

CubifAE-3D: Monocular Camera Space Cubification on Autonomous Vehicles for Auto-Encoder based 3D Object Detection

Shubham Shrivastava and Punarjay Chakravarty

Ford Greenfield Labs, Palo Alto

{sshriwa5, pchakra5}@ford.com

Abstract

We introduce a method for 3D object detection using a single monocular image. Depth data is used to pre-train an RGB-to-Depth Auto-Encoder (AE). The embedding learnt from this AE is then used to train a 3D Object Detector (3DOD) CNN which is used to regress the parameters of 3D object poses after the encoder from the AE generates a latent embedding from the RGB image. We show that we can pre-train the AE using paired RGB and depth images from simulation data once and subsequently only train the 3DOD network using real data, comprising of RGB images and 3D object pose labels (without the requirement of dense depth). Our 3DOD network utilizes a particular ‘cubification’ of 3D space around the camera, where each cuboid is tasked with predicting N object poses, along with their class and confidence values. The AE pre-training and this method of dividing the 3D space around the camera into cuboids give our method its name - CubifAE-3D. We demonstrate results for monocular 3D object detection on the Virtual KITTI 2, KITTI, and nuScenes datasets for Autonomous Vehicle (AV) perception.

1. Introduction

Detecting objects around the ego-vehicle is one of the fundamental tasks for the perception system on an AV. The perception of objects such as other vehicles, pedestrians, bicycles, and their relative positions in 3D space around the ego-vehicle are then used by the vehicles path planning system to chart a collision free route through it.

Detecting objects as 2D bounding boxes in monocular camera images is a relatively mature technology, and techniques like YOLO [19] work reasonably well in the real world. However, the same cannot be said for detecting these objects in 3D space using monocular cameras. Most methods for 3D object detection available today use high-cost sensors such as LIDAR [30, 24, 22, 31].



Figure 1. Object pose predictions using CubifAE-3D on Monocular RGB camera images from nuScenes (top), KITTI (middle), and Virtual KITTI 2 datasets (bottom).

Some work has also suggested the use of depth-maps to generate pseudo-lidar representation, which can subsequently be used for 3D object detection using state-of-the-art LIDAR based object detection methods [23, 29].

Here, we present a method for performing 3D object detection using a single RGB camera during inference. We assume that a dense depth map (as obtained from stereo / RGB-D camera) is available during training. Sparse depth maps (as obtained from LIDAR) could also be used for training by first transforming it into a dense representation [13, 26]. We first pre-train an RGB-to-depth auto-encoder in a fully-supervised manner and let the latent space learn an RGB-to-depth embedding, after which, the Decoder is

disconnected and the output of the latent space is fed to the 3D Object Detection (3DOD) network, which is trained to predict 3D object poses. We call our model CubifAE-3D, *pronounced Cubify-3D*. The first part of the name refers to the *cubification* or voxellization of monocular camera space as a pre-processing step, and the AE refers to the Auto-Encoding of RGB-to-depth space.

Furthermore, we show that the RGB-to-depth auto-encoder can be pre-trained using *simulation* data. This pre-trained latent encoding of the RGB image can subsequently be used to train the 3DOD network on a separate real dataset. The only annotation needed for the real dataset therefore, is the object pose itself corresponding to each camera frame. We do not use LIDAR or stereo data for training the network on the real dataset (though these sensing sources might be required to annotate the 3D labels themselves). Fig.1 illustrates the 3D object pose predictions from a single RGB image made by our model on the *nuScenes* [3], *KITTI* [6] and *Virtual KITTI 2* [2] datasets.

2. Related Work

3D object detection for AVs has its inception with LIDAR as the primary sensor, in the Darpa Urban Challenge [5] of 2007. The winning entry in that contest - Boss, a robotic SUV from CMU [5], accumulated a few successive LIDAR scans and detected the ground plane. Anything above the ground plane was an obstacle. In addition to using the ground plane, a prior 3D map of the scene can also be used to determine the location of objects in the scene by using localization information to extract the relevant portion of the prior map and subtracting the current point cloud from it [18, 27]. These approaches were popular in the AV community before the advent of Deep Learning for object detection.

More recently, Deep Learning approaches have been used for 3D object detection in LIDAR. These approaches directly regress the parameters of the 3D bounding box circumscribing the detected object from the LIDAR point cloud and do not require prior maps. They operate either directly in the LIDAR point-cloud [16, 31] or in a Bird's Eye View (BEV) projection of the point cloud [28] or a combination of camera image with the BEV representation of the point cloud [8, 9]. The BEV represents a point-cloud in a top-down view, has the advantage of preserving object size and scale throughout the representation and is faster to process, allowing real-time detections. The BEV can be treated as an image, that can be scanned using 2D convolutions that are faster to compute compared to the 3D convolutions [31] required in the native point cloud.

Due to the expense of LIDAR scanners and the simultaneous advances in Deep Learning, camera-only techniques for 3D object detection are gaining currency over the past few years. Pseudo-LIDAR [23, 29] demonstrated that

they could retrieve LIDAR-like point clouds from cameras. Camera images (monocular and stereo) are converted into a depth map, which is unprojected into the 3D camera space to get a point-cloud. 3D object poses are subsequently detected directly in the point cloud using a Point-net-like architecture [16] or converted to a BEV image, before being fed to a CNN for detecting the 3D pose of objects using an AVOD-like architecture [8]. Pseudo-LIDAR++ [29] improves on the 3D pose estimation accuracy by adding a very sparse (4-beam) LIDAR to constrain the errors in the point-cloud generated by stereo depth map unprojection.

Determining the 6-Dof pose of an object from a 2D image is essentially an ill-posed problem, because the camera projection equation squashes the depth dimension so that a 2D point in an image corresponds to a ray in 3D space, and the 3D world point corresponding to its projected 2D point could lie along any point along that ray. Approaches like Mousavian et. al. [15] and mono-GR-net [17] combine Deep Learning with the geometric properties for the rigid bodies of vehicles to constrain this problem. Mousavian et. al. [15] uses Deep Learning to predict the orientation angle and the dimensions of the 3D bounding box corresponding to a 2D detection. This 3D bounding box is then translated to the right location in the camera coordinate frame by minimizing its projection error compared to the 2D detection.

