knowledge \mathbf{G}^{s_r} . The attentive VFE output (AVFEO) layer takes \mathbf{H} , \mathbf{c}^{s_r} and \mathbf{G}^{s_r} to retrieve the voxel-wise feature and then project it to a pseudo-image feature \mathbf{I}^{s_r} . Same as the AVFE layer, all AVFEO layers share the same parameters. In the dynamic feature projection, \mathbf{G}^{s_r} is the key for the scale-wise feature retrieval, guiding the AVFEO layer to be attentive to project \mathbf{H} into corresponding scale s_r . Besides, \mathbf{c}^{s_r} is used for the efficient graph-based implementation of the AVFEO layer (Sec. 3.2).

3.2. Attentive Voxel Feature Encoding

Standard VFE layer. Traditional VFE [35, 10] first voxelizes data with point number in each voxel limited and al-

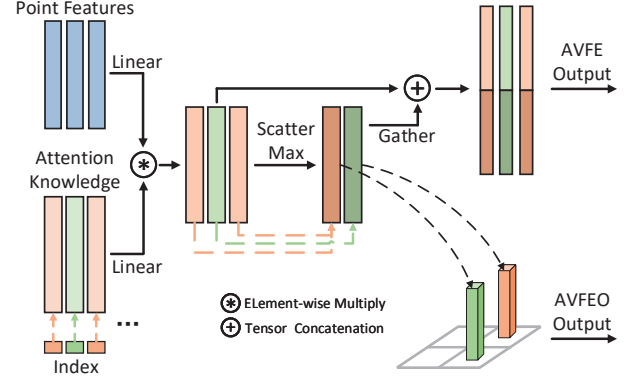


Figure 4. AVFE and AVFEO layers. The index tensor is used in indexing operation: Gather and Scatter.

locates fixed preassigned tensors for each voxel. After that, linear transforms and max pooling are conducted in each voxel to aggregate features. Then pseudo-image features are collected according to the physical location of voxels.

While VFE method provides a simple means of integrating sparse point cloud, it has several drawbacks:

- There are points dropped stochastically during voxelization, leading to tremendous information losses and instability prediction behaviours;
- Zero padding is conducted when point number is lower than limitation, which can greatly enlarge amounts of computation and memory consumption;

Voxel-wise attention feature. As the expansion of raw point-wise features \mathbf{p}_i , attention knowledge $\mathbf{g}_j \in \mathbb{R}^q$ for voxel \mathbf{v}_j is obtained by voxel-wise feature aggregation, which guides our feature encoding network to concern more about the interrelationships within a voxel. Given index \mathbf{c} , the attention knowledge \mathbf{g}_i of point \mathbf{p}_i is represented as:

$$\mathbf{g}_i = \left(\left(\mathbf{x}_i - \frac{\sum_{U_j} \mathbf{x}_k}{N_{U_j}} \right) \oplus \mathbf{f}_i \right) \oplus \frac{\sum_{U_j} \mathbf{p}_k}{N_{U_j}}, \quad (2)$$

where \oplus represents the tensor concatenate. The collection U_j , referred to the index set of points that lies in the same voxel as \mathbf{p}_i , is denoted as $U_j = \{k | 0 < k < N \& \mathbf{c}_k = \mathbf{c}_i\}$. The collection process of U has a lot of loops, making this aggregation step unbearably slow. We will discuss the efficient index-based implementation in Sec. 3.3.

Shown in the Eq. 2, the attention knowledge for a point \mathbf{p} in a voxel \mathbf{v} is composed of two parts: a) the mean of features from all the points lie in \mathbf{v} ; and b) the feature of \mathbf{p} whose point coordinate \mathbf{x} is subtracted by the average coordinate of \mathbf{v} . The attention feature \mathbf{g} participates in both encoder and decoder as the bridge information.

