DSGN: Deep Stereo Geometry Network for 3D Object Detection

Yilun Chen[1] Shu Liu[2] Xiaoyong Shen[2] Jiaya Jia[1,2]
[1]The Chinese University of Hong Kong [2]SmartMore
{ylchen, leojia}@cse.cuhk.edu.hk {sliu, xiaoyong}@smartmore.com

Abstract

Most state-of-the-art 3D object detectors heavily rely on LiDAR sensors because there is a large performance gap between image-based and LiDAR-based methods. It is caused by the way to form representation for the prediction in 3D scenarios. Our method, called Deep Stereo Geometry Network (DSGN), significantly reduces this gap by detecting 3D objects on a differentiable volumetric representation – 3D geometric volume, which effectively encodes 3D geometric structure for 3D regular space. With this representation, we learn depth information and semantic cues simultaneously. For the first time, we provide a simple and effective one-stage stereo-based 3D detection pipeline that jointly estimates the depth and detects 3D objects in an end-to-end learning manner. Our approach outperforms previous stereo-based 3D detectors (about 10 higher in terms of AP) and even achieves comparable performance with several LiDAR-based methods on the KITTI 3D object detection leaderboard. Our code will be (or is) available at `https://github.com/chenyilun95/DSGN`.

1. Introduction

3D scene understanding is a challenging task in 3D perception, which serves as a basic component for autonomous driving and robotics. Due to the great capability of LiDAR sensors to accurately retrieve 3D information, we witness fast progress on 3D object detection. Various 3D object detectors were proposed [9, 23, 58, 26, 27, 33, 39, 52, 54] to exploit LiDAR point cloud representation. The limitation of LiDAR is on the relatively sparse resolution of data with several laser beams and on the high price of the devices.

In comparison, video cameras are cheaper and are with much denser resolutions. The way to compute scene depth on stereo images is to consider disparity via stereo correspondence estimation. Albeit recently several 3D detectors based on either monocular [36, 7, 6, 30, 48] or stereo [25, 45, 37, 56] setting push the limit of image-based 3D object detection, the accuracy is still left far behind compared with the LiDAR-based approaches.

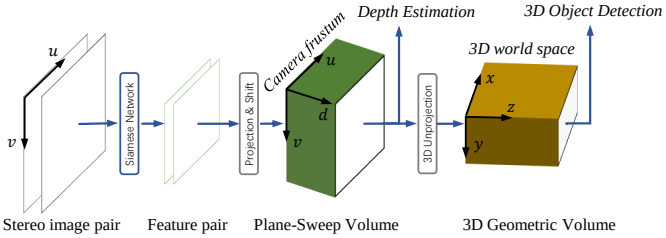

Figure 1. DSGN jointly estimates depth and detects 3D objects from a stereo image pair. It intermediately generates a plane-sweep volume and 3D geometric volume to represent 3D structure in two different 3D space.

Challenges One of the greatest challenges for image-based approaches is to give appropriate and effective representation for predicting 3D objects. Most recent work [25, 36, 48, 37, 40, 2] divides this task into two sub ones, *i.e.*, depth prediction and object detection. Camera projection is a process that maps 3D world into a 2D image. One 3D feature in different object poses causes local appearance changes, making it hard for a 2D network to extract stable 3D information.

Another line of solutions [45, 56, 47, 30] generate intermediate point cloud followed by a LiDAR-based 3D object detector. This 3D representation is less effective since the transformation is non-differentiable and incorporates several independent networks. Besides, the point cloud faces the challenge of object artifacts [18, 47, 56] that limits the detection accuracy of the following 3D object detector.

Our Solution In this paper, we propose a stereo-based end-to-end 3D object detection pipeline (Figure 1) – Deep Stereo Geometry Network (DSGN), which relies on space transformation from 2D features to an effective 3D structure, called 3D geometric volume (3DGV).

The insight behind 3DGV lies in the approach to construct the 3D volume that encodes 3D geometry. 3D geometric volume is defined in 3D world space, transformed from a plane-sweep volume (PSV) [10, 11] constructed in the camera frustum. The pixel-correspondence constraint can be well learned in PSV, while 3D features for real-world

objects can be learned in 3DGV. The volume construction is fully differentiable and thus can be jointly optimized for learning of both stereo matching and object detection.

This volumetric representation has two key advantages. First, it is easy to impose the pixel-correspondence constraint and encode full depth information into 3D real-world volume. Second, it provides 3D representation with geometry information that makes it possible to learn 3D geometric features for real-world objects. As far as we know, there was no study yet to explicitly investigate the way of encoding 3D geometry into an image-based detection network. Our contribution is summarized as follows.

- To bridge the gap between 2D image and 3D space, we establish stereo correspondence in a plane-sweep volume and then transform it to 3D geometric volume for capability to encode both 3D geometry and semantic cues for prediction in 3D regular space.
- We design an end-to-end pipeline for extracting *pixel-level* features for stereo matching and *high-level* features for object recognition. The proposed network jointly estimates scene depth and detects 3D objects in 3D world, enabling many practical applications.
- Without bells and whistles, our simple and fully-differentiable network outperforms all other stereo-based 3D object detectors (10 points higher in terms of AP) on the official KITTI leaderboard [13].

2. Related Work

We briefly review recent work on stereo matching and multi-view stereo. Then we survey 3D object detection based on LiDAR, monocular images, and stereo images.