Deep MANTA [4] and Mono3D++ [7] use 2D vehicle part detectors in monocular images to further constrain the determination of their 3D poses using optimization. They have Deep-learnt vehicle part detectors for the location of vehicle parts in the 2D image and use the actual 3D locations of these parts relative to the vehicle coordinate frame (learnt offline) and a 3D-to-2D matching technique like the Perspective-n-Point Projection algorithm [10] to determine the pose of the vehicle in the camera coordinate frame. Mono3D++ [7], in addition to 3D part locations, adds additional priors like depth (from monocular depth estimation) and ground plane detection. Conv-nets are used to predict a coarse orientation and scale of the 3D bounding box (corresponding to its 2D detection), in addition to detecting 2D part-landmarks. These are then combined with the 3D wireframe model of the car along with the depth and ground-plane priors to get the final 3D pose of the vehicle in an energy minimization framework that is solved using the Ceres solver [1].

Our *CubifAE-3D* network takes inspiration from the fast 2D object detector YOLO [19]. YOLO discretizes the camera image space into a 2D grid, where each grid cell is allowed to make a fixed number of 2D bounding box detections, from which only a few detections are selected, based on their confidence score and intersection-over-union (IoU) with other objects. We divide the space around the camera into a 3D grid, that starts by subdividing the camera image into a 4x4 grid and extends outwards into the 3rd (z) axis.

The sub-divisions or *cuboids* in our 3D grid are each allowed to predict N 3D object poses and their confidence scores. In fact, our work also resembles OFT-Net [20], which divides the 3D space around the camera into voxels and pools those 2D images features into the voxels. However, this is an intermediate representation and OFT-Net finally projects these features into an orthographic (BEV), ground-plane grid, with each 2D (ground plane) grid cell being tasked with detecting the position, dimension, orientation and confidence of 3D bounding boxes.

A major advantage of these methods that operate in the 3D voxel space [16, 31] or indeed the orthographic/BEV/ground-plane space [23, 28, 29, 20], over methods that operate in the camera image space is that they can detect objects farther away from the camera, and we further extend this advantage and demonstrate that the accuracy of our method remains constant with distance from the camera.

3. Method

Our method of performing 3D object detection relies on first learning the latent space embeddings for per-pixel RGB-to-depth predictions in an image. This is achieved by training an auto-encoder to predict the dense depth map from a single RGB image. Once trained, the decoder is detached, and the latent space embedding is fed to our 3DOD network. The high-level model architecture is shown in Fig.2 and a detailed model architecture is shown in Fig.3.

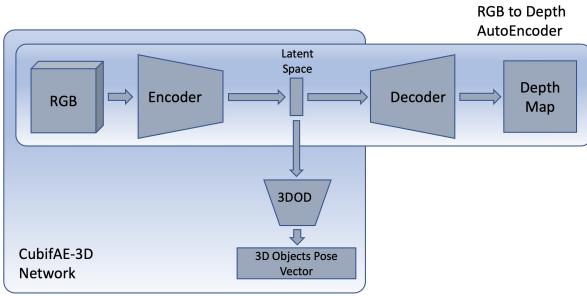


Figure 2. CubifAE-3D high-level architecture

By training the auto-encoder first, we force its latent space to learn a compact RGB-to-depth embedding representation which is encoded in the latent space. A model which then operates on these encodings is thus able to formulate a relationship between the structures present in the RGB image and its real world depth. We then *cubify* the monocular camera space and train our 3DOD model to detect object poses (*xcenter*, *ycenter*, *zcenter*, *width*, *height*, *length*, *orientation*). So, at the test time, only an RGB image is needed for detecting object poses. *Orientation* term in the predicted pose refers to the rotation about *y*-axis in the camera frame. For the sake of simplicity, we assume rotation about the other two axes to be *zero*. An additional

classifier network with a small number of parameters is then used to classify all of these detected objects at once by resizing and stacking the object crops and feeding them to the *classifier* model. This is done instead of predicting the classes directly as a part of the vector corresponding to each object from the 3DOD network in the favour of reducing number of parameters in the fully-connected layers and hence the inference time. We apply *non-max suppression* to further filter out the object pose predictions with high *IoU* by retaining only the objects with the highest *confidence*.

In our experiments, we train the RGB-to-depth AE only in simulation and do not use/require dense depth information from the real datasets. We realize that an RGB-to-depth model trained on one dataset is incompatible with another because of the focal length mismatch. However, because we only use the encoder part of the RGB-to-depth AE as an embedding, our 3DOD network automatically learns to compensate for this. It seems to learn the linear relationships between relative focal lengths and relative depth scales between the simulation and real datasets, regardless of differences in camera intrinsics. We train the latent representation once, using one set of simulation data and use the same latent representation in the 3DOD network which then learns to predict object poses on new real data.

3.1. RGB-to-Depth Auto-Encoder

The RGB-to-depth auto-encoder as shown in Fig. 2 is trained with a combination of a Mean Squared Error (Equation 1) and an Edge-Aware Smoothing Loss (Equation 2) to perform depth map prediction from monocular RGB images. The network contains a U-Net [21] like encoder-decoder architecture with skip connections and is shown in Fig.3.

$$mse_{loss} = \frac{\lambda_{mse}}{u * v} \sum_{i=0, j=0}^{u, v} (d_{i,j} - \hat{d}_{i,j})^2 \quad (1)$$

$$\begin{aligned} eas_{loss} = & \frac{\lambda_{eas}}{u * v} \sum_{i=0, j=0}^{u, v} |\partial_x \hat{d}_{i,j}| e^{-|\partial_x I_{i,j}|} + \\ & |\partial_y \hat{d}_{i,j}| e^{-|\partial_y I_{i,j}|} \end{aligned} \quad (2)$$

This model learns to predict accurate depth maps from monocular RGB images. The edge-aware smoothing loss penalizes high edge gradients if the image gradient is low and vice-versa, thus forcing the depth to be continuous and locally smooth within object boundaries while ensuring a clear depth-difference at the edges. This improves the accuracy of depth along the silhouette of the objects and avoids noisy holes within the object boundaries. Prediction results of this network is shown in Fig.4.