Learning to be flexible and attentive. Shown in Fig. 4, the input point features \mathbf{F} and the attention knowledge \mathbf{G} are conducted by linear layers, followed by a element-wise matrix multiply. After that, multiplied features \mathbf{M} in the same voxel are aggregated by max pooling layer to form the voxel-wise features $\mathbf{G}' \in \mathbb{R}^{N \times q}$. For AVFE layer, the output is the point-wise concatenation of \mathbf{M} and \mathbf{G}' . For

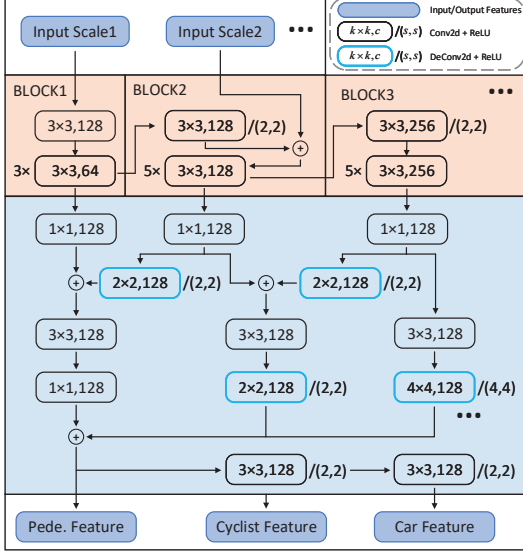


Figure 5. The backbone network architecture of HVNet. Multi-scale features are first fused shallowly in the main stream network, and then fused deeply in the proposed FFPN network.

AVFEO layer, each feature of \mathbf{G}' is projected to a pixel $(x_I^{s_r}, y_I^{s_r})$ in pseudo image \mathbf{I}^{s_r} , given as:

$$x_I^{s_r} = \mathbf{c}_j^{s_r} \left\lfloor \frac{W}{v_W s_r} \right\rfloor, y_I^{s_r} = \mathbf{c}_j^{s_r} \bmod \left\lfloor \frac{W}{v_W s_r} \right\rfloor, \quad (3)$$

where s_r is the target scale, $\lfloor \cdot \rfloor$ denotes exact division and \bmod denotes the module operation.

3.3. Efficient Index-based Implementation

The Hybrid Voxel Feature Extraction module is an index-based learning system, where the irregular graph data (point cloud) is grouped by physical correlation. In order to handle the sparse structures, we propose HSV operation to transform sparse matrix into dense matrix and corresponding indexes. Verified by experiments, the efficiency of the HVNet then hinges heavily on how well The index strategy of U and the parallel stream processing schedules. Therefore, the key index-based propagation operators, Scatter and Gather, is implemented on GPUs.

The Gather conducts the sparse point data propagation within voxel and behave as tensor slice according to the cursor vector \mathbf{c} . For implementation, tensor slice operation in PyTorch [19] is fast enough for Gather.

The Scatter manipulates all values from source tensor into output at the indices specified in \mathbf{c} along a given axis. In our approach, Scatter Mean is used in voxel wise attention and Scatter Max is used in AVFE and AVFEO layers. Take Scatter Max as an example: $\text{out}_i = \max(\text{out}_i, \max(\{\text{src}_j | \mathbf{c}_j = i, j \in \mathbf{c}\}))$, where out and src are output and input respectively, \mathbf{c} is the ‘group-index’ that references the location of src . In the implementation of Scatter, atomic lock of GPU Global Memory is used to ensure the consistency of $\arg \max$ results for reproducibility.

3.4. Backbone Network

We use a 2D convolutional backbone network for further detection. The backbone network has two main parts: 1) the main stream network that aggregates multi-scale features \mathbf{I}^{S_R} ; 2) the FFPN network that refines the feature map and generates multi-class features in the same time.

Main stream network with multi-scale feature aggregation. The main stream is shown in the orange part in Fig. 5, characterized by a concatenation of several blocks. In Block_2 and Block_3, the first layer has a 2×2 stride to reduce spatial resolution of features. Given pseudo-image feature set \mathbf{I}^{S_R} where $s_{r+1} = 2s_r$, \mathbf{I}^{s_1} is taken as input of Block_1 and \mathbf{I}^{s_2} is aggregated in Block_2 by tensor concatenation with output of the first layer in Block_2. More scales can be added by the similar way. We take the output of the last three blocks $\mathbf{B}_1 \in \mathbb{R}^{N_W \times N_L \times C_1}$, $\mathbf{B}_2 \in \mathbb{R}^{\frac{N_W}{2} \times \frac{N_L}{2} \times C_2}$ and $\mathbf{B}_3 \in \mathbb{R}^{\frac{N_W}{4} \times \frac{N_L}{4} \times C_3}$ as the input of FFPN.

