

# HVPR: Hybrid Voxel-Point Representation for Single-stage 3D Object Detection

Jongyoun Noh

Sanghoon Lee

Bumsub Ham\*

School of Electrical and Electronic Engineering, Yonsei University

## Abstract

We address the problem of 3D object detection, that is, estimating 3D object bounding boxes from point clouds. 3D object detection methods exploit either voxel-based or point-based features to represent 3D objects in a scene. Voxel-based features are efficient to extract, while they fail to preserve fine-grained 3D structures of objects. Point-based features, on the other hand, represent the 3D structures more accurately, but extracting these features is computationally expensive. We introduce in this paper a novel single-stage 3D detection method having the merit of both voxel-based and point-based features. To this end, we propose a new convolutional neural network (CNN) architecture, dubbed HVPR, that integrates both features into a single 3D representation effectively and efficiently. Specifically, we augment the point-based features with a memory module to reduce the computational cost. We then aggregate the features in the memory, semantically similar to each voxel-based one, to obtain a hybrid 3D representation in a form of a pseudo image, allowing to localize 3D objects in a single stage efficiently. We also propose an Attentive Multi-scale Feature Module (AMFM) that extracts scale-aware features considering the sparse and irregular patterns of point clouds. Experimental results on the KITTI dataset demonstrate the effectiveness and efficiency of our approach, achieving a better compromise in terms of speed and accuracy.

## 1. Introduction

3D object detection is an essential technique for scene understanding, which aims at predicting 3D bounding boxes of objects in a scene. It can also be exploited as a basic building block for many applications, including autonomous driving and robotics. Recent approaches to 3D object detection focus on learning discriminative 3D representations using point clouds acquired from a LiDAR sensor. Point clouds provide accurate depth information of objects, but they are sparse with the densities largely varying

\*Corresponding author.

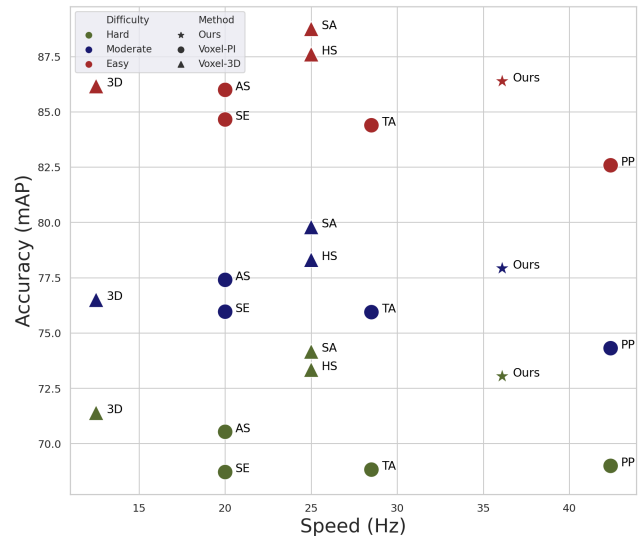


Figure 1: Runtime and accuracy comparison of detection results on the KITTI [10] test set. We compare our model with voxel-based methods on the car class for three difficulty levels. Voxel-based methods using pseudo image representations (Voxel-PI) are shown as circles, and 3D voxel-based methods (Voxel-3D) are plotted as triangles. Our method gives a better compromise in terms of accuracy and runtime for all cases. SE: SECOND [42]; PP: PointPillars [18]; TA: TANet [23]; AS: Associate-3D [7]; 3D: 3DIoULoss [49]; SA: SA-SSD [12]; HS: HotSpotNet [3]. Best viewed in color.

w.r.t distances from the sensor.

There are many attempts to learn 3D feature representations using point clouds with deep neural networks, which can be categorized into two groups: Voxel-based and point-based methods. The first approaches [7, 12, 18, 35, 43, 46] transform raw point clouds to structured grid representations, e.g., birds-eye-view (BEV) [35, 43] or voxels [12, 18, 33, 46, 50], to extract 3D representations using convolutional neural networks (CNNs). The voxelization and downsampling operations in these methods allow to extract compact 3D features efficiently, but they fail to preserve fine-grained 3D structures of objects. The second approaches [29, 34, 38, 44, 45], on the other hand, exploit raw point clouds directly to extract point-wise features using e.g., PointNet++ [30, 31]. They provide more discrim-

inative 3D representations than the voxel-based one, and thus give better detection results. Processing a large-scale point cloud data, however, needs lots of computational cost.

We present in this paper a novel single-stage 3D object detection framework that integrates voxel-based and point-based features effectively and efficiently to obtain discriminative 3D representations. To this end, we introduce a new CNN architecture, dubbed HVPR, that consists of a two-stream encoder for voxel-based and point-based features and a memory module. It augments the encoder for point-based features with the memory module to reduce the computational cost. Namely, we update and store point-based features from the encoder to memory items during training, and do not use the encoder at test time, avoiding heavy computation. Specifically, we aggregate the point-wise features in memory items, semantically similar to each voxel-based one, to obtain a pseudo image representation, enabling exploiting hybrid 3D representations efficiently for localizing objects in a single stage. This also encourages voxel-based features to incorporate fine-grained representations of point-based features, which are particularly effective for localizing small or largely occluded objects acquired from sparse point clouds. We also introduce a detection network with an Attentive Multi-scale Feature Module (AMFM). Given the hybrid pseudo image, it extracts multi-scale features, and AMFM refines them using 3D scale representations to obtain scale-aware features, which are crucial especially for 3D object detection, due to the sparse and irregular patterns of point clouds. Extensive experimental results on standard benchmarks demonstrate the effectiveness and efficiency of our approach to exploiting hybrid feature representations. The main contributions of this paper can be summarized as follows:

- We introduce a novel single-stage framework for 3D object detection using hybrid 3D representations. We propose to use a memory module to augment point-based features, maintaining the efficiency of a single-stage method.
- We introduce AMFM that provides scale-aware features considering the sparse and irregular patterns of point clouds explicitly, allowing to consider complex scale variations across objects for 3D object detection.
- We demonstrate that our approach gives a better compromise in terms of speed and accuracy, compared to the state of the art. Our model runs at 36.1fps, while achieving competitive performance on the KITTI dataset [10].