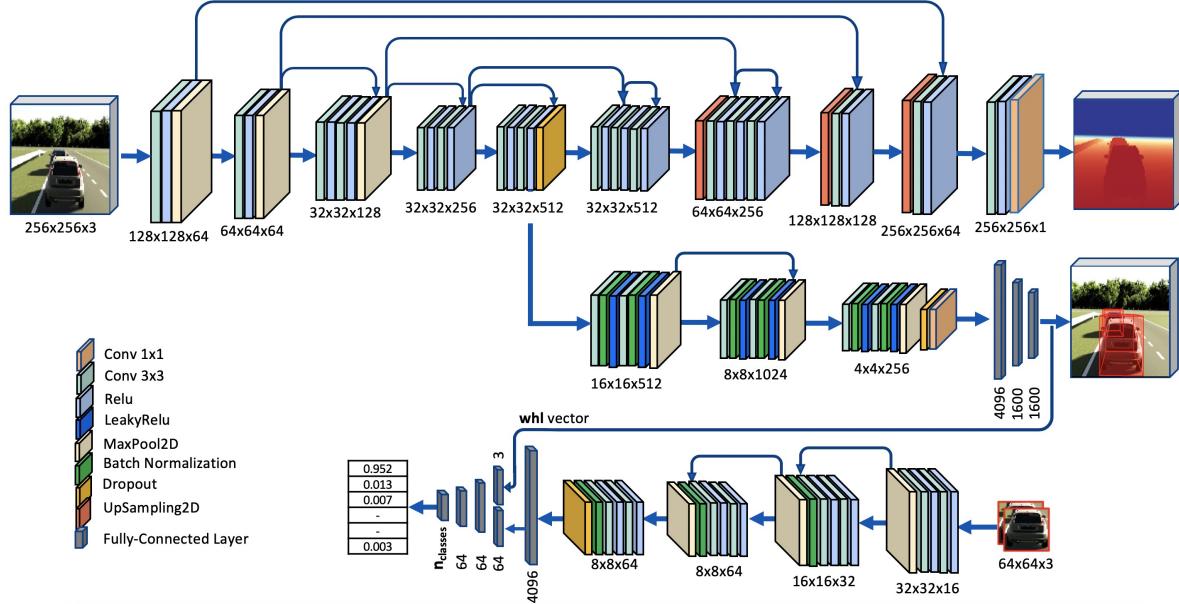


Figure 3. Detailed model architecture of CubifAE-3D. The RGB-to-depth auto-encoder (top branch) is first trained in a supervised way with a combination of *MSE* and *Edge-Aware Smoothing Loss*. Once trained, the decoder is detached, encoder weights are frozen, and the encoder output is fed to the 3DOD model (middle branch), which is trained with a combination of *xyzloss*, *whlloss*, *orientationloss*, *iouloss*, and *confloss*. A 2D bounding-box is obtained for each object by projecting its detected 3D bounding-box onto the camera image plane, cropped, and resized to 64x64 and fed to the *classifier* model (bottom branch) along with the normalized *whl* vector for class prediction. The dimensions indicated correspond to the output tensor for each block.

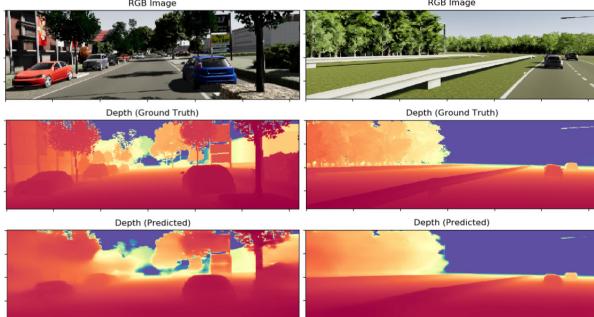


Figure 4. Monocular RGB to Depth Map prediction. (1) First row contains the input RGB image, (2) Second row contains the ground-truth Depth Map, and (3) Third row contains the depth map predictions from our RGB to Depth auto-encoder

3.2. The 3DOD Network

Once the RGB-to-depth auto-encoder (AE) is trained, the head (encoder) of this network is detached, its weights are frozen, and the latent space is then fed to the 3DOD network for object pose estimation. This combined network as shown in Fig.5 thus learns to perform 3D object detection based on just a monocular RGB image as an input. We do not need the ground-truth depth map from real datasets such as *nuScenes* and *KITTI*; we instead pre-train the RGB-to-depth AE in simulation and then use the latent represen-

tation to train the 3DOD network. The camera intrinsics between the simulated and real datasets do not match, but our 3DOD network, given ground truth 3D pose labels for the real dataset, automatically learns to compensate for the depth scale difference caused by the focal length mismatch.

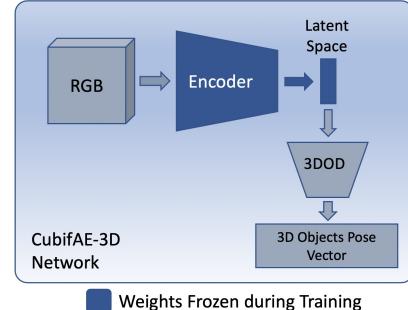


Figure 5. CubifAE-3D Model Architecture (encoder weights are frozen during the training of the 3DOD network).

We prepare training labels for the 3DOD network in a way that allows each part of the network to only be responsible for detecting objects within a certain physical space relative to the ego-vehicle camera. We *cubify* the 3D region-of-interest (ROI) of the ego-camera into a 3-dimensional grid. This 3D grid is of size $4 \times M$, where the camera image plane is divided into 4 regions along the (x,y) dimensions of the camera coordinate frame, with z axis further quantified

into M cuboids for each of these 4 regions. Each cuboid in this $4 \times M$ grid is responsible for predicting up to N objects in an increasing order of z (depth) from the center of the ego-camera. The model predicts a vector of length 8 for each possible object ($confidence, x_{center}, y_{center}, z_{center}, width, height, length, orientation$). This results in the model output being a vector of size $4 \times M \times N \times 8$. The intuition behind dividing the visible ROI into a 4-dimensional grid in the x and y directions comes from the fact that the center of the image corresponds to the optic center of the lens, and hence a 4×4 grid in the image space demarcates the 3D space in the same directions. This simplifies our normalization process given the x and y maximum limits. This also ensures that each grid in the camera plane contains an object within the corresponding 3D space; for example, the top-left grid contains objects with their center at *negative* x and *negative* y 3D coordinates, and the top-right grid contains objects with their center at *positive* x and *negative* y 3D coordinates.