Feature fusion pyramid network. The FPN [15] has been proved to be a valid means for multi-scale feature embedding. Due to the sparse data distribution in point cloud and the small resolution of \mathbf{I}_R , the FPN structure plays a more important role in our approach. Therefore, an enhanced feature fusion pyramid network is proposed, whose main structure is shown in Fig. 3. Instead fusing features from top to bottom layer by layer in [15], we first concatenate features at the smallest scale to obtain an intermediate feature \mathbf{B}_f , represented as:

$$\mathbf{B}_f = \sum_{i=1}^{N_B} \Psi_i (\mathbf{B}_i + 1_{i < N_B} \Gamma_{i+1} (\mathbf{B}_{i+1})), \quad (4)$$

where $+$ means tensor concatenate, 1 means the indicator function, N_B is the number of \mathbf{B}_i , Γ_i represents the deconv function of each input feature map \mathbf{B}_i for scale alignment and Ψ_i denotes the conv function before being concatenated together. The class-specific pyramid features are given as $\mathbf{B}_o^{(i)} = \Upsilon_i (\mathbf{B}_f)$, where Υ_i denotes the conv layers with various stride. Compared with F-SSD [12], we fuse features in two stages: a) layer by layer fusion in Eq. 4 and b) down-sample convolutional chain. Furthermore, within one forward propagation, class-specific pyramid features are able to be obtained, where $\mathbf{B}_o^{(1)}$ for *Pedestrian* class, $\mathbf{B}_o^{(2)}$ for *Cyclist* class and $\mathbf{B}_o^{(3)}$ for *Car* class respectively.

3.5. Detection Head and Loss Design

We use the detection head in the SSD [17] to detect 3D objects as [10]. In our setting, the positive anchors are selected by matching to the ground truth by Rotated Intersection over Union (RIoU) [4, 33] in bird eye view (BEV). Each pyramid feature $\mathbf{B}_o^{(i)}$ is wrapped by three parallel branches which are 3×3 convolution layers $\phi_c^{(i)}$, $\phi_l^{(i)}$ and $\phi_h^{(i)}$ to get classification probability, location offsets and height regression respectively, whose output channel numbers are $N_{\text{anc}} \times N_{\text{cls}}$, $N_{\text{anc}} \times 8$ for location corner offsets in

Method	Speed (Hz)	Input		Car BEV AP (%)			Pedestrian BEV AP (%)			Cyclist BEV AP (%)		
		L	I	Easy	Moder.	Hard	Easy	Moder.	Hard	Easy	Moder.	Hard
F-PointNet [20]	5.9	✓	✓	91.17	84.67	74.77	57.13	49.57	45.48	77.26	61.37	53.78
AVOD-FPN [9]	10	✓	✓	90.99	84.82	79.62	58.49	50.32	46.98	69.39	57.12	51.09
MMF [13]	12.5	✓	✓	93.67	88.21	81.99	N/A	N/A	N/A	N/A	N/A	N/A
IPOD [31]	5	✓		89.64	84.62	79.96	60.88	49.79	45.43	78.19	59.40	51.38
F-ConvNet [26]	2.1	✓		91.51	85.84	76.11	57.04	48.96	44.33	84.16	68.88	60.05
STD [32]	12.5	✓		94.74	89.19	86.42	60.02	48.72	44.55	81.36	67.23	59.35
PointRCNN [22]	10	✓		92.13	87.39	82.72	54.77	46.13	42.84	82.56	67.24	60.28
Fast Point R-CNN [2]	16.7	✓		90.87	87.84	80.52	N/A	N/A	N/A	N/A	N/A	N/A
SECOND [28]	20	✓		89.39	83.77	78.59	55.99	45.02	40.93	76.50	56.05	49.45
HRI-VoxelFPN [24]	50	✓		92.75	87.21	79.82	N/A	N/A	N/A	N/A	N/A	N/A
PointPillars [10]	42.4	✓		90.07	86.56	82.81	57.60	48.64	45.78	79.90	62.73	55.58
PIXOR++ [29]	35	✓		93.28	86.01	80.11	N/A	N/A	N/A	N/A	N/A	N/A
Ours	31	✓		92.83	88.82	83.38	54.84	48.86	46.33	83.97	71.17	63.65