Stereo Matching In the field of stereo matching on binocular images, methods of [21, 4, 57, 14, 43, 46] process the left and right images by a Siamese network and construct a 3D cost volume to compute the matching cost. Correlation-based cost volume is applied in recent work [31, 55, 51, 14, 28, 42]. GC-Net [21] forms a concatenation-based cost volume and applies 3D convolution to regress disparity estimates. Recent PSMNet [4] further improves the accuracy by introducing pyramid pooling module and stacks hourglass modules [32]. State-of-the-art methods already achieved less than 2% 3-pixel error on KITTI 2015 stereo benchmark.

Multi-View Stereo Methods of [5, 53, 19, 20, 17, 16] reconstruct 3D objects in a multi-view stereo setting [1, 3].

MVSNet [53] constructs plane-sweep volumes upon a *camera frustum* to generate the depth map for each view. Point-MVSNet [5] instead intermediately transforms the plane-sweep volume to point cloud representation to save computation. Kar *et al*. [20] proposed the differentiable projection and unprojection operation on multi-view images.

LiDAR-based 3D Detection LiDAR sensors are very powerful, proven by several leading 3D detectors. Generally two types of architectures, i.e., voxel-based approaches [58, 9, 24, 54] and point-based approaches [34, 35, 39, 52, 49], were proposed to process point cloud.

Image-based 3D Detection Another line of detection is based on images. Regardless of monocular- or stereo-based setting, methods can be classified into two types according to intermediate representation existence.

3D detector with depth predictor: the solution relies on 2D image detectors and depth information extraction from monocular or stereo images. Stereo R-CNN$_{\mathrm{Stereo}}$ [25] formulates 3D detection into multiple branches/stages to explicitly resolve several constraints. We note that the keypoint constraint may be hard to generalize to other categories like *Pedestrian*, and the dense alignment for stereo matching directly operating raw RGB images may be vulnerable to occlusion.

MonoGRNet$_{Mono}$ [36] consists of four subnetworks for progressive 3D localization and directly learning 3D information based solely on semantic cues. MonoDIS$_{\mathrm{Mono}}$ [40] disentangles the loss for 2D and 3D detection. It achieves both tasks in an end-to-end manner. M3D-RPN$_{\mathrm{Mono}}$ [2] applies multiple 2D convolutions of non-shared weights to learn location-specific features for joint prediction of 2D and 3D boxes. Triangulation$_{\mathrm{Stereo}}$ [37] directly learns offset from predefined 3D anchors on bird's eye view and establishes object correspondence on RoI-level features. Due to low resolutions, pixel correspondence is not fully exploited.

3D representation based 3D Detector: 3DOP$_{\mathrm{Stereo}}$ [7, 8] generates point cloud by stereo and encodes the prior knowledge and depth in an energy function. Several methods [45, 56, 47, 30] transform the depth map to Pseudo-LiDAR (point cloud) intermediately followed by another independent network. This pipeline yields large improvement over previous methods. OFT-Net$_{\mathrm{Mono}}$ [38] maps image feature into an orthographic bird's eye view representation and detects 3D objects on bird's eye view.

3. Our Approach

In this section, we first explore the proper representation for 3D space and motivate our network design. Based on the discussion, we present our complete 3D detection pipeline under a binocular image pair setting.

3.1. Motivation

Due to perspective, objects appear smaller with the increase of distance, which makes it possible to roughly estimate the depth according to the relative scale of objects sizes and the context. However, 3D objects of the same category may still have various sizes and orientations. It

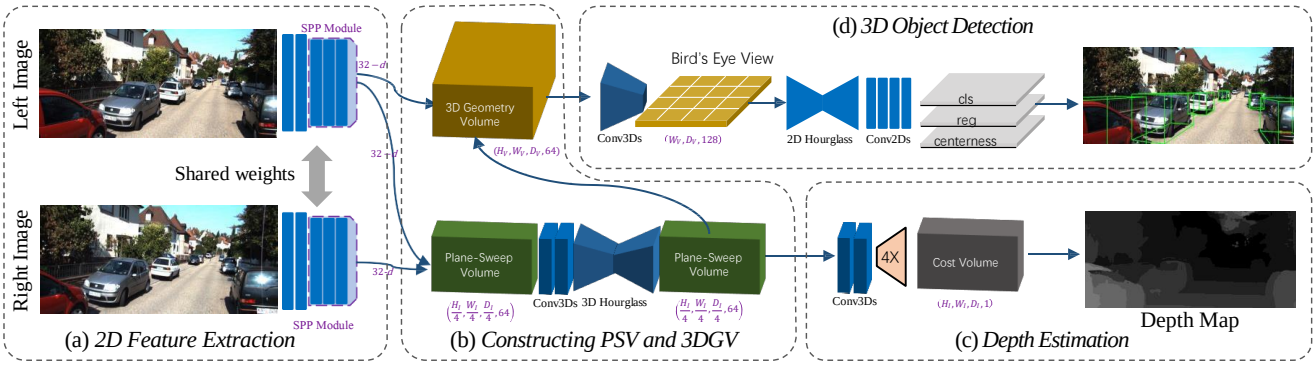

Figure 2. Overview of Deep Stereo Geometry Network (DSGN). The whole neural network consists of four components. (a) A 2D image feature extractor for capture of both *pixel-* and *high-level* feature. (b) Constructing the plane-sweep volume and 3D geometric volume. (c) Depth Estimation on the plane-sweep volume. (d) 3D object detection on 3D geometric volume.

greatly increases the difficulty to make accurate prediction.