This *Cubification* of the camera space is visualized in Fig.6. We assume the standard camera coordinate system for our work (as used in computer vision applications), where, x points to the right and is aligned with the image x -axis, y points downward and is aligned with the image y -axis, and z points in the direction that the camera is pointed at.

We normalize 3D coordinates of the center of the object (x, y, z) and dimensions ($width, height, length$) between (0, 1) in accordance with a prior that is computed from data statistics. Orientation is normalized by simply scaling the values from $(-\pi, \pi)$ to (0, 1). We use a Sigmoid activation function at the output of the final fully-connected layer for these predictions. Each predicted object ROI with high confidence is then projected to the 2D camera plane, cropped, and resized to a 64×64 patch, which is then fed to a *classifier* network that predicts a class for the detected object.

The total loss function for this model is a weighted sum of 5 individual loss terms for minimizing the detection loss of object center coordinates, object dimensions, orientation, and their detection confidence. We take inspiration from YOLO [19] for the development of these loss functions.

Additionally, we also try to maximize 3D intersection-over-union (IoU) explicitly by minimizing a cross-entropy loss called iou_{loss} . Attempting to minimize this loss improves training by allowing the network to converge faster while improving the accuracy and mean IoU as shown in Fig.7. However, a very small weight, λ_{iou} is chosen for this loss function since it can quickly become unstable because of the \log term.

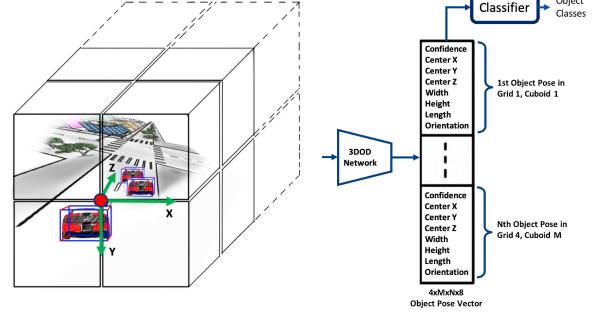


Figure 6. *Cubification* of the camera space: The perception region of interest is divided into a $4 \times 4 \times M$ grid (4×4 in the x and y directions aligned with the camera image plane, where each grid has stacked on it, M cuboids in the z direction). Each cuboid is responsible for predicting up to N object poses. The object coordinates and dimensions are then normalized between 0 and 1 in accordance with a prior that is computed from data statistics.

$$xyz_{loss} = \frac{\lambda_{xyz}}{n_{true_objs}} \sum_{i=0}^4 \sum_{j=0}^M \sum_{k=0}^N T_{ijk}^{n_{objs}} [(x_{ijk} - \hat{x}_{ijk})^2 + (y_{ijk} - \hat{y}_{ijk})^2 + (z_{ijk} - \hat{z}_{ijk})^2] \quad (3)$$

$$w h l_{loss} = \frac{\lambda_{w h l}}{n_{true_objs}} \sum_{i=0}^4 \sum_{j=0}^M \sum_{k=0}^N T_{ijk}^{n_{objs}} [(\sqrt{w_{ijk}} - \sqrt{\hat{w}_{ijk}})^2 + (\sqrt{h_{ijk}} - \sqrt{\hat{h}_{ijk}})^2 + (\sqrt{l_{ijk}} - \sqrt{\hat{l}_{ijk}})^2] \quad (4)$$

$$orientation_{loss} = \frac{\lambda_{orientation}}{n_{true_objs}} \sum_{i=0}^4 \sum_{j=0}^M \sum_{k=0}^N T_{ijk}^{n_{objs}} [(o_{ijk} - \hat{o}_{ijk})^2] \quad (5)$$

$$iou_{loss} = \frac{\lambda_{iou}}{n_{true_objs}} \sum_{i=0}^4 \sum_{j=0}^M \sum_{k=0}^N -T_{ijk}^{n_{objs}} [\log(iou_{ijk})] \quad (6)$$

$$conf_{loss} = \frac{\lambda_{conf}}{4 * N * M} \sum_{i=0}^4 \sum_{j=0}^M \sum_{k=0}^N T_{ijk}^{n_{objs}} [(c_{ijk} - \hat{c}_{ijk})^2] + (1 - T_{ijk}^{n_{objs}})[(c_{ijk} - \hat{c}_{ijk})^2] \quad (7)$$

$$total_{loss} = xyz_{loss} + whl_{loss} + orientation_{loss} + iou_{loss} + conf_{loss} \quad (8)$$

In equations [3 - 8], M is the number of *cuboids* for each (of the 4) 2D image grid-sections, N is the maximum number of possible objects per cuboid, and n_{true_objs} is the number of ground-truth objects. T_{ijk} in these equations indicates whether or not an object appears at that position in the ground-truth label vector, its value being 1 if it does, 0 otherwise.

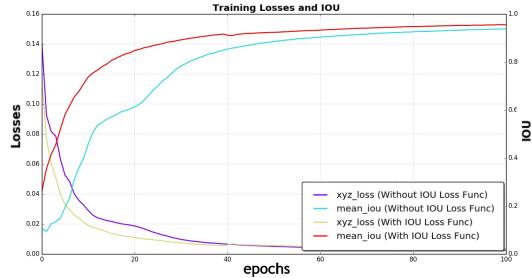


Figure 7. Introducing the iou_{loss} improves error convergence and mean intersection-over-union (IoU). x -axis shows number of training epochs, left y -axis shows the xyz_{loss} which is the mean-squared error of object center coordinate predictions (*lower is better*), right y -axis shows the mean 3D *intersection-over-union* between detected objects and ground-truth (*higher is better*).