Table 1. Performance of bird eye views on KITTI test set. Here ‘L’ denotes LiDAR input and ‘I’ denotes RGB image. We compare with detectors on the KITTI leaderboard evaluated by 40 recall positions. Methods are divided into three types: LiDAR & image, two-stage LiDAR only and one-stage. The bold results means the best in all methods and the blue results means the best among one-stage methods.

BEV and $N_{\text{anc}} \times 2$ for z center and height.

Different from most voxel-based methods [10, 35, 24] that predict center x, y and w, h, l, θ , HVNet utilizes location corner offsets relative to anchors in BEV as localization objective, represented as $\Delta\vec{\text{cor}} = \{\Delta\vec{x}_i, \Delta\vec{y}_i | i = 1, 2, 3, 4\}$. Δcor is a vector of \mathbb{R}^8 during propagation. Suppose that the location branch ϕ_l predicts the offset Δcor , then the localization loss is given by $\mathcal{L}_{\text{loc}} = \sum \text{SmoothL1}(\Delta\vec{\text{cor}} - \Delta\text{cor})$.

As to the classification branch ϕ_c , given the class probability p_a of an anchor, focal loss [16] is used to handle the unbalance between positive and negative samples, represented as $\mathcal{L}_{\text{cls}} = -\alpha(1 - p_a)^\gamma \log(p_a)$. Given prediction z, h from ϕ_h , the vertical loss is denoted as $\mathcal{L}_h = \sum_{b \in (z, h)} \text{SmoothL1}(\Delta b)$.

Therefore, the total loss is given by

$$\mathcal{L} = \frac{1}{N_{\text{pos}}} (\lambda_{\text{loc}} \mathcal{L}_{\text{loc}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_h \mathcal{L}_h). \quad (5)$$

4. Experiments

HVNet is evaluated on the challenging KITTI benchmark [5]. We first introduce the setup details of HVNet in Sec. 4.1. In Sec. 4.2, we compare HVNet with state-of-the-art methods. A detailed ablation study is also provided in Sec. 4.3 to verify the validity of each component.

4.1. Setup

Dataset. KITTI dataset [5] consists of 7481 training images and 7518 test images as well as the corresponding point cloud with categories *Car*, *Pedestrian* and *Cyclist*.

Metric. KITTI’s metric is defined as average precision (AP) over the 40 recall positions on the PR curve [4]. Labels are divided into three subsets (*Easy*, *Moderate*, *Hard*) on the basis of object size, occlusion and truncation levels. The leaderboard rank is based on results of *Moderate* difficult.

Experiment details. The physical detection range is limited within minimum (0, −32, −3) and maximum

(64, 32, 2). The size of a voxel is $v_L = 0.2$, $v_W = 0.2$, thus the resolution of BEV feature is 320×320 . In the encoder and decoder, scale set $S_T = \{0.5, 1, 2\}$ and $S_R = \{1, 2, 4\}$. Besides, feature dimension is $q = 64$ for \mathbf{H}^{st} and $N_H = 128$ for \mathbf{I}^{st} . Anchor size is designed as [0.8, 0.8, 1.7] for *Pedestrian*, 0.8, 1.8, 1.5 for *Cyclist* and [1.7, 3.5, 1.56], [2.0, 6.0, 1.56] for *Car*. Each class has the same anchor orientation in $[0, \pi/4, \pi/2, 3\pi/4]$. In the training phase, we choose anchors that have RIoU with ground truth larger than [0.35, 0.35, 0.5] for *Pedestrian*, *Cyclist* and *Car* respectively as positive samples, and those lower than [0.25, 0.25, 0.35] as negative samples. As to the test phase, prediction score threshold is set to 0.2, and the rotated NMS [18] threshold is set to [0.02, 0.02, 0.4]. In the loss design, α and γ for focal loss are set to [0.75, 0.75, 0.25] and [2., 2., 2.] respectively. Three loss weights are $\lambda_{\text{loc}} = 1$, $\lambda_{\text{cls}} = 1$ and $\lambda_h = 1.5$. HVNet is trained for 70 epochs with Adam [8] optimizer, the initial learning rate is $\text{lr} = 2e^{-4}$ with weight decay $1e^{-4}$. We first employ warmup strategy [7] with 300 warmup iterations and 1./3 warmup ratio. Besides, lr decays in a ratio of 0.1 in the 40-th and 60-th epoch.