Our code and models are available online: <https://cvlab.yonsei.ac.kr/projects/HVPR>.

## 2. Related work

**Multi-sensor based 3D detection.** Multi-sensory data, obtained from *e.g.*, RGB and depth sensors, provides complementary information for 3D object detection. For ex-

ample, RGB and depth images give semantic and structural information of objects, respectively. Many approaches attempt to leverage features from RGB images and point clouds jointly for 3D object detection. MV3D [5] generates object proposals from BEV representations, and refines them using features from point clouds and RGB images. AVOD [16] instead incorporates these features to extract object proposals, which provides better detection results. To further enhance 3D representations, proxy tasks are also exploited [20], such as a ground estimation with point clouds and a depth completion with RGB images. Recent methods [13, 47] also handle the case when multi-sensory data is not registered. They leverage correspondences between features from point clouds and RGB images to better exploit the complementary information.

**LiDAR based 3D detection.** Recent approaches to 3D object detection use point clouds alone from a LiDAR sensor. They can be divided into two categories, voxel-based and point-based methods, depending on how to transform point clouds to 3D representations for localizing objects. First, voxel-based approaches convert irregular point clouds into ordered grid representations. Early works [35, 43] exploit a BEV image, obtained by projecting raw point clouds, which provides a compact 3D representation and preserves the scales of objects. Hand-crafted BEV representations are, however, limited to obtain accurate 3D features for object localization. Recent methods represent point clouds in a form of 3D voxels [7, 12, 18, 46, 50]. They often use 3D convolutions to extract features from the voxel representation, which is computationally inefficient. To alleviate this problem, PointPillars [18] propose to use a pseudo-image. Representing 3D voxels in a form of an image allows to exploit 2D convolutions, resulting in more efficient 3D object detection, compared to using 3D voxels directly. The voxel-based approaches mainly have two limitations: (1) Fine-grained 3D structures are lost during voxelization; (2) The localization performance largely depends on the size of voxel grids. As the grid size becomes smaller, fine-grained 3D representations can be obtained but at the cost of more runtime. Second, point-based approaches leverage raw point clouds directly. They extract point-based features, typically using PointNet++ [30, 31], that provide more accurate spatial representations than 3D voxels. These approaches thus perform better than the voxel-based one in terms of detection accuracy. It is, however, computationally demanding to extract point-based features directly from a large amount of point clouds for *e.g.*, outdoor sceneries. To circumvent this issue, point-based approaches adopt a two-stage pipeline. Specifically, they estimate 3D object proposals either using point-based features [34, 45] or off-the-shelf 2D object detectors [29, 38] in the first stage. The object proposals are then refined with the point-based features inside the proposals for final predictions in the second stage.

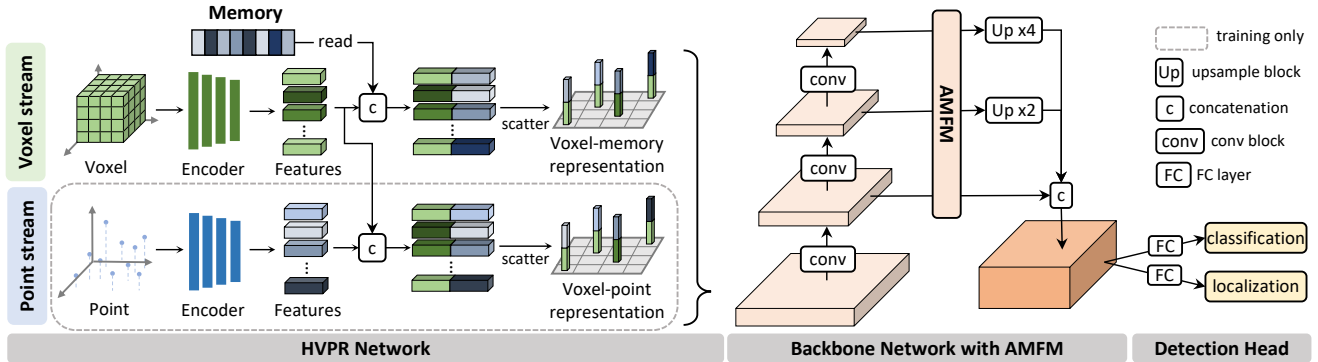


Figure 2: An overview of our framework. The HVPR network inputs point clouds and generates two types of hybrid 3D features via a two-stream encoder: Voxel-point and voxel-memory representations. The former representations are obtained by aggregating point-based features for individual voxel-based ones. For the later ones, we also perform the aggregation but with memory items, instead of using point-based features. That is, we augment the point-based features using a memory module, and exploit voxel-memory representations, *i.e.*, hybrid 3D features, at test time for fast inference. The backbone network with AMFM inputs the voxel-memory representations to extract multiple scale-aware features, and the detection head predicts 3D bounding boxes and object classes. See text for details.