3.3. Classifier

The output of our 3DOD model is a vector of objects poses along with a *confidence* value for their prediction. We compute 8 vertices for each bounding-box and project them onto the 2D camera plane. A 2D bounding-box is then obtained for each detected object, cropped, and resized to a 64×64 patch. All object crops are then stacked and fed to the *classifier* network which is responsible for predicting a class for each detected object using *Softmax* activation function. Additionally, we also feed the normalized **whl** (*width*, *height*, *length*) vector for each detected object to the network which is concatenated with the output of first fully-connected layer as shown in Fig.3. This further helps the network establish a relationship between object classes and their dimensions. Our *classifier* model has a few residual blocks followed by fully-connected layers and contains only 430k parameters.

4. Experimental Details

We start with training the RGB-to-depth auto-encoder on the *Virtual KITTI 2* dataset by considering depths up to 100 meters. Any depth value higher than that is set to 100 meters. The dense depth map is then normalized between $[-0.5, 0.5]$. Once trained, the pre-trained encoder of RGB-to-depth auto-encoder is used to train our 3DOD model. We perform 3D Object Detection experiments on the *Virtual KITTI 2* [2], *KITTI* [6], and *nuScenes* [3] datasets. For *Virtual KITTI 2*, we use 18,943 training

and 1,049 validation frames for training. 1000 samples were randomly chosen from the dataset and separated as a test set. We employ a similar strategy for evaluating our method on the *KITTI* and *nuScenes* datasets as well, where we randomly sample 10% of the *training* data and use it as a *test* set. We do not consider small movable objects in *nuScenes*, such as *traffic cone*, *barrier*, and *debris* classes for our detection and classification tasks. During the training of our *classifier* model, we randomly perturb the center of object position (x, y, z) by adding a small amount of noise in the range of $[-0.2$ meters, 0.2 meters] to improve the robustness of classification. We also perform data augmentation in terms in the saturation and brightness space by first converting the RGB images into HSV colorspace and then multiplying the *saturation* (S) and *value* (V) channel by a factor randomly chosen between $[0.5, 1.5]$, and then converting it back to RGB. An example of this augmentation is shown in Fig.8. Other types of commonly used data augmentation techniques such as cropping, rotation, and offsetting the image plane, would need more sophisticated algorithms because of the requirement for the determination of new object poses as a result of image warping and will be considered as future work.



Figure 8. Samples of augmented data (KITTI dataset). *top-left*: original image, *top-right*: lower saturation, higher brightness, *bottom-left*: higher saturation, higher brightness, *bottom-right*: lower saturation, lower brightness.

For our work presented here, we use 5 cuboids per 2D (image plane) grid section, and allow the detection of a maximum of 10 objects per cuboid. As a result, our model is able to predict up to $4 \times 5 \times 10 = 200$ objects within a 100 meter range. We specify a region-of-interest (ROI) and only consider objects within the ROI for our experiments. Furthermore, we obtain a *prior* on object dimensions from the dataset statistics. The ROI around the ego-camera used in our work is: $[x = 40m, y = 10m, z = 100m]$. We compute *priors* for each dataset separately and normalize them between $[0, 1]$ for training. *Priors* used for the datasets were:

Virtual KITTI 2: $[width_{min} = 1.13, width_{max} = 3.02, height_{min} = 1.22, height_{max} = 4.20, length_{min} = 2.22, length_{max} = 16.44]$.

nuScenes: $[width_{min} = 0.27, width_{max} = 3.53, height_{min} = 0.40, height_{max} = 4.46, length_{min} = 0.30, length_{max} = 14.46]$.

KITTI: $[width_{min} = 0.30, width_{max} = 3.01,$

$height_{min} = 0.76$, $height_{max} = 4.20$, $length_{min} = 0.20$, $length_{max} = 35.24$.

We train our models with a batch size of 16, learning rate of 10^{-4} and decay rate of 0.001. The Adam Optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ is used for error optimization. Lambda weights used for loss functions are: $\lambda_{mse} = 0.8$, $\lambda_{eas} = 0.2$, $\lambda_{xyz} = 5.0$, $\lambda_{whl} = 5.0$, $\lambda_{orientation} = 1.0$, $\lambda_{iou} = 0.1$, $\lambda_{conf} = 0.5$. Our complete model (*Encoder of RGB-to-depth AE + 3DOD + Classifier*), has $31.8M$ parameters and runs at 7.2 FPS on a single NVIDIA RTX 2070 GPU during inference. To avoid overfitting, we use dropout layers with rate 0.5 in our auto-encoder and 3DOD model. In the classifier model we employ *L2 regularization* with a factor of 10^{-4} .

5. Results

We compute the *Mean Average Precision* for *Virtual KITTI 2*, *KITTI*, and *nuScenes* dataset using 101-point interpolation as used in MS COCO [11] for *3D IoU* thresholds of 0.3, 0.5, and 0.7 by varying the *confidence* threshold between [0.0, 1.0] across all classes. The Precision-Recall curves for varying IOU thresholds obtained on the *nuScenes* dataset is shown in Fig.9. Additionally, we use the KITTI object detection evaluation kit for evaluating our 3D detection performance on the KITTI object dataset on *Easy*, *Moderate*, and *Hard* difficulty levels (Table 2). We also compare our results with other state-of-the-art methods.

Table 1. Monocular 3D Object Detection results on the Virtual KITTI 2, KITTI, and nuScenes datasets. mAP_x stands for % of *Mean Average Precision* (*mAP*) for *3D IoU* threshold of x . Higher is better.

Method	Dataset	mAP_3	mAP_5	mAP_7
CubifAE-3D	vKITTI 2	86.6	66.7	34.1
CubifAE-3D	nuScenes	77.6	68.6	54.0
CubifAE-3D	KITTI	83.8	59.2	27.0

We also attempt to model the error distribution with respect to the absolute distance from the ego-vehicle for each class. We find that, because of our *cubification* techniques, the error is highly localized within each *cuboid* and does not increase with increasing distance from the ego-vehicle. Errors are found to be the minimum at the center of each *cuboid* and maximum at the beginning and end of each *cuboid* as evident from Fig.10. This figure shows the plot of error distribution for the *car* class. Note that the *cuboid* size in the depth (z) direction was chosen to be 20 *meters* and hence error peaks in z can be seen at 0, 20, 40, 60, 80, and 100 *metres*.