3D data augmentation. Global random flipping, rotation, scaling and translation are first applied to the whole point cloud data, where the flipping probability is set to 0.5, the rotation angle ranges are uniform distribution from $[-\pi/2, \pi/2]$, the scaling ratio is between [0.95, 1.05] and the location translation obeys normal distribution with mean 0 and standard deviation [0.2, 0.2, 0.2] for (x, y, z) . Followed by SECOND [28], several new boxes from ground truth and corresponding points in other frames, where 8 for *Cyclist*, 8 for *Pedestrian* and 15 for *Car*, will be fixed into current training frame except boxes which have physical collision with boxes in the current frame.

4.2. Experimental Results

Quantitative Analysis. We compare with 3D object detectors in three types: a) LiDAR & image based approaches;

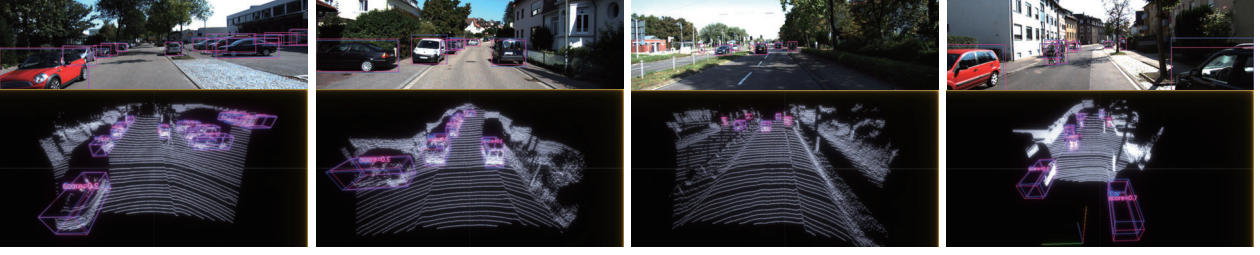


Figure 6. Qualitative analysis on KITTI validation set with Kitti Viewer [28]. We show 3D bounding boxes on point cloud along with projected 2D bounding boxes on image. In each image, blue boxes indicates the ground truth, red boxes indicates detections by HVNet.

Method	Car BEV AP(%)			Car 3D AP(%)		
	Easy	Moder.	Hard	Easy	Moder.	Hard
MV3D [1]	86.55	78.10	76.67	71.29	62.68	56.56
F-PointNet [20]	88.16	84.02	76.44	83.76	70.92	63.65
VoxelNet [35]	89.60	84.81	78.57	81.97	65.46	62.85
SECOND [28]	89.96	87.07	79.66	87.43	76.48	69.10
PointPillars [10]	N/A	87.7	N/A	N/A	77.4	N/A
Ours	93.29	88.90	87.72	87.21	77.58	71.79

Table 2. Performance (AP) of BEV and 3D on KITTI validation set for *Car*. Our method achieves the state of the art in most case.

Classes	Tasks	Easy	Moder.	Hard
<i>Car</i>	2D	96.50	92.04	89.34
	BEV	76.53	70.75	68.37
	3D	69.13	64.81	59.42
<i>Pedestrian</i>	2D	93.90	80.54	77.41
	BEV	89.98	77.13	72.27
	3D	87.21	73.75	68.98

Table 3. Performance on KITTI validation set for *Pedestrian*, *Cyclist* in 2D, BEV and 3D tasks; and *Car* in 2D task.

b) LiDAR only two-stage approaches and c) one-stage approaches, shown in three columns respectively in Tab. 1. Most methods in a) and b) are relatively slow for inference. From the table we see that HVNet outperforms all other approaches in *mAP* and *Cyclists*. HVNet also achieves attractive performance for *Car* and *Pedestrian* under a real-time runtime, even when compared with two-stage approaches. Among one-stage approaches, HVNet achieves the state of the art in *Car* and *Cyclist*, leading the second best HRI-VoxelFPN [24] and PointPillars [10] by over 1.61% and 8.44% respectively in *moderate*. More details for our test results are in KITTI [leaderboard](#). Furthermore, we draw the performance vs. speed figure in Fig. 1 according KITTI leaderboard to have more intuitive comparison.