Besides, the visual effect of *foreshortening* causes that nearby 3D objects are not scaled evenly in images. A regular cuboid car appears like an irregular frustum. These two problems impose major challenges for 2D neural networks to model the relationship between 2D imaging and real 3D objects [25]. Thus, instead of relying on 2D representation, by reversing the process of projection, an intermediate 3D representation provides a more promising way for 3D object understanding. The following two representations can be typically used in 3D world.

Point-based Representation Current state-of-the-art pipelines [45, 56, 30] generate intermediate 3D structure of point cloud by depth prediction approaches [12, 4, 21] and apply LiDAR-based 3D object detectors. The main possible weakness is that it involves several independent networks and potentially loses information during intermediate transformation, making the 3D structure (such as cost volume) boiled down to point cloud.

This representation often encounters streaking artifacts near object edges [18, 47, 56]. Besides, the network is hard to be differentiated for multi-object scenes [5, 54].

Voxel-based Representation Volumetric representation, as another way of 3D representation, is investigated less intensively. OFT-Net$_{\text{mono}}$ [38] directly maps the image feature to the 3D voxel grid and then collapses it to the feature on bird's eye view. However, this transformation keeps the 2D representation for this view and does not explicitly encode the 3D geometry of data.

Our Advantage The key to establishment of an effective 3D representation relies on the ability to encode accurate 3D geometric information of the 3D space. A stereo camera provides an explicit pixel-correspondence constraint for computing depth. Aiming to design a unified network to exploit this constraint, we explore deep architectures capable of extracting both *pixel-level* features for stereo correspondence and *high-level* features for semantic cues.

On the other hand, the pixel-correspondence constraint is supposedly imposed along the projection ray through each pixel where the depth is considered to be *definite*. To this end, we create an intermediate plane-sweep volume from a binocular image pair to learn stereo correspondence constraint in camera frustum and then transform it to a 3D volume in 3D space. In this 3D volume with 3D geometric information lifted from the plane-sweep volume, we are able to well learn 3D features for real-world objects.

3.2. Deep Stereo Geometry Network

In this subsection, we describe our overall pipeline – Deep Stereo Geometry Network (DSGN) as shown in Figure 2. Taking the input of a binocular image pair (I_L, I_R), we extract features by a Siamese network and construct a plane-sweep volume (PSV). The pixel-correspondence is learned on this volume. By differentiable warping, we transform PSV to a 3D geometric volume (3DGV) to establish 3D geometry in *3D world space*. Then the following 3D neural network on the 3D volume learns necessary structure for 3D object detection.

3.2.1 Image Feature Extraction

Networks for stereo matching [21, 4, 14] and object recognition [15, 41] have different architecture designs for their respective tasks. To ensure reasonable accuracy of stereo matching, we adopt the main design of PSMNet [4].

Because the detection network requires a discriminative feature based on high-level semantic features and large context information, we modify the network for grasping more high-level information. Besides, the following 3D CNN for cost volume aggregation takes much more computation, which gives us room to modify the 2D feature extractor without introducing extra heavy computation overhead in the overall network.

Network Architecture Details Here we use the notations `conv_1`, `conv_2`, ..., `conv_5` following [15]. The key modification for 2D feature extractor is as follows.

- Shift more computation from `conv_3` to `conv_4` and `conv_5`, *i.e.*, changing the numbers of basic blocks of `conv_2` to `conv_5` from $\{3, 16, 3, 3\}$ to $\{3, 6, 12, 4\}$.
- The SPP module used in PSMNet concatenates the output layers of `conv_4` and `conv_5`.
- The output channel number of convolutions in `conv_1` is 64 instead of 32 and the output channel number of a basic residual block is 192 instead of 128.

Full details of our 2D feature extraction network are included in the supplementary material.

3.2.2 Constructing 3D Geometric Volume

To learn 3D convolutional features in 3D regular space, we first create a 3D geometric volume (3DGV) by warping a plane-sweep volume to 3D regular space. Without loss of generality, we discretize the region of interest in *3D world space* to a 3D voxel occupancy grid of size (W_V, H_V, D_V) along the right, down and front directions in camera view. W_V, H_V, D_V denote the width, height and length of the grid, respectively. Each voxel is of size (v_w, v_h, v_d).

Plane-Sweep Volume In binocular vision, an image pair (I_L, I_R) is used to construct a disparity-based cost volume for computing matching cost, which matches a pixel i in the left image I_L to the correspondence in the right image I_R horizontally shifted by an integral disparity value d. The depth is inversely proportional to disparity.

It is thus hard to distinguish among distant objects due to the similar disparity values [25, 45, 56]. For example, objects 40-meter and 39-meter away have almost no difference (< 0.25pix) on disparity on KITTI benchmark [13].

In a different way to construct the cost volume, we follow the classic plane sweeping approach [10, 11, 53] to construct a plane-sweep volume by concatenating the left image feature F_L and the reprojected right image feature $F_{R->L}$ at equally spaced depth interval, which avoids imbalanced mapping of features to 3D space.

The coordinate of PSV is represented by (u, v, d), where (u, v) represents (u, v)-pixel in the image and it adds another axis orthogonal to the image plane for depth. We call the space of (u, v, d) grid *camera frustum space*. The depth candidates d_i are uniformly sampled along the depth dimension with interval v_d following the pre-defined 3D grid. Concatenation-based volume enables the network to learn semantic features for object recognition.