We show qualitative results of our model on *nuScenes* dataset in Fig.13, *KITTI* dataset in Fig.12, and *Virtual KITTI*

Table 2. Monocular 3D Object Detection Results on the KITTI validation split for *car*, *pedestrian*, and *cyclist*. These metrics are evaluated by $AP|_{R11}$ at 0.7 *IoU* threshold for *car*, and 0.5 *IoU* threshold for *pedestrian* and *cyclist*. Higher is better.

Method	Class	Easy	Mod.	Hard
OFT-Net[20]	Car	4.07	3.27	3.29
FQNet[12]	Car	5.98	5.50	4.75
Xu et al.[25]	Car	7.85	5.39	4.73
Mono3D++[7]	Car	10.60	7.90	5.70
ROI-10D[14] (w/o depth)	Car	10.12	1.76	1.30
ROI-10D[14] (w/depth)	Car	7.79	5.16	3.95
ROI-10D[14] (w/depth, syn.)	Car	9.61	6.63	6.29
ours	Car	8.64	7.94	6.42
ours	Pedestrian	5.50	5.43	4.82
ours	Cyclist	6.65	6.54	4.40

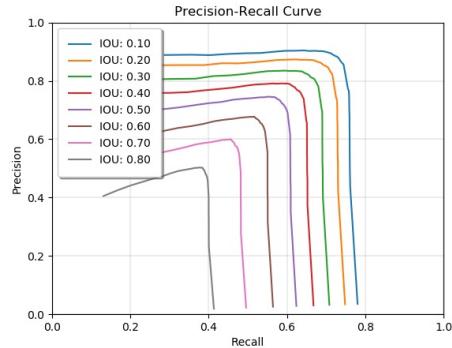


Figure 9. Precision-Recall curves for varying *IoU* on the nuScenes validation set across all classes for distances up to 100 *meters*.

2 dataset in Fig.14.

6. Discussion and Future Work

CubifAE-3D leverages the power of simulation to learn an embedding for RGB-to-depth, which is then used to train a network to predict object poses from RGB images on real datasets. We find that our method does surprisingly well on objects that are far from the ego-vehicle and is represented by only a few pixels, as is evident from our results on the KITTI *moderate* and *hard* difficulty levels. Additionally, with the *cubification* technique, we are able to localize the position error to within each *cuboid* so that it does not increase with increasing distance from the ego-camera. This encourages us to believe that with further quantification of the 3D space, we will be able to achieve higher accuracy and this will be explored in our future work. A non-uniform or a coarse-to-fine *cubification* technique is also something of

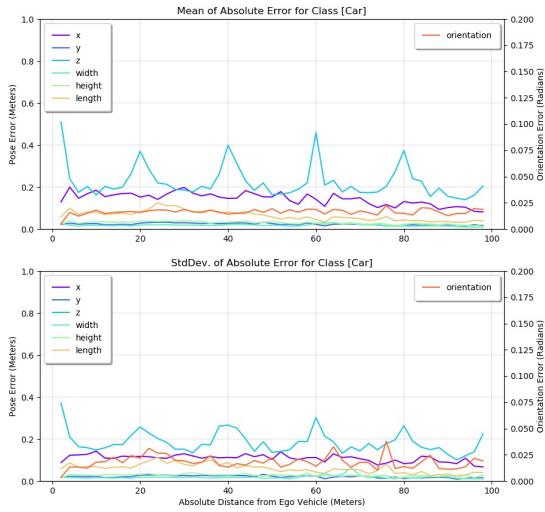


Figure 10. Distribution of errors for the *car* class. Because of the *cubification* method, our errors are localized within each cuboid and does not increase with respect to absolute distance from the ego-vehicle.

interest to fill the gap between short-range and long-range performance and reduce errors at cuboid boundaries. Furthermore, we would like to perform *projective data augmentation* in order to virtually zoom and move the camera through space to increase the amount of labelled 3D pose training data so that the 3DOD network can generalize better on real datasets.

The techniques presented in this paper: pre-training an RGB-to-depth embedding from simulation and *cubifying* 3D space are also likely to be useful for other AV perception tasks that are dependent on pixel-level depth embeddings, without the need for a ground-truth dense depth map, like free space/occupancy grid estimation or 6 DoF monocular localization of the ego-camera/AV.

We show that our method already outperforms some state-of-the-art methods for monocular 3D object detection and we are hopeful that our continuing improvements to *CubifAE-3D* will serve as a baseline for monocular 3D object detection.

7. Conclusions

3D object detection is a primary task in the perception pipeline for an AV and we present *CubifAE-3D*, a method for doing this using monocular imagery. Our method utilizes dense depth information corresponding to RGB images from Gaming Engine based simulators and this is used to pre-train a RGB-to-depth auto-encoder (AE). The encoder of this pre-trained AE is then detached and used as

an embedder for RGB images. This embedding is subsequently used to train our 3D Object Detector (3DOD), which is trained in a fully supervised manner and requires RGB images and corresponding 3D object pose annotations. Our 3DOD *cubifies* the 3D space around the camera into a set of cuboid grids, where each cuboid in the grid is trained to output N 3D object poses and their confidence scores. Once trained, our aggregate detector *CubifAE-3D* (comprising of the AE-encoder and the 3DOD) is able to generate accurate 3D object poses from RGB images. The particular scheme of *cubification* of the 3D space gives results whose accuracy is maintained with distance from the camera. We believe that *CubifAE-3D* represents an important advancement in the field, for fast, accurate 3D object detection from monocular images. The ability to do this robustly and accurately from cheap cameras (instead of expensive LIDAR) that are already present in a surround-view configuration in present generation commercial vehicles will be crucial in the development and democratization of AV technology.