Only a few methods expose the results on the validation set. The comparison results for *Car* are reported in Tab. 2. Among methods that expose results, our approach achieves the best performance in both BEV and 3D tasks. As almost no currently published method presents validation results for *Pedestrian* and *Cyclist*, we show only our validation results for these two classes in Tab 3 in all the three tasks: 2D, BEV and 3D. Overall, evaluation both on test and validation set shows that our approach can produce high-accuracy detections with a fast inference speed. Besides that, we con-

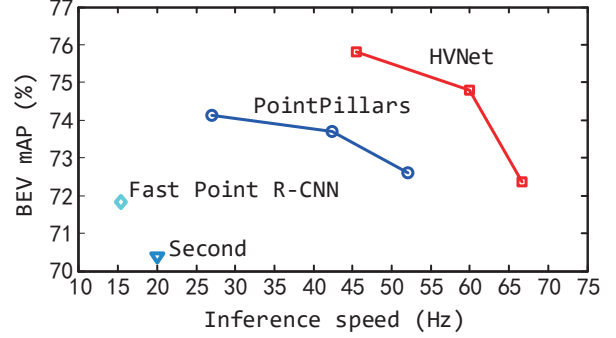


Figure 7. Voxel scale study on BEV of KITTI validation set. For PointPillar we use our own implementation in Pytorch. Blue circle shows the results of PointPillar. We choose voxel scale at $\{0.16^2, 0.20^2, 0.24^2\}m^2$. Red rectangle illustrates results from Hybrid Voxel Feature Extractor. From left to right, we use voxel scales at $\{0.1, 0.2\}m$, $\{0.2, 0.3\}m$, $\{0.2, 0.4\}m$, feature projection scales at $0.2m$, $0.3m$, $0.4m$ respectively

duct experiments on our Hybrid Voxel Feature Extractor architecture at different settings, compared with PointPillar through different grid size in Fig. 7. We can easily see the power of our architecture. Compared with best result from PointPillar at grid size $0.16m$, our model show great advantages at both *mAP* and inference speed. Even at a coarse scale for feature projection scale of grid size $0.4m$, our model can achieve comparable results of PointPillar at grid size $0.24m$ and save a lot of runtime cost.

Qualitative Analysis. We present several 3D detections on the KITTI validation set along with projected 2D bounding boxes on 2D image in Fig. 6. HVNet can produce high-quality 3D detections for all classes via a single pass of the network. Moreover, good detections are also given by HVNet for scenes that have point cloud occlusion or strongly dense objects. Generally, these qualitative results demonstrate the effectiveness of our approach.

4.3. Ablation Studies

Multiple components study. To analyze how different components contribute to the final performance, we conduct an ablation study on KITTI validation set. We use the BEV mean AP of three categories (*Car*, *Pedestrian* and *Cyclist*) as the evaluation matrix, shown in Tab. 4. Our baseline model, shown in the first line of Tab. 4, is a detector with a single scale feature extractor and a SSD [17] like

Model	voxel scale S_T			projection scale S_R			AVFE	FPN	FFPN	α	BEV mAP (%)		
	0.5S	1S	2S	1S	2S	4S					Easy	Moder.	Hard
Baseline	✓			✓						0.25	79.39	71.7	66.77
+ Attention	✓			✓			✓			0.25	81.36	73.76	69.18
+ FPN	✓			✓			✓	✓		0.25	81.84	74.34	69.83
+ FFPN	✓			✓			✓		✓	0.25	82.39	75.18	70.78
* Focal loss	✓			✓			✓		✓	0.75	82.82	75.71	71.54
+ Scale 2*	✓	✓		✓			✓		✓	0.75	84.34	76.88	73.01
+ Scale 2	✓	✓		✓	✓		✓		✓	0.75	85.26	77.88	74.46
+ Scale 3	✓	✓	✓	✓	✓	✓	✓		✓	0.75	86.31	78.59	75.08

Table 4. Ablation study on KITTI validation set for the mAP . 0.5S means the scale of 0.5. Δ mAP denotes the change in *Moderate* mAP compared with the corresponding controlled experiment. The maximum improvement is achieved by increasing S_T and S_R .