Recently, PV-RCNN [33] and Fast Point R-CNN [6] further reduce the computational cost of point-based approaches. They propose to exploit voxel-based features to generate 3D object proposals in the first stage, while refining them to estimate final 3D bounding boxes with point-based features in the second stage. Since these methods exploit 3D voxels and raw point clouds separately in different stages, they do not fully leverage the complementary information of the two representations. Moreover, the point-based approaches are still slower than voxel-based methods, even leveraging a single-stage framework [38, 44].

Our approach belongs to the voxel-based method in the sense that it uses a pseudo image representation of point clouds, similar to PointPillars [18]. In contrast to previous voxel-based methods, we also exploit point-based features for the pseudo image representation in an efficient way, which boosts the detection performance significantly, while retaining fast runtime. Our approach is similar to recent point-based methods [6, 33, 44] in that it exploits both voxel-based and point-based features [6, 33]. On the contrary, we leverage both voxel-based and point-based representations jointly within a single-stage framework. We weave both representations in a form of a pseudo image, allowing to exploit 2D convolutions for efficient 3D object detection. Moreover, we further reduce the computational cost for extracting point-based features via augmenting a feature encoder with a memory module.

**Memory networks.** Memory networks [39] have been introduced to capture long-term dependencies in sequential data by augmenting a neural network with an external memory module that can be read and written to. Based on this idea, recent works develop more advanced architectures using continuous memory representations [36] or a key-value memory [26]. A number of approaches have adopted the memory-augmented networks to solve various computer vi-

sion tasks including visual question answering [9, 17, 25], one-shot learning [2, 14, 32], anomaly detection [11, 27], and person recognition [41, 48].

### 3. Approach

Our model mainly consists of three components (Fig. 2): a HVPR network (Sec. 3.1), a backbone network with AMFM (Sec. 3.2), and a detection head (Sec. 3.3). Given point clouds, the HVPR network outputs hybrid voxel-point representations in a form of pseudo-images. To this end, we exploit a two-stream encoder to extract voxel-based and point-based features. For each voxel-based feature, we aggregate point-based ones, based on their similarities, and obtain hybrid voxel-point representations. Extracting point-based features is, however, computationally demanding. To overcome this, we augment point-based features using a memory module. Specifically, we store various prototypes of point-based features in memory items, and aggregate the prototypical features in the memory to obtain voxel-memory representations. The memory items are updated by encouraging the aggregated prototypical and point-based features to be similar. Note that we exploit the voxel-memory representations only at test time, instead of using point-based ones directly, enabling a fast object detection. The backbone network inputs the voxel-memory representations in a form of a pseudo image, and extracts multi-scale feature maps. The AMFM refines the feature maps using 3D scale representations, and provides scale-aware features. The detection head predicts 3D object bounding boxes and object classes using the scale-aware features. In the following, we describe our single-stage detection framework in detail.

#### 3.1. HVPR Network

**Voxel-based feature.** We assume that the physical dimension of a 3D scene that point clouds lie on is within the range

of  $W \times H \times L$ , where  $W$ ,  $H$ , and  $L$  are width, height, and length, respectively, in three-dimensions. Supposing that the size of each voxel is  $v_W \times v_H \times v_L$ , we define voxel grids of size  $W' \times H' \times L'$ , where  $W' = W/v_W$ ,  $H' = H/v_H$ ,  $L' = L/v_L$ . Following [18], we voxelize point clouds in a x-y plane only, *i.e.*, by setting the value of  $v_L$  to be the same as  $L$ . We represent a voxel as a tensor of size  $N_{\text{vox}} \times D$ , where  $N_{\text{vox}}$  and  $D$  are the number of point clouds within each voxel and the size of augmented point clouds<sup>1</sup>, respectively. We exploit a tiny PointNet [30] as an encoder for voxel-based features. The encoder takes a set of voxels and extracts features of all point clouds in each voxel, providing an output of size  $C \times N \times N_{\text{vox}}$ , where  $N$  is a total number of voxels. It then applies a max pooling operator over the point clouds in each voxel to obtain a voxel-based feature map of size  $C \times N$ , where we denote by  $\mathbf{f}_{\text{vox}}(n)$  each voxel-based feature of size  $C \times 1$ , where  $n = 1, \dots, N$ .

**Point-based feature.** As a point stream, we adopt a PointNet++ [31], widely used to other 3D applications, such as 3D semantic segmentation [1, 37] and 3D object detection [29, 34, 38, 44, 45], that extracts point-based features from raw point clouds directly. To this end, PointNet++ uses set abstraction (SA) and feature propagation (FP) layers. It first downsamples a set of points gradually, with a number of SA layers, to produce novel features with fewer points. The FP layers then propagate sub-sampled features to recover the initial points entirely. We follow these procedures to obtain a point-based feature for each point cloud. We denote by  $\mathbf{f}_{\text{pts}}(m)$  each point-based feature of size  $C \times 1$ , where  $m = 1, \dots, M$  and  $M$  is a total number of point clouds.

### 3.1.1 Voxel-point representation

We integrate voxel-based and point-based features,  $\mathbf{f}_{\text{vox}}$  and  $\mathbf{f}_{\text{pts}}$ , to obtain a hybrid voxel-point representation. To this end, we compute the dot product between all pairs of voxel-based and point-based features, resulting in a 2-dimensional correlation map of size  $N \times M$  as follows:

$$C(n, m) = \mathbf{f}_{\text{vox}}(n)^\top \mathbf{f}_{\text{pts}}(m). \quad (1)$$

For each voxel-based feature, we select the nearest  $K$  point-based features, according to the correlation scores. We denote by  $\hat{\mathbf{f}}_{\text{pts}}(n, k)$  the  $k$ -th similar point-wise feature w.r.t the  $n$ -th voxel-based one, where  $k = 1, \dots, K$ . We then compute the matching probabilities between these voxel-based and point-based features as follows:

$$P(n, k) = \frac{\exp(\mathbf{f}_{\text{vox}}(n)^\top \hat{\mathbf{f}}_{\text{pts}}(n, k))}{\sum_{k'} \exp(\mathbf{f}_{\text{vox}}(n)^\top \hat{\mathbf{f}}_{\text{pts}}(n, k'))}. \quad (2)$$

<sup>1</sup> We parameterize a point cloud using its positions at x, y, z coordinates, and a reflectance intensity of laser. Following [18], we further augment each point cloud with average distances of all point clouds in the voxel and x-y offsets from the center position of the voxel, resulting in a  $D$ -dimensional feature vector.