We apply 3D convolution to this volume and finally get a matching cost volume for all depth. To ease computation, we apply only one 3D hourglass module, contrary to the three used in PSMNet [4]. We note that the resulting performance degradation can be compensated in the following detection network since the overall network is differentiable.

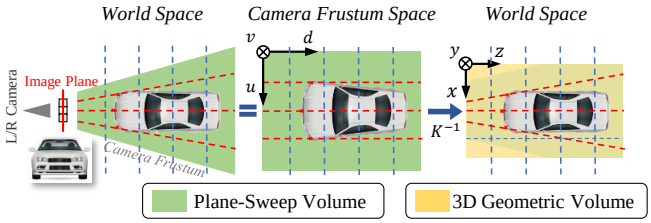

Figure 3. Illustration of volume transformation. The image is captured at the image plane (red solid line). PSV is constructed by projecting images at equally spaced depth (blue dotted lines) in left camera frustum, which is shown in the *3D world space* (left) and *camera frustum space* (middle). Car is shown to be distorted in the middle. Mapping by the camera intrinsic matrix K, PSV is warped to 3DGV, which restores the car.

3D Geometric Volume With known camera internal parameters, we transform the last feature map of PSV before computing matching cost from *camera frustum space* (u, v, d) to *3D world space* (x, y, z) by reversing 3D projection with

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1/f_x & 0 & -c_u/f_x \\ 0 & 1/f_y & -c_v/f_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ud \\ vd \\ d \end{pmatrix} \quad (1)$$

where f_x, f_y are the horizontal and vertical focal lengths. This transformation is fully-differentiable and saves computation by eliminating background outside the pre-defined grid, such as the sky. It can be implemented by warp operation with *trilinear* interpolation.

Figure 3 illustrates the transformation process. The common pixel-correspondence constraint (red dotted lines) is imposed in *camera frustum* while object recognition is learned in regular *3D world space* (*Euclidean space*). There obviously is difference in these two representations.

In the last feature map of plane-sweep volume, a low-cost voxel (u, v, d) means the high probability of object existing at depth d along the ray through the focal point and image point (u, v). With the transformation to regular *3D world space*, the feature of low cost suggests that this voxel is occupied in the front surface of the scene, which can serve as a feature for 3D geometric structure. Thus it is possible for the following 3D network to learn 3D object features on this volume.

This operation is fundamentally different from differentiable unprojection [20], which directly lifts the image feature from 2D image frame to 3D world by *bilinear* interpolation. Our goal is to lift geometric information from cost volume to 3D world grid. We make pixel-correspondence constraint easy to be imposed along the projection ray.

The contemporary work [56] applies a similar idea to

construct depth-cost-volume like plane-sweep volume. Differently, we aim to avoid imbalanced warping from plane-sweep volume to 3D geometric volume, and deal with the streaking artifact problem. Besides, our transformation keeps the distribution of depth instead of deducting it to a depth map. Our strategy intriguingly avoids object artifacts.

3.2.3 Depth Regression on Plane-Sweep Cost Volume

To compute the matching cost on the plane-sweep volume, we reduce the final feature map of plane-sweep volume by two 3D convolutions to get 1D cost volume (called plane-sweep cost volume). Soft arg-min operation [21, 4, 57] is applied to compute the expectation for all depth candidates with probability $\sigma(-c_d)$ as

$$\hat{d} = \sum_{d \in \{z_{\min}, z_{\min}+v_d, \ldots, z_{\max}\}} d \times \sigma(-c_d) \tag{2}$$

where the depth candidates are uniformly sampled within pre-defined grid $[z_{\min}, z_{\max}]$ with interval v_d. The softmax function encourages the model to pick a single depth plane per pixel.

3.2.4 3D Object Detector on 3D Geometric Volume

Motivated by recent one-stage 2D detector FCOS [44], we extend the idea of *centerness* branch in our pipeline and design a distance-based strategy to assign targets for the real world. Because objects of the same category are of similar size in 3D scene, we still keep the design of anchors.

Let $\mathcal{V} \in \mathbb{R}^{W \times H \times D \times C}$ be the feature map for 3DGV of size (W, H, D) and denote the channels as C. Considering the scenario of autonomous driving, we gradually downsample along the height dimension and finally get the feature map $\mathcal{F}$ of size (W, H) for bird's eye view. The network architecture is included in the supplementary material.

For each location (x, z) in $\mathcal{F}$, several anchors of different orientations and sizes are placed. Anchors $\mathbf{A}$ and ground-truth boxes $\mathbf{G}$ are represented by the location, prior size and orientation, *i.e.*, $(x_\mathbf{A}, y_\mathbf{A}, z_\mathbf{A}, h_\mathbf{A}, w_\mathbf{A}, l_\mathbf{A}, \theta_A)$ and $(x_\mathbf{G}, y_\mathbf{G}, z_\mathbf{G}, h_\mathbf{G}, w_\mathbf{G}, l_\mathbf{G}, \theta_\mathbf{G})$. Our network regresses from anchor and gets the final prediction $(h_\mathbf{A} e^{\delta h}, w_\mathbf{A} e^{\delta w}, l_\mathbf{A} e^{\delta l}, x_\mathbf{A} + \delta x, y_\mathbf{A} + \delta y, z_\mathbf{A} + \delta z, \theta_\mathbf{A} + \pi / N_\theta \tanh(\delta\theta))$, where N_θ denotes the number of anchor orientations and $\delta\cdot$ is the learned offset for each parameter.

