MoNet3D: Towards Accurate Monocular 3D Object Localization in Real Time

Xichuan Zhou [1] **Yicong Peng** [1] **Chunqiao