Finally, we aggregate the nearest  $K$  point-based features with corresponding matching probabilities as follows:

$$\mathbf{g}_{\text{pts}}(n) = \sum_k P(n, k) \hat{\mathbf{f}}_{\text{pts}}(n, k), \quad (3)$$

where  $\mathbf{g}_{\text{pts}}(n)$  is an aggregated point-based feature of size  $C \times 1$ . It is semantically similar to the corresponding voxel-based one, but contains more accurate 3D structural information for objects. This enables imposing the fine-grained 3D representation to the voxel-based feature that loses the structure during voxelization. Specifically, we concatenate voxel-based and aggregated point-based features, and obtain a voxel-point pseudo image of size  $H \times W \times 2C$  by scattering the concatenated features back to the corresponding voxel locations, similar to PointPillars [18]. In this context, the aggregated feature also has an effect of feature augmentation, providing more discriminative 3D representations, especially for the voxels obtained from sparse point clouds.

### 3.1.2 Voxel-memory representation

Point-wise features obtained by PointNet++ [31] contain fine-grained 3D representations, but this requires lots of computational cost. PointRCNN [34], for instance, composed of 4 pairs of SA and FP layers, requires 54ms with a Titan V GPU to obtain point-based features [44], which already exceeds the overall runtime of most voxel-based 3D detection methods [12, 18, 46]. To address this problem, we augment point-based features using a memory module, where each memory item stores various prototypes of the features, to obtain voxel-memory representations.

Concretely, we denote by  $\mathbf{f}_{\text{mem}}(t)$  each item in the memory, the size of which is  $C \times 1$ , where  $t = 1, \dots, T$ , and  $T$  is a total number of memory items. We integrate the voxel-based features  $\mathbf{f}_{\text{vox}}$  and the memory items  $\mathbf{f}_{\text{mem}}$  to obtain a voxel-memory pseudo image in a similar way to the voxel-point one in Sec. 3.1.1. Specifically, we consider the memory as a set of prototypes for point-based features. To read memory items, we use voxel-based features as queries. We then compute matching probabilities between voxel-based features and memory items, and aggregate the top  $K$  items with corresponding probabilities for each voxel-based feature, which is analogous to the aggregated point-based feature in (3). We concatenate the voxel-based features with the aggregated items, and scatter the concatenated features back to the original voxel locations to form a voxel-memory pseudo image. To update memory items, we encourage the aggregated point-based features and memory items to be similar as follows:

$$\mathcal{L}_{\text{mem}} = \sum_n \|\mathbf{g}_{\text{pts}}(n) - \mathbf{g}_{\text{mem}}(n)\|_2, \quad (4)$$

where we denote by  $\mathbf{g}_{\text{mem}}(n)$  an aggregated memory item, and  $\|\cdot\|_2$  computes the L2 norm of a vector. Note that we

do not use the voxel-point pseudo image at test time. We instead use the voxel-memory pseudo image only, avoiding extracting point-based features at test time. As will be shown in our experiments, obtaining hybrid 3D representations using a memory is much more cheaper than exploiting point-based features directly, while retaining the same level of quality.

### 3.2. Backbone Network with AMFM

Our backbone network consists of a series of 2D convolutional layers and AMFM (Fig. 2). It inputs our hybrid 3D representations, *i.e.*, the voxel-memory pseudo image, and gives multi-scale feature maps. The AMFM then refines them using spatial attention maps, and provides scale-aware features. We obtain final feature representations for predicting 3D object bounding boxes by concatenating the scale-aware feature maps along the channel dimension.

**AMFM.** There are large scale variations across 3D object instances in a scene. Multi-scale features help to consider the scale variations for 3D object detection, but they are not scale-aware, even when obtained from a feature pyramid [19]. To address this problem, we introduce AMFM that provides scale-aware features using 3D scale information explicitly. The basic idea of AMFM is to exploit spatial attention maps [40], while leveraging 3D scale features explicitly, for suppressing scale-confused features. Specifically, AMFM refines a multi-scale feature by element-wise multiplication with the attention map. It then combines the initial and refined multi-scale features via a skip connection. We illustrate in Fig. 3 how AMFM works in the backbone network in detail. In the following, we describe how to extract 3D scale features, and then present how to obtain a spatial attention map using 3D scale features in detail.