Distance-based Target Assignment Taking object orientation into consideration, we propose distance-based target assignment. The distance is defined as the distance of 8 corners between anchor and ground-truth boxes as

$$distance(\mathbf{A}, \mathbf{G}) = \frac{1}{8} \sum_{i=1}^{8} \sqrt{(x_{\mathbf{A}_i} - x_{\mathbf{G}_i})^2 + (z_{\mathbf{A}_i} - z_{\mathbf{G}_i})^2)}$$

In order to balance the ratio of positive and negative samples, we let the anchors with top N nearest distance to ground-truth as positive samples, where $N = \gamma \times k$ and k is the number of voxels inside ground-truth box on bird's eye view. γ adjusts the number of positive samples. Our centerness is defined as the exponent of the negative normalized distance of eight corners as

$$centerness(\mathbf{A}, \mathbf{G}) = e^{-\text{norm}(distance(\mathbf{A}, \mathbf{G}))}, \tag{3}$$

where norm denotes min-max normalization.

3.3. Multi-task Training

Our network with stereo matching network and 3D object detector is trained in an end-to-end fashion. We train the overall 3D object detector with a multi-task loss as

$$Loss = \mathcal{L}_{depth} + \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{centerness}. \tag{4}$$

For the loss of depth regression, we adopt smooth L_1 loss [21] in this branch as

$$\mathcal{L}_{depth} = \frac{1}{N_D} \sum_{i=1}^{N_D} smooth_{L_1} \left(d_i - \hat{d}_i \right), \tag{5}$$

where N_D is the number of pixels with ground-truth depth (obtained from the sparse LiDAR sensor).

For the loss of classification, focal loss [29] is adopted in our network to deal with the class imbalance problem in 3D world as

$$\mathcal{L}_{cls} = \frac{1}{N_{pos}} \sum_{(x,z) \in \mathcal{F}} \textit{Focal\_Loss}(p_{\mathbf{A}_{(x,z)}}, p_{\mathbf{G}_{(x,z)}}), \tag{6}$$

where N_{pos} denotes the number of positive samples. Binary cross-entropy (BCE) loss is used for centerness.

For the loss of 3D bounding box regression, smooth L_1 Loss is used for the regression of bounding boxes as

$$\begin{aligned} \mathcal{L}_{reg} = \frac{1}{N_{pos}} \sum_{(x,z) \in F_{pos}} & \text{centerness}(\mathbf{A}, \mathbf{G}) \times \\ & \text{smooth}_{L_1}(l1\_\text{distance}(\mathbf{A}, \mathbf{G})) \end{aligned} \tag{7}$$

where F_{pos} denotes all positive samples on bird's eye view.

We try two different regression targets with and without jointly learning all parameters.

- *Separably optimizing box parameters*. The regression loss is directly applied to the offset of $(x, y, z, h, w, l, \theta)$.
- *Jointly optimizing box corners*. For jointly optimizing box parameters, the loss is made on the average *L1* distance of eight box corners between predicted boxes from 3D anchors and ground-truth boxes following that of [33].

In our experiments, we use the second regression target for *Car* and the first regression target for *Pedestrian* and *Cyclist*. Because it is hard for even human to accurately predict or annotate the orientation of objects like *Pedestrian* from an image, other parameter estimation under joint optimization can be affected.

4. Experiments

Datasets Our approach is evaluated on the popular KITTI 3D object detection dataset [13], which is union of $7,481$ stereo image-pairs and point clouds for training and $7,518$ for testing. The ground-truth depth maps are generated from point clouds following [45, 56]. The training data has annotation for *Car*, *Pedestrian* and *Cyclist*. The KITTI leaderboard limits the access to submission to the server for evaluating test set. Thus, following the protocol in [9, 25, 45], the training data is divided into a training set (3,712 images) and a validation set (3,769 images). All ablation studies are conducted on the split. For the submission of our approach, our model is trained from scratch on the 7K training data only.

Evaluation Metric KITTI has three levels of difficulty setting of easy, moderate (main index) and hard, according to the occlusion/truncation and the size of an object in the 2D image. All methods are evaluated for three levels of difficulty under different IoU criteria per class , *i.e.*, IoU ≥ 0.7 for *Car* and IoU ≥ 0.5 for *Pedestrian* and *Cyclist* for 2D, bird's eye view and 3D detection.

Following most image-based 3D object detection setting [45, 25, 37, 36, 2, 40], the ablation experiments are conducted on *Car*. We also report the results of *Pedestrian* and *Cyclist* for reference in the supplementary file. KITTI benchmark recently changes evaluation where AP calculation uses 40 recall positions instead of the 11 recall positions proposed in the original Pascal VOC benchmark. Thus, we show the main test results following the official KITTI leaderboard. We generate the validation results using the original evaluation code for fair comparison with other approaches in ablation studies.

4.1. Implementation

Training Details By default, models are trained on 4 *NVIDIA Tesla V100* (32G) GPUs with batch-size 4 – that is, each GPU holds one pair of stereo images of size 384×1248. We apply *ADAM* [22] optimizer with initial learning rate 0.001. We train our network for 50 epochs and the learning rate is decreased by 10 at 50-th epoch. The overall training time is about 17 hours. The data augmentation used is horizontal flipping only.

Following other approaches [56, 45, 58, 50, 39, 54], another network is trained for *Pedestrian* and *Cyclist*, we first pre-train the network with all training images for the stereo network and then apply fine-tune with 3D box annotation for both branches because only about $1/3$ images have annotations of these two objects.