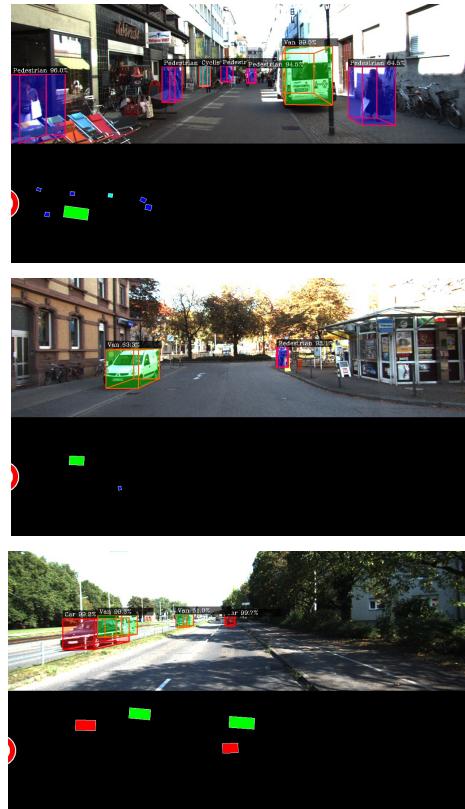


Figure 11. Samples of qualitative results on the KITTI dataset. More results to follow on the next page. The top part of each image shows a bounding box obtained as a 2D projection of their 3D poses (red: car; green: van; blue: pedestrian). The bottom part (black background) shows a birds-eye view of the object poses with the ego-vehicle positioned at the center of red circle drawn on the left; pointing towards the right of the image.

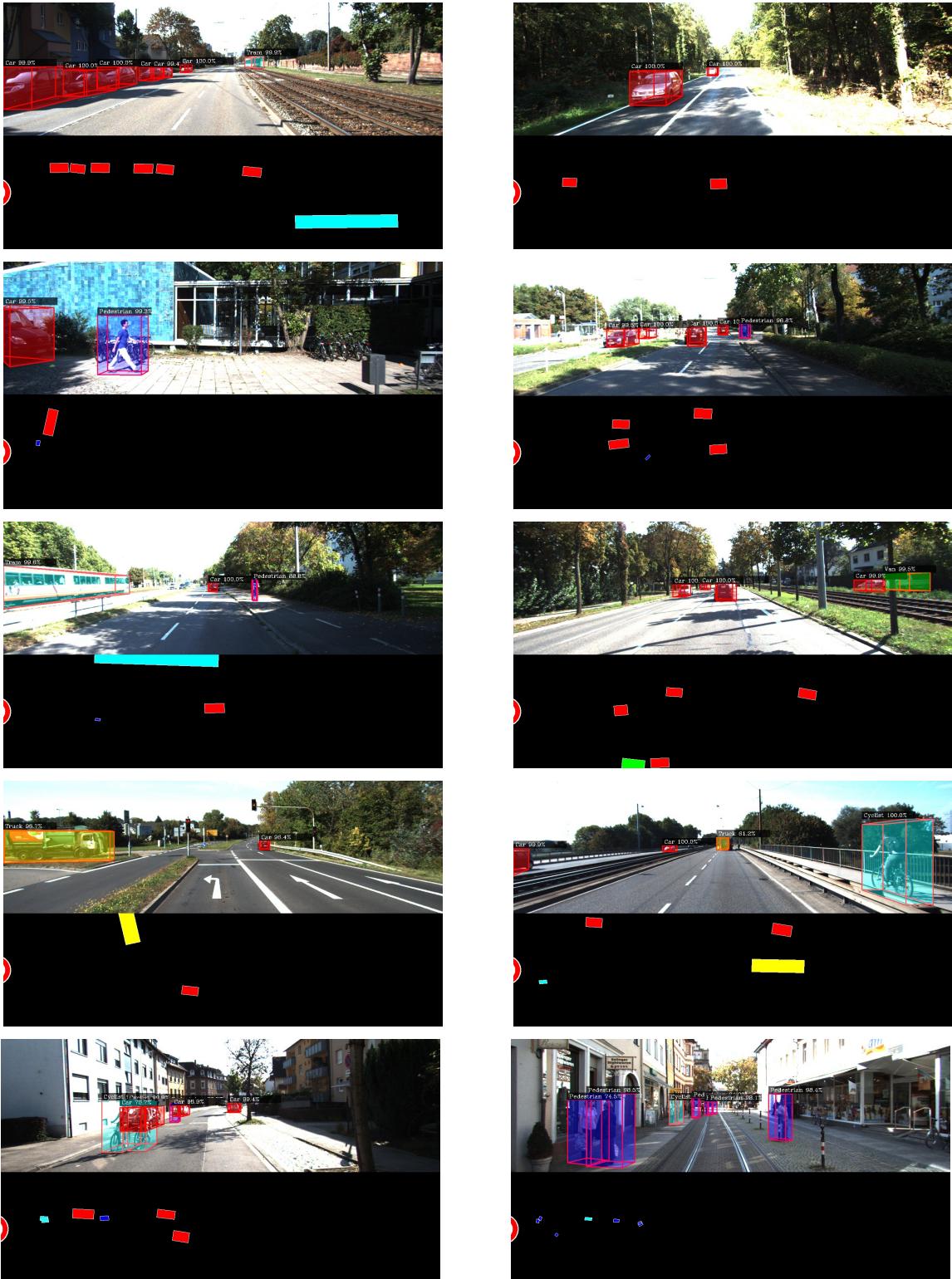


Figure 12. Qualitative results on the *KITTI* dataset. The top part of each image shows a bounding box obtained as a 2D projection of their 3D poses (red: car; yellow: truck, green: van, blue: pedestrian, cyan: tram, cyclist, and others). The bottom part (black background) shows a birds-eye view of the object poses with the ego-vehicle positioned at the center of red circle drawn on the left; pointing towards the right of the image.



Figure 13. Qualitative results on the *nuScenes* dataset across various scenes and lighting conditions. The top part of each image shows a bounding box obtained as a 2D projection of their 3D poses (red: car, yellow: truck, green: motorcycle, blue: pedestrian, cyan: bus, bicycle, and others). The bottom part (black background) shows a birds-eye view of the object poses with the ego-vehicle positioned at the center of red circle drawn on the left; pointing towards the right of the image.