We observe that 3D point clouds are sparse with the densities largely varying w.r.t distances from a LiDAR sensor. Namely, the sparse and irregular patterns of point clouds and their distances from a sensor reflect scale information of 3D objects. Based on this idea, we represent each voxel with the number of point clouds within a voxel, and an absolute position of point clouds averaged on each voxel and its distance from a sensor. We input a set of voxel representations to a network, and obtain voxel-wise features. To be specific, we exploit a tiny PointNet [30] similar to the encoder for voxel-based features. We then scatter the voxel-wise features back to the original locations to obtain a 3D scale feature map of the same size as the voxel-memory pseudo image. We downsample the 3D scale feature with a strided convolution, before feeding it to AMFM, such that it has the same spatial resolution as a corresponding multi-scale feature map. To obtain a spatial attention map, we apply max pooling and average pooling layers to the 3D scale feature along the channel dimension, highlighting informative regions in the feature map [15]. We concatenate the pooled features, and feed the concatenated one to a con-

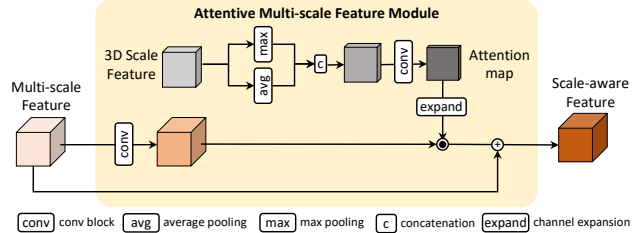


Figure 3: Illustration of AMFM. AMFM refines multi-scale features with a spatial attention map. It then combines the initial and refined features via a skip connection to obtain scale-aware representations. To compute the attention map, we leverage 3D scale features considering scale variations of objects in a 3D scene. See text for details.

volutional layer, followed by applying a sigmoid function, to produce the attention map.

### 3.3. Detection Head and Loss

Following [18], we adopt the detection head of SSD [22] to localize 3D objects. It consists of two fully-connected layers for regressing 3D object bounding boxes and classifying object classes inside the boxes. We parameterize a 3D object bounding box as  $(x, y, z, w, l, h, \theta)$ , where  $(x, y, z)$ ,  $(w, l, h)$  and  $\theta$  are the center coordinate, the size, the heading angle of the box, respectively. We compute 2D intersection over union (IoU) scores [8] between anchor boxes and the ground truth, and pick the anchors with the IoU scores being larger than a predefined threshold. Our detection head regresses residuals between the anchor boxes and the ground truth directly, defined as follows:

$$\begin{aligned} \Delta x &= \frac{x^{\text{gt}} - x^{\text{a}}}{d}, \Delta y = \frac{y^{\text{gt}} - y^{\text{a}}}{d}, \Delta z = \frac{z^{\text{gt}} - z^{\text{a}}}{h^{\text{a}}} \\ \Delta w &= \log \frac{w^{\text{gt}}}{w^{\text{a}}}, \Delta l = \log \frac{l^{\text{gt}}}{l^{\text{a}}}, \Delta h = \log \frac{h^{\text{gt}}}{h^{\text{a}}} \\ \Delta \theta &= \sin(\theta^{\text{gt}} - \theta^{\text{a}}), \end{aligned} \quad (5)$$

where  $d = \sqrt{(w^{\text{a}})^2 + (l^{\text{a}})^2}$ , and we use the superscripts, gt and a, to indicate parameters for ground-truth and anchor boxes, respectively.

To train our model, we use four terms for regression ( $\mathcal{L}_{\text{reg}}$  and  $\mathcal{L}_{\text{dir}}$ ), classification ( $\mathcal{L}_{\text{cls}}$ ), and memory update ( $\mathcal{L}_{\text{mem}}$  in (4)), as follows:

$$\mathcal{L} = \frac{1}{N_{\text{pos}}} (\lambda_{\text{reg}} \mathcal{L}_{\text{reg}} + \lambda_{\text{dir}} \mathcal{L}_{\text{dir}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{mem}} \mathcal{L}_{\text{mem}}), \quad (6)$$

where  $N_{\text{pos}}$  is the number of positive anchors, and  $\lambda$  is a balancing parameter for the corresponding loss. We define the regression term with the residual between anchor and ground-truth boxes as follows:

$$\mathcal{L}_{\text{reg}} = \sum_{r \in (x, y, z, w, l, h, \theta)} \text{SmoothL1}(\Delta r). \quad (7)$$

Note that  $\Delta\theta$  is defined with a sine function in (5), penalizing the same 3D bounding boxes but with opposite directions differently. To prevent this, we also exploit the angle classification term ( $\mathcal{L}_{\text{dir}}$ ) [42]. As a classification term ( $\mathcal{L}_{\text{cls}}$ ), we use the focal loss [21], which has shown the effectiveness to handle a class imbalance between positive and negative samples, with the default parameters in [21].

## 4. Experiments

### 4.1. Implementation details

**Dataset and evaluation protocol.** We evaluate our model on the KITTI 3D object detection benchmark [10]. The KITTI dataset provides 7,481 training and 7,518 test samples for object classes of cars, pedestrians and cyclists. Object instances across different classes are further classified into easy, moderate and hard splits, depending on the object size, the degree of occlusion and the maximum truncation level. We evaluate our model on the car and the pedestrian classes, most commonly used object class and the hardest one for 3D object detection, respectively. Following [23], we split the training set into two splits with the ratio of around 5:1 to train our model for testing via an official evaluation server. For other experiments including an ablation study, we follow the experimental protocol in [4, 8], and split the original training set of 7,481 samples into 3,712 and 3,769 ones for training and validation, respectively. We report the mean average precision (mAP) on both validation and test splits with 40 recall positions on a PR curve [10]. At test time, we apply non maximum suppression (NMS) with a threshold of 0.1 to eliminate multiple overlapping bounding boxes.