Implementation Details For constructing plane-sweep volume, the image feature map is shrunk to 32D and downsampled by 4 for both left and right images. Then by reprojection and concatenation, we construct the volume of shape $(W_I/4, H_I/4, D_I/4, 64)$, where the image size is $(W_I = 1248, H_I = 384)$ and the number of depth is $D_I = 192$. It is followed by one 3D hourglass module [4, 32] and extra 3D convolutions to get the matching cost volume of shape $(W_I/4, H_I/4, D_I/4, 1)$. Then interpolation is used to upsample this volume to fit the image size.

To construct 3D geometric volume, We discretize the region in range $[-30.4, 30.4] \times [-1, 3] \times [2, 40.4]$ (meters) to a 3D voxel occupancy grid of size $(W_V = 300, H_V = 20, D_V = 192)$ along the right (X), down (Y) and front (Z) directions in camera's view. 3D geometric volume is formed by warping the last feature map of PSV. Each voxel is a cube of size $(0.2, 0.2, 0.2)$ (meter).

Other implementation details and the network architecture are included in the supplementary file.

4.2. Main Results

We give comparison with state-of-the-art 3D detectors in Tables 1 and 2. Without bells and whistles, our approach outperforms all other image-based methods on 3D and BEV object detection. We note that Pseudo-LiDARs [45, 56] is with pre-trained PSMNet [4] on a large-scale synthetic scene flow dataset [31] (with 30,000+ pairs of stereo images and dense disparity maps) for stereo matching. Stereo R-CNN [25] uses ImageNet pre-trained ResNet-101 as backbone and has input images of resolution 600×2000.

Differently, our model is trained from scratch only on these 7K training data with input of resolution 384×1248. Also, Pseudo-LiDARs [45, 56] approaches apply two independent networks including several LiDAR-based detectors, while ours is just one unified network.

DSGN without explicitly learning 2D boxes surpasses those applying strong 2D detectors based on ResNet-101 [25] or DenseNet-121 [2]. It naturally achieves duplicate removal by non-maximum suppression (NMS) in 3D space, which coincides with the common belief that there is no collision between regular objects.

More intriguingly, as shown in Table 1, DSGN even achieves comparable performance on BEV detection and better performance on 3D detection on KITTI easy regime with MV3D [9] (with LiDAR input only) – a classic LiDAR-based 3D object detector, *for the first time*. This result demonstrates a promising future application at least in the scenario of low-speed autonomous driving.

The above comparison manifests the effectiveness of 3D geometric volume, which serves as a link between 2D im-

Modality	Method	**3D Detection AP (%)**			BEV Detection AP (%)			2D Detection AP (%)		
		Easy	**Moderate**	Hard	Easy	**Moderate**	Hard	Easy	**Moderate**	Hard
LiDAR	MV3D (LiDAR) [9]	68.35	54.54	49.16	86.49	78.98	72.23	–	–	–
Mono	OFT-Net [38]	1.61	1.32	1.00	7.16	5.69	4.61	–	–	–
	MonoGRNet [36]	9.61	5.74	4.25	18.19	11.17	8.73	88.65	77.94	63.31
	M3D-RPN [2]	14.76	9.71	7.42	21.02	13.67	10.23	89.04	85.08	69.26
	AM3D [30]	16.50	10.74	9.52	25.03	17.32	14.91	92.55	**88.71**	77.78
Stereo	3DOP [7]	–	–	–	–	–	–	93.04	88.64	**79.10**
	Stereo R-CNN* [25]	47.58	30.23	23.72	61.92	41.31	33.42	93.98	85.98	71.25
	PL: AVOD* [45]	54.53	34.05	28.25	67.30	45.00	38.40	85.40	67.79	58.50
	PL++: P-RCNN* [56]	61.11	42.43	36.99	78.31	58.01	51.25	94.46	82.90	75.45
	DSGN (Ours)	**73.50**	**52.18**	**45.14**	**82.90**	**65.05**	**56.60**	**95.53**	86.43	78.75

Table 1. Comparison of main results on KITTI test set (official KITTI leaderboard). The results are evaluated using new evaluation metric on the KITTI leaderboard. Several methods undergoing old evaluation are not available on the leaderboard. PL/PL++* uses extra Scene Flow dataset to pre-train the stereo matching network and Stereo R-CNN* uses ImageNet pre-trained model.

Modality	Method	**3D Detection AP (%)**			BEV Detection AP (%)			2D Detection AP (%)		
		Easy	**Moderate**	Hard	Easy	**Moderate**	Hard	Easy	**Moderate**	Hard
LiDAR	MV3D (LiDAR) [9]	71.29	56.60	55.30	86.18	77.32	76.33	88.41	87.76	79.90
Mono	OFT-Net [38]	4.07	3.27	3.29	11.06	8.79	8.91	–	–	–
	MonoGRNet [36]	13.88	10.19	7.62	43.75	28.39	23.87	–	78.14	–
	M3D-RPN [2]	20.27	17.06	15.21	25.94	21.18	17.90	90.24	83.67	67.69
	AM3D [30]	32.23	21.09	17.26	43.75	28.39	23.87	–	–	–
Stereo	MLF [48]	–	9.80	–	–	19.54	–	–	–	–
	3DOP [7]	6.55	5.07	4.10	12.63	9.49	7.59	–	–	–
	Triangulation [37]	18.15	14.26	13.72	29.22	21.88	18.83	–	–	–
	Stereo R-CNN* [25]	54.1	36.7	31.1	68.5	48.3	41.5	**98.73**	**88.48**	71.26
	PL: F-PointNet* [45]	59.4	39.8	33.5	72.8	51.8	33.5	–	–	–
	PL: AVOD* [45]	61.9	45.3	39.0	74.9	56.8	49.0	–	–	–
	PL++: AVOD* [56]	63.2	46.8	39.8	77.0	63.7	56.0	–	–	–
	PL++: PIXOR* [56]	–	–	–	79.7	61.1	54.5	–	–	–
	PL++: P-RCNN* [56]	67.9	50.1	45.3	82.0	**64.0**	57.3	–	–	–
	DSGN (Ours)	**72.31**	**54.27**	**47.71**	**83.24**	63.91	**57.83**	89.25	83.59	**78.45**