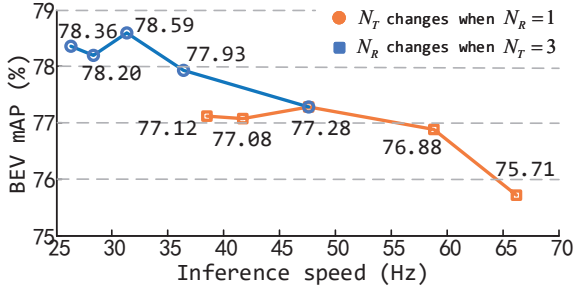


Figure 8. Ablation study on KITTI validation set for hybrid scale. The jacinth line means N_T changes from 5 to 1 when $N_R = 1$. The blue line means N_R changes from 5 to 1 when $N_T = 3$.

backbone. Besides, the VFE layers [35] replace the AVFE and AVFEO layers. It's given that adding attention to VFE layer is able to bring a 2.06 mAP gain in BEV *Moderate* compared with baseline. FPN is effective with a 0.58 gain but the proposed FFPN bring a larger improvement of a 1.42 gain. We also adopt focal loss for classification. However, the default α results in the confidence score degradation for *Pedestrian* and *Cyclist*. Therefore, we change $\alpha = 0.75$ by experiments. As to the Hybrid Voxel Feature Extractor, increasing voxel/feature projection scale number to 2 gives the maximum performance boost of **2.17** in mAP. Furthermore, keeping going up to 3 scales gives another 0.71 gain.

Hybrid Voxel Feature Extractor. Given that the Hybrid Voxel strategy plays an important role, it is important to make S_T and S_R enough for feature encoding while not consuming much computation. Thus, we conduct a series of experiments with various scale number. Note that as the scale number increases, the block number in backbone in Fig. 5 increases as well. Shown in Fig. 8, it's a nice trade-off between speed and performance when $N_T = N_R = 3$, demonstrated that the effectiveness of scale projection between S_T and S_R . Furthermore, we visualize the \mathbf{H}^{S_T} with/without attention in Fig. 9, which shows that the activation of target region is greater with attention strategy.

VFE [10]	HSV	Multi-class Head	Time (ms)
✓	×	×	34
×	✓	×	32
×	✓	✓	35

Table 5. Inference speed for HVNet with different module.

Inference speed. The inference time of HVNet is 32ms

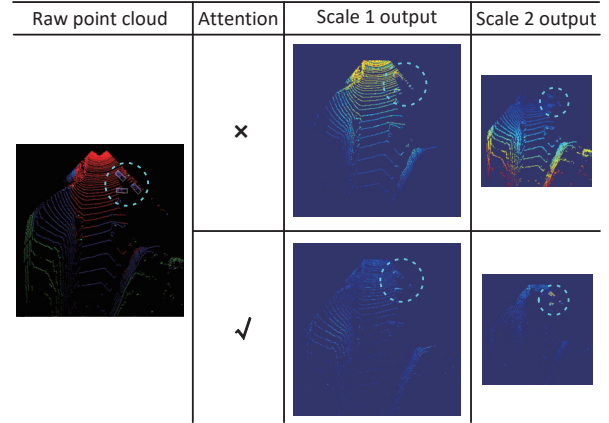


Figure 9. Multi-scale features \mathbf{H}^{S_T} with/without attention. With attention mechanism, our output feature map can suppress the background area and enhance the shape feature of objects

for a single class in average on a 2080Ti GPU where the Hybrid Voxel Feature Extractor takes 12ms, the backbone network takes 11ms and the head with NMS takes the left 9ms. The required time for each module is changed with the number of input points. In our approaches, HSV and index-based implementation are proposed to accelerate the feature encoding, whose effectiveness is shown in Tab. 5. We employ the VFE layer in PointPillars [10] as baseline. Utilize the HSV and index-based implementation save 2ms in average. Furthermore, the head in our model only takes extra 3ms for extending multi-class detection in one forward.