**Training.** Following the previous works [18, 23, 42, 50], we train separate networks for the car and the pedestrian classes. We describe hereafter the training setting for the car class, and the one for the pedestrian is included in the supplementary material. We train our network for 100 epochs, with a learning rate of  $3e-3$  and a weight decay of  $1e-2$ . We use a cosine annealing technique [24] as a learning rate scheduler. Batch size is set to 8 for the network. All models are trained end-to-end using PyTorch [28]. We use data augmentation techniques during training. Specifically, we apply random flipping along the  $x$ -axis, global scaling, where a scaling factor is randomly chosen within a range of  $[0.95, 1.05]$ , and global rotation along the  $z$ -axis, where a rotation angle is sampled randomly from a range of  $[-\frac{\pi}{4}, \frac{\pi}{4}]$ . We also use a ground-truth augmentation technique [42] that randomly selects ground-truth bounding boxes within the entire training dataset and associates them for each 3D scene.

**Parameter setting.** We set the parameters for our model by a grid search in terms of mAP on the validation split:  $K = 20$ ,  $C = 64$ ,  $T = 2,000$ ,  $\lambda_{\text{reg}} = 2.0$ ,  $\lambda_{\text{dir}} = 0.2$ ,  $\lambda_{\text{cls}} = 1.0$  and  $\lambda_{\text{mem}} = 1.0$ . For other parameters, we

use the same setting as in [18] as follows: We assume that the dimension of a 3D scene ( $W, H, L$ ) is within a range of  $[(0, 70.4), (-40, 40), (-3, 1)]$  meters. We set the size of anchors as  $1.6 \times 3.9 \times 1.5$ . To define positive and negative pairs, we compute IoU scores between the anchors and the ground truth. We choose the anchors with the IoU scores being larger than 0.6, as positive boxes, while those lower than 0.45 are used as negative ones. We use  $(0.16, 0.16, 4)$  as the size of a voxel,  $v_W \times v_H \times v_L$ . The number of point clouds within each voxel  $N_{\text{vox}}$  and the size of augmented point clouds  $D$  are set to 32 and 9, respectively. Please refer to the supplementary material for more implementation details including our network architecture.

### 4.2. Results

We compare in Table 1 our model with the state of the art for LiDAR based 3D detection. We report mAP scores of car and pedestrian classes on the KITTI test dataset [10]. All numbers including ours are obtained from an official evaluation server, except for [42, 50], which are taken from [3]. We classify 3D object detection methods into point-based and voxel-based approaches, according to features used to represent 3D objects. The voxel-based approaches are further categorized into two groups based on whether they exploit 3D voxels explicitly or implicitly in a form of a pseudo image, where our model belongs to the latter one. We mainly compare our model with voxel-based methods, especially with the ones exploiting pseudo image representations. For comparison, we also report results of point-based and other voxel-based approaches. From this table, we observe four things: (1) Our model for the car class gives the best results among voxel-based methods using pseudo images. This demonstrates the effectiveness of our approach to exploiting a hybrid 3D representation and scale-aware features for 3D object detection. (2) Our method is fastest among voxel-based approaches, except PointPillars [18], while providing competitive results. Although PointPillars is faster than ours (42.4Hz vs. 36.1Hz), it is outperformed by our model by a significant margin in terms of mAP for all difficulty levels. This suggests that our model offers a good compromise in terms of speed and accuracy. (3) The performance improvements over other voxel-based methods using pseudo images are particularly significant in a hard split, where 3D objects are typically small in size and largely occluded. For example, the mAP gain of our model w.r.t the second best method [7] (73.04 vs. 70.53) is much larger than those for easy and moderate splits of the car class. Also, our model gives the best result among other voxel-based methods using pseudo images on the hard split of the pedestrian class. This confirms that augmenting point-based features provides more discriminative 3D representations, especially for objects taken from sparse point clouds. (4) Our model for the pedestrian class outperforms PointPillars [18] significantly, while providing

Table 1: Quantitative comparison with the state of the art in terms of mAP(%) and runtime on the KITTI test set [10]. We mainly compare our model with voxel-based methods using pseudo image representations (Voxel-PI). For comparison, we also report results for point-based and (Point) other voxel-based methods using 3D voxel representations (Voxel-3D), although they typically provide better results, but at the cost of much slower runtime, than exploiting pseudo image representations. Numbers in bold for Voxel-PI indicate the best performance and underscored ones are the second best.

Type	Stage	Methods	Reference	Runtime (Hz)	GPU	Car			Pedestrian		
						Easy	Mod.	Hard	Easy	Mod.	Hard
Point	Two	PointRCNN [34]	CVPR 2019	10	TITAN XP	86.96	75.64	70.70	47.98	39.37	36.01
	Two	FastPointRCNN [6]	ICCV 2019	16.7	Tesla P40	85.29	77.40	70.24	-	-	-
	One	F-ConvNet [38]	IROS 2019	2.1	GTX 1080	87.36	76.39	66.69	52.16	43.38	38.80
	Two	STD [45]	ICCV 2019	12.5	TITAN V	87.95	79.71	75.09	53.29	42.47	38.35
	One	3DSSD [44]	CVPR 2020	25	TITAN V	88.36	79.57	74.55	54.64	44.27	40.23
	Two	PV-RCNN [33]	CVPR2020	12.5	GTX 1080Ti	90.25	81.43	76.82	52.17	43.29	40.29
Voxel-3D	One	VoxelNet [50]	CVPR 2018	4.5	TITAN X	77.47	65.11	57.73	39.48	33.69	31.51
	One	3DIoULoss [49]	3DV 2019	12.5	-	86.16	76.50	71.39	-	-	-
	One	SA-SSD [12]	CVPR 2020	25	GTX 2080Ti	88.75	79.79	74.16	-	-	-
	One	HotSpotNet [3]	ECCV 2020	25	-	87.60	78.31	73.34	53.10	45.37	41.47
Voxel-PI	One	SECOND [42]	Sensors 2018	20	GTX 1080Ti	84.78	75.32	68.70	45.31	35.52	33.14
	One	PointPillars [18]	CVPR 2019	<b>42.4</b>	GTX 1080Ti	82.58	74.31	68.99	51.45	41.92	38.89
	One	TANet [23]	AAAI 2020	28.5	TITAN V	84.39	75.94	68.82	<b>53.72</b>	<b>44.34</b>	<u>40.49</u>
	One	Associate-3D [7]	CVPR 2020	20	GTX 1080Ti	<u>85.99</u>	<u>77.40</u>	<u>70.53</u>	-	-	-
	One	Ours		<u>36.1</u>	GTX 2080Ti	<b>86.38</b>	<b>77.92</b>	<b>73.04</b>	<u>53.47</u>	<u>43.96</u>	<b>40.64</b>