Table 2. Comparison of main results on KITTI *val* set. As described in Section 4, we use original KITTI evaluation metric here. PL/PL++* uses extra Scene Flow dataset to pre-train the stereo matching network and Stereo R-CNN* uses ImageNet pre-trained model.

ages and 3D space by combining the depth information and semantic feature.

Inference Time On a NVIDIA Tesla V100 GPU, the inference time of DSGN for one image pair is 0.682s on average, where 2D feature extraction for left and right images takes 0.113s, constructing the plane-sweep volume and 3D geometric volume takes 0.285s, and 3D object detection on 3D geometric volume takes 0.284s. The computation bottleneck of DSGN lies on 3D convolution layers.

4.3. Ablation Study

4.3.1 Ablation study of 3D Volume Construction

One of the main obstacles to construct an effective 3D geometric representation is the appropriate way of learning 3D geometry. We therefore investigate the effect of following three key components to construct a 3D volume.

Input Data Monocular-based 3D volume only has the potential to learn the correspondence between 2D and 3D feature, while stereo-based 3D volume can learn extra 2D feature correspondence for pixel-correspondence constraint.

Constructing 3D Volume One straightforward solution to construct 3D volume is by directly projecting the image feature to 3D voxel grid [20, 38] (denoted as IMG→3DV). Another solution in Figure 3 transforms plane-sweep volume or disparity-based cost volume to 3D volume, which provides a natural way to impose pixel-correspondence constraint along the projection ray in camera frustum (denoted as IMG→(PS)CV→3DV).

Supervising Depth Supervised with or without the point cloud data, the network learns the depth explicitly or implicitly. One way is to supervise the voxel occupancy of 3D grid by ground-truth point cloud using *binary cross-entropy loss*. The second is to supervise depth on the plane-sweep cost volume as explained in Section 3.3.

For fair comparison, the models IMG→3DV and IMG→(PS)CV→3DV have the same parameters by adding the same 3D hourglass module for the model IMG→3DV. In addition, several important facts can be revealed from

Input	Transformation	Supervision	AP_{3D} / AP_{BEV} / AP_{2D}
Mono	IMG→3DV	×	6.22 / 11.98 / 58.23
		3DV	13.66 / 19.92 / 65.89
Stereo	IMG→3DV	×	11.03 / 15.17 / 57.30
		3DV	42.57 / 54.55 / 81.86
	IMG→CV→3DV	CV	45.89 / 58.40 / 81.71
	IMG→PSCV→3DV	×	38.48 / 52.85 / 77.83
		PSCV	**54.27 / 63.91 / 83.59**

Table 3. Ablation study of depth encoded approaches. "PSCV" and "3DV" with "Supervision" header represent that the constraint is imposed in (plane-sweep) cost volume and 3D volume, respectively. The results are evaluated in moderate level.

Table 3 and are explained in the following.

Supervision of point cloud is important. The approaches under the supervision of the LiDAR point cloud consistently perform better than those without supervision, which demonstrates the importance of 3D geometry for image-based approaches.

Stereo-based approaches work much better than monocular ones under supervision. The discrepancy between stereo and monocular approaches indicates that direct learning of 3D geometry from semantic cues is a quite difficult problem. In contrast, image-based approaches without supervision make these two lines yield similar performance, which indicates that supervision only by 3D bounding boxes is insufficient for learning of 3D geometry.

Plane-sweep volume is a more suitable representation for 3D structure. Plane-sweep cost volume (54.27 AP) performs better than disparity-based cost volume (45.89 AP). It shows that balanced feature mapping is important during the transformation to 3D volume.

Plane-sweep volume, as an intermediate encoder, more effectively contains depth information. The inconsistency between IMG→PSCV→3DV and IMG→3DV manifests that plane-sweep volume as the intermediate representation can effectively help learning of depth information. The observation explains that the *soft arg-min* operation encourages the model to pick a single depth plane per pixel along the projection ray, which shares the same spirit as the assumption that only one depth-value is true for each pixel. Another reason can be that PSCV and 3DV have different matching densities – PSCV intermediately imposes the dense pixel correspondence over all image pixels. In contrast, only the left-right pixel pairs through the voxel centers are matched on 3DV.

From above comparison of volume construction, we observe that the three key facts affect the performance of computation pipelines. The understanding and recognition of how to construct a suitable 3D volume is still at the very early stage. More study is expected to reach comprehensive understanding of the volume construction from the multi-view images.