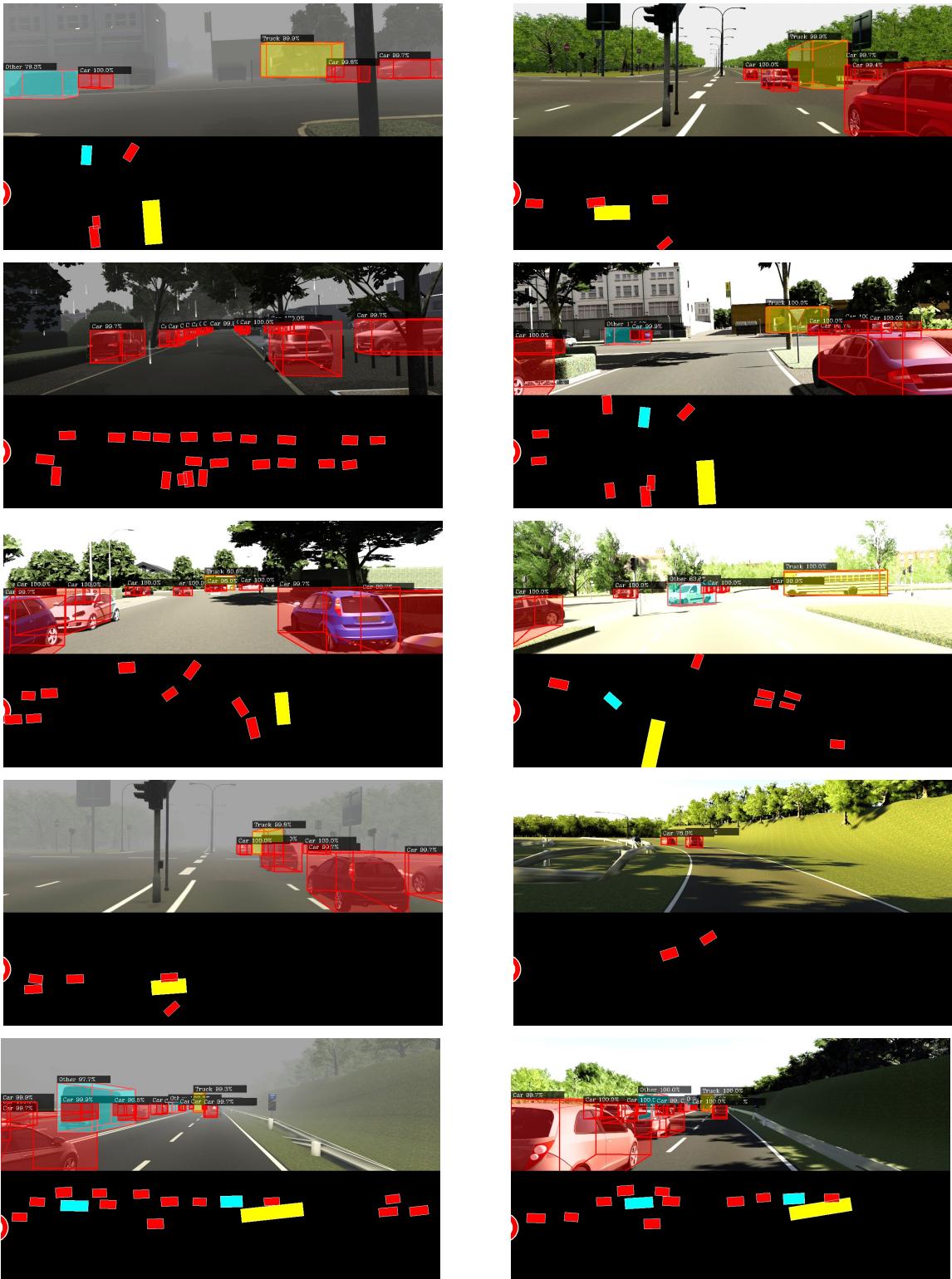


Figure 14. Qualitative results on the *Virtual KITTI 2* dataset across various scenes, weather, and lighting conditions. The top part of each image shows a bounding box obtained as a 2D projection of their 3D poses (red: car, yellow: truck, cyan: van). The bottom part (black background) shows a birds-eye view of the object poses with the ego-vehicle positioned at the center of red circle drawn on the left; pointing towards the right of the image.

References

- [1] Sameer Agarwal, Keir Mierle, and Others. Ceres solver. <http://ceres-solver.org>. 2
- [2] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 2, 6
- [3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giacarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving, 2019. 2, 6
- [4] Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Céline Teuliere, and Thierry Chateau. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2040–2049, 2017. 2
- [5] Michael S Darms, Paul E Rybski, Christopher Baker, and Chris Urmson. Obstacle detection and tracking for the urban challenge. *IEEE Transactions on intelligent transportation systems*, 10(3):475–485, 2009. 2
- [6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2, 6
- [7] Tong He and Stefano Soatto. Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8409–8416, 2019. 2, 7
- [8] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018. 2
- [9] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019. 2
- [10] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155, 2009. 2
- [11] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollr. Microsoft coco: Common objects in context, 2014. 7
- [12] Lijie Liu, Jiwen Lu, Chunjing Xu, Qi Tian, and Jie Zhou. Deep fitting degree scoring network for monocular 3d object detection, 2019. 7
- [13] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image, 2017. 1
- [14] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape, 2018. 7
- [15] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017. 2
- [16] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2, 3
- [17] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8851–8858, 2019. 2
- [18] B Ravi Kiran, Luis Roldao, Benat Irastorza, Renzo Verastegui, Sebastian Suss, Senthil Yogamani, Victor Talpaert, Alexandre Lepoutre, and Guillaume Trehard. Real-time dynamic object detection for autonomous driving using prior 3d-maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2
- [19] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015. 1, 2, 5
- [20] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. *arXiv preprint arXiv:1811.08188*, 2018. 3, 7
- [21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 3
- [22] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud, 2018. 1
- [23] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 1, 2, 3
- [24] Zhixin Wang and Kui Jia. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection, 2019. 1
- [25] B. Xu and Z. Chen. Multi-level fusion based 3d object detection from monocular images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2345–2353, 2018. 7
- [26] Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li. Depth completion from sparse lidar data with depth-normal constraints, 2019. 1
- [27] Bin Yang, Ming Liang, and Raquel Urtasun. Hdnet: Exploiting hd maps for 3d object detection. In *Conference on Robot Learning*, pages 146–155, 2018. 2
- [28] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7652–7660, 2018. 2, 3
- [29] 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. *arXiv preprint arXiv:1906.06310*, 2019. 1, 2, 3

- [30] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Sudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds, 2019. [1](#)
- [31] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018. [1](#), [2](#), [3](#)