Table 2: Quantitative comparison for variants of our model. We compute mAP (%) and measure runtime (Hz) on the KITTI validation set [10]. VoxF: Voxel-based feature; PtsF: Point-based feature; MEM: Memory module; 3DSF: 3D scale feature.

VoxF	PtsF	Mem	AMFM	3DSF	Easy	Mod.	Hard	Time
✓					87.66	78.82	75.32	39.5
✓	✓				88.97	80.35	77.28	6.8
✓	✓	✓			87.96	80.44	77.92	37.9
✓	✓	✓	✓		89.51	80.86	78.22	31.3
✓	✓	✓	✓	✓	<b>91.14</b>	<b>82.05</b>	<b>79.49</b>	36.1

a competitive result with the best voxel-based method using pseudo images [23]. This shows that our model can generalize to other object classes. Note that our 3D representations from the HVPR network can be easily incorporated into other voxel-based methods [12, 49, 50], which may boost the mAP performance.

### 4.3. Discussion

In this section, we show more experimental results to analyze individual components of our model in detail. Following previous works [44, 45, 50], we report quantitative results of the car class in terms of mAP on the validation split of KITTI [10]. We measure the mAP performance with 40 recall positions.

**Ablation study.** We show an ablation analysis on different components of our model in Table 2. As our baseline in the first row, we use PointPillars [18], reproduced by ourselves, for ablation studies. Similar to ours, PointPillars leverage a pseudo image to represent a 3D scene, but it exploits voxel-based features alone. We can see from

the first and second rows that a hybrid representation using additional point-based features boosts the detection performance remarkably, which provides mAP gains of 1.31, 1.53 and 1.94 for easy, moderate and hard splits, respectively. The improvements for the moderate and hard splits are more significant than the easy one. This suggests that 1) the aggregated point-based features allow to obtain more discriminative 3D representations, especially for small or occluded objects with sparse points, and 2) voxel-based and point-based features are complementary to each other. Extracting point-based features, however, is computationally expensive, which degrades runtime performance from 39.5Hz to 6.8Hz. The third row demonstrates that our memory module boosts the runtime performance significantly, while retaining the same level of quality. The average runtime is almost the same as the baseline (39.5Hz vs. 37.9Hz). This shows the effectiveness and efficiency of a memory module to augment point-based features. Our model in the fourth row exploits AMFM but without using 3D scale features. That is, it computes an attention map directly using multi-scale features. From the third and fourth rows, we can see that refining multi-scale features with a spatial attention map gives better results. The last two rows show that 3D scale features further boosts the detection performance drastically for all cases. The mAP gains for each split are 1.63, 1.19 and 1.27, respectively. The reason is that the features allow to leverage 3D scale information explicitly, which is particularly important to consider complex scale variations of objects in a scene for 3D object detection.

**Memory items.** We show in Table 3 quantitative results for different combinations of a total number of memory



Figure 4: Qualitative results on the validation split of KITTI [10]. Our predictions and ground-truth bounding boxes are shown in green and red, respectively. Our method localizes small and/or occluded objects well, except the heavily occluded ones, *e.g.*, in the left-bottom of the top-middle image. We also show 2D bounding boxes projected from 3D detection results. Best viewed in color.

Table 3: Quantitative comparison for different combinations of a total numbers of memory items ( $T$ ) and the number of nearest items retrieved ( $K$ ). We report mAP (%) for the car class on the validation split of KITTI [10].

	$K = 5$	$K = 10$	$K = 20$	$K = 30$
$T = 1000$	80.22	80.74	81.32	80.82
$T = 2000$	80.68	80.91	82.05	81.18
$T = 3000$	80.68	80.82	81.56	81.45

items  $T$  and the number of nearest items retrieved  $K$ . We compare the performance of our final model, with a memory module and AMFM, in terms of mAP, and report results of the car class on the moderate split. We can clearly see that the mAP performance is robust over the total number of memory items. We can also observe that aggregating more items for hybrid 3D representations provides better results, but using less similar items are not helpful. For example, mAP scores increase, until we reach the number of nearest items  $K$  to the value of 20. After that, using more items rather lowers the scores.

**Qualitative results.** We visualize in Fig. 4 detection results on the validation split of KITTI [10]. We can see that our model is robust to complex scale variations. That is, it localizes 3D objects well, including small ones captured with sparse point clouds. This confirms the effectiveness

of our hybrid 3D representations and scale-aware features once again. More qualitative results are available in the supplementary material.

## 5. Conclusion

We have introduced a novel single-stage 3D detection method using hybrid 3D representations. We have integrated voxel-based and point-based features effectively to obtain 3D representations of a scene in a form of a pseudo image. For efficient object detection, we have proposed to use a memory module to augment point-based features. We have also presented AMFM that provides scale-aware representations using 3D scale features, which boosts the performance of 3D object detection significantly. An extensive experimental analysis on KITTI demonstrates that our model achieves the best results among voxel-based methods using pseudo images, and it provides a better compromise in terms of accuracy and speed than other 3D detection approaches.