Networks	Targets	Depth Error (meters)		AP_{3D} / AP_{BEV} / AP_{2D}
		Mean	Median	
PSMNet-PSV*	Depth	0.5337	0.1093	—-
DSGN		**0.5279**	**0.1055**	—-
PSMNet-PSV*	Both	0.5606	0.1157	46.41 / 57.57 / 80.67
DSGN		0.5586	0.1104	**54.27 / 63.91 / 83.59**

Table 4. Influence on depth estimation, evaluated on KITTI val images. PSMNet-PSV* is a variant of PSMNet [4], which uses one 3D hourglass module instead of three of them for refinement considering limited memory space and takes the plane-sweep approach to construct cost volume.

4.3.2 Influence on Stereo Matching

We conduct experiments for investigating the influence of depth estimation, which is evaluated on KITTI *val* set following [45]. The average and median value of absolute depth estimation errors within the pre-defined range of $[z_{\text{min}}, z_{\text{max}}]$ is shown in Table 4. A natural baseline for our approach is PSMNet-PSV* modified from PSMNet [4] whose 2D feature extractor takes 0.041s while ours takes 0.113s.

Trained with depth estimation branch only, DSGN performs slightly better than PSMNet-PSV* with the same training pipeline in depth estimation. For joint training of both tasks, both approaches suffer from larger and similar depth error (0.5586 meter for DSGN vs. 0.5606 meter for PSMNet-PSV*). Differently, DSGN outperforms the alternatives by 7.86 AP on 3D object detection and 6.34 AP on BEV detection. The comparison indicates that our 2D network extracts better high-level semantic features for object detection.

5. Conclusion

We have presented a new 3D object detector on binocular images. It shows that an end-to-end stereo-based 3D object detection is feasible and effective. Our unified network encodes 3D geometry via transforming the plane-sweep volume to a 3D geometric one. Thus, it is able to learn high-quality geometric structure features for 3D objects on the 3D volume. The joint training lets the network learn both *pixel-* and *high-level* features for the important tasks of stereo correspondence and 3D object detection.

Without bells and whistles, our one-stage approach outperforms other image-based approaches and even achieves comparable performance with a few LiDAR-based approaches on 3D object detection. The ablation study investigates several key components for training 3D volume in Table 3. Although the improvement is clear and explained, our understanding of how the 3D volume transformation works will be further explored in our future work.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. volume 120, pages 153–168. Springer, 2016.

[2] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. 2019.

[3] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, pages 5410–5418, 2018.

[5] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. 2019.

[6] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *CVPR*, 2016.

[7] 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 *NIPS*, pages 424–432, 2015.

[8] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals using stereo imagery for accurate object class detection. volume 40, pages 1259–1272. IEEE, 2017.

[9] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *CVPR*, 2017.

[10] Robert T Collins. A space-sweep approach to true multi-image matching. In *CVPR*, pages 358–363. IEEE, 1996.

[11] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *CVPR*, pages 5515–5524, 2016.

[12] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011, 2018.

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[14] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, 2019.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[16] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *CVPR*, pages 2821–2830, 2018.

[17] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: end-to-end deep plane sweep stereo. 2019.

[18] Saif Imran, Yunfei Long, Xiaoming Liu, and Daniel Morris. Depth coefficients for depth completion. In *CVPR*, June 2019.

[19] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *ICCV*, pages 2307–2315, 2017.

[20] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in neural information processing systems*, pages 365–376, 2017.

[21] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, pages 66–75, 2017.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2014.

[23] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Waslander. Joint 3d proposal generation and object detection from view aggregation. 2018.

[24] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. 2018.

[25] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *CVPR*, pages 7644–7652, 2019.

[26] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3d object detection. In *CVPR*, pages 7345–7353, 2019.

[27] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *ECCV*, 2018.

[28] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *CVPR*, pages 2811–2820, 2018.

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, 2017.

[30] Xinzhu Ma, Zhihui Wang, Haojie Li, Wanli Ouyang, and Pengbo Zhang. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. 2019.

[31] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016.

[32] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016.

[33] 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.

[34] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. 2017.

[35] Charles R. Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, 2017.

[36] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization. In *AAAI*, volume 33, pages 8851–8858, 2019.

[37] Zengyi Qin, Jinglu Wang, and Yan Lu. Triangulation learning network: from monocular to stereo 3d object detection. 2019.

[38] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. 2019.

[39] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. 2018.

[40] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel Lopez-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *ICCV*, 2019.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2014.

[42] Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang. Edgestereo: A context integrated residual pyramid network for stereo matching. In *Asian Conference on Computer Vision*, 2018.

[43] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018.

[44] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. 2019.

[45] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, pages 8445–8453, 2019.

[46] Yan Wang, Zihang Lai, Gao Huang, Brian H Wang, Laurens van der Maaten, Mark Campbell, and Kilian Q Weinberger. Anytime stereo image depth estimation on mobile devices. In *ICRA*, pages 5893–5900. IEEE, 2019.

[47] Xinshuo Weng and Kris Kitani. Monocular 3d object detection with pseudo-lidar point cloud. 2019.

[48] Bin Xu and Zhenzhong Chen. Multi-level fusion based 3d object detection from monocular images. In *CVPR*, pages 2345–2353, 2018.

[49] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *CVPR*, 2018.

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

[51] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *ECCV*, pages 636–651, 2018.

[52] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. In *ICCV*, 2019.

[53] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, pages 767–783, 2018.

[54] Chen Yilun, Liu Shu, Shen Xiaoyong, and Jia Jiaya. Fast point r-cnn. In *ICCV*, 2019.

[55] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *CVPR*, pages 6044–6053, 2019.

[56] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. 2019.

[57] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *CVPR*, pages 185–194, 2019.

[58] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, 2018.