5. Conclusion

In this work, we propose HVNet, a novel one-stage 3D object detector. HVNet aggregates hybrid scale voxel grids into unified point-wise features, and then projects them into different scale pseudo-image features under the guidance of attention knowledge. The key to HVNet is that it decouples the feature extraction scales and the pseudo-image projection scales. Further more, a backbone with feature fusion pyramid network takes pseudo-images and fuses features to generate compact representations for different categories. Experimental studies show that our method achieves state-of-the-art mAP with a real-time speed.

References

- [1] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 7
- [2] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Fast Point R-CNN. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 3, 6
- [3] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner. Vote3Deep: Fast object detection in 3D point clouds using efficient convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1355–1361, 2017. 2
- [4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 5, 6
- [5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1, 2, 6
- [6] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 6
- [8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [9] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander. Joint 3D Proposal Generation and Object Detection from View Aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8, 2018. 1, 2, 6
- [10] 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*, pages 12697–12705, 2019. 1, 2, 3, 4, 5, 6, 7, 8
- [11] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo R-CNN Based 3D Object Detection for Autonomous Driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7644–7652, 2019. 1
- [12] Zuoxin Li and Fuqiang Zhou. Fssd: feature fusion single shot multibox detector. *arXiv preprint arXiv:1712.00960*, 2017. 5
- [13] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-Task Multi-Sensor Fusion for 3D Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7345–7353, 2019. 1, 2, 3, 6
- [14] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep Continuous Fusion for Multi-Sensor 3D Object Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 641–656, 2018. 3
- [15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3, 5
- [16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 6
- [17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 5, 7
- [18] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 850–855. IEEE, 2006. 6
- [19] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5
- [20] 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. 1, 2, 6, 7
- [21] Zengyi Qin, Jinglu Wang, and Yan Lu. Triangulation Learning Network: From Monocular to Stereo 3D Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7615–7623, 2019. 1
- [22] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019. 1, 3, 6
- [23] Andrea Simonelli, Samuel Rota Rota Bulò, Lorenzo Porzi, Manuel López-Antequera, and Peter Kotschieder. Disentangling Monocular 3D Object Detection. *arXiv:1905.12365 [cs]*, 2019. 1
- [24] Bei Wang, Jianping An, and Jiayan Cao. Voxel-FPN: Multi-scale voxel feature aggregation in 3D object detection from point clouds. *arXiv:1907.05286 [cs, stat]*, 2019. 2, 3, 6, 7
- [25] Dominic Zeng Wang and Ingmar Posner. Voting for voting in online point cloud object detection. In *Robotics: Science and Systems*, volume 1, pages 10–15607, 2015. 2
- [26] Zhixin Wang and Kui Jia. Frustum ConvNet: Sliding Frustums to Aggregate Local Point-Wise Features for Amodal 3D Object Detection. *arXiv:1903.01864 [cs]*, 2019. 1, 3, 6
- [27] Bin Xu and Zhenzhong Chen. Multi-Level Fusion Based 3D Object Detection From Monocular Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2345–2353, 2018. 1
- [28] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1, 2, 6, 7

- [29] Bin Yang, Ming Liang, and Raquel Urtasun. HDNET: Exploiting HD Maps for 3D Object Detection. In *Conference on Robot Learning*, pages 146–155, 2018. 1, 6
- [30] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2
- [31] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. IPOD: intensive point-based object detector for point cloud. *CoRR*, abs/1812.05276, 2018. 1, 3, 6
- [32] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. STD: Sparse-to-Dense 3D Object Detector for Point Cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 3, 6
- [33] Dingfu Zhou, Jin Fang, Xibin Song, Chenye Guan, Junbo Yin, Yuchao Dai, and Ruigang Yang. IoU Loss for 2D/3D Object Detection. *arXiv:1908.03851 [cs]*, 2019. 5
- [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. *arXiv preprint arXiv:1910.06528*, 2019. 3
- [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. 1, 2, 3, 4, 6, 7, 8