**Acknowledgments.** This research was partly supported by R&D program for Advanced Integrated-intelligence for Identification (AIID) through the National Research Foundation of KOREA (NRF) funded by Ministry of Science and ICT (NRF-2018M3E3A1057289), and Institute for Information and Communications Technology Promotion (IITP) funded by the Korean Government (MSIP) under Grant 2016-0-00197.



## References

- [1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, 2019. 4
- [2] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In *CVPR*, 2018. 3
- [3] 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 *ECCV*, 2020. 1, 6, 7
- [4] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3D object proposals for accurate object class detection. In *NeurIPS*, 2015. 6
- [5] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3D object detection network for autonomous driving. In *CVPR*, 2017. 2
- [6] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Fast point r-cnn. In *ICCV*, 2019. 3, 7
- [7] Liang Du, Xiaoqing Ye, Xiao Tan, Jianfeng Feng, Zhenbo Xu, Errui Ding, and Shilei Wen. Associate-3Ddet: Perceptual-to-conceptual association for 3D point cloud object detection. In *CVPR*, 2020. 1, 2, 6, 7
- [8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88, 2010. 5, 6
- [9] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, 2019. 3
- [10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 1, 2, 6, 7, 8
- [11] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *ICCV*, 2019. 3
- [12] Chenhang He, Hui Zeng, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Structure aware single-stage 3D object detection from point cloud. In *CVPR*, 2020. 1, 2, 4, 7
- [13] Tengpeng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. EP-Net: Enhancing point features with image semantics for 3D object detection. In *ECCV*, 2020. 2
- [14] Lukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. In *ICLR*, 2017. 3
- [15] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 5
- [16] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3D proposal generation and object detection from view aggregation. In *IROS*, 2018. 2
- [17] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Roman Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*, 2016. 3
- [18] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 1, 2, 3, 4, 5, 6, 7
- [19] Yazhao Li, Yanwei Pang, Jianbing Shen, Jiale Cao, and Ling Shao. NETNet: Neighbor erasing and transferring network for better single shot object detection. In *CVPR*, 2020. 5
- [20] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3D object detection. In *CVPR*, 2019. 2
- [21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 6
- [22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *ECCV*. Springer, 2016. 5
- [23] Zhe Liu, Xin Zhao, Tengpeng Huang, Ruolan Hu, Yu Zhou, and Xiang Bai. TANet: Robust 3D object detection from point clouds with triple attention. In *AAAI*, 2020. 1, 6, 7
- [24] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6
- [25] Chao Ma, Chunhua Shen, Anthony Dick, Qi Wu, Peng Wang, Anton van den Hengel, and Ian Reid. Visual question answering with memory-augmented networks. In *CVPR*, 2018. 3
- [26] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. Key-value memory networks for directly reading documents. In *EMNLP*, 2016. 3
- [27] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, 2020. 3
- [28] 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. 6
- [29] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3D object detection from rgb-d data. In *CVPR*, 2018. 1, 2, 4
- [30] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017. 1, 2, 4, 5
- [31] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 1, 2, 4
- [32] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016. 3
- [33] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-voxel feature set abstraction for 3D object detection. In *CVPR*, 2020. 1, 3, 7
- [34] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointRCNN: 3D object proposal generation and detection from point cloud. In *CVPR*, 2019. 1, 2, 4, 7
- [35] Martin Simon, Stefan Milzy, Karl Amendey, and Horst-Michael Gross. Complex-YOLO: An euler-region-proposal

- for real-time 3d object detection on point clouds. In *ECCVW*, 2018. 1, 2
- [36] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *NeurIPS*, 2015. 3
- [37] Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. Associatively segmenting instances and semantics in point clouds. In *CVPR*, 2019. 4
- [38] Zhixin Wang and Kui Jia. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection. In *IROS*, 2019. 1, 2, 3, 4, 7
- [39] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *ICLR*, 2015. 3
- [40] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *ECCV*, 2018. 5
- [41] Jiangyue Xia, Anyi Rao, Qingqiu Huang, Linning Xu, Jiangtao Wen, and Dahua Lin. Online multi-modal person search in videos. In *ECCV*, 2020. 3
- [42] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely embedded convolutional detection. *Sensors*, 18(10), 2018. 1, 6, 7
- [43] Bin Yang, Wenjie Luo, and Raquel Urtasun. PIXOR: Real-time 3D object detection from point clouds. In *CVPR*, 2018. 1, 2
- [44] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3DSSD: Point-based 3D single stage object detector. In *CVPR*, 2020. 1, 3, 4, 7
- [45] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. STD: Sparse-to-dense 3D object detector for point cloud. In *ICCV*, 2019. 1, 2, 4, 7
- [46] Maosheng Ye, Shuangjie Xu, and Tongyi Cao. HVNet: Hybrid voxel network for lidar based 3D object detection. In *CVPR*, 2020. 1, 2, 4
- [47] Jin Hyeok Yoo, Yeocheol Kim, Ji Song Kim, and Jun Won Choi. 3D-CVF: Generating joint camera and lidar features using cross-view spatial feature fusion for 3D object detection. In *ECCV*, 2020. 2
- [48] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019. 3
- [49] Dingfu Zhou, Jin Fang, Xibin Song, Chenye Guan, Junbo Yin, Yuchao Dai, and Ruigang Yang. Iou loss for 2D/3D object detection. In *3DV*, 2019. 1, 7
- [50] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3D object detection. In *CVPR*, 2018. 1, 2, 6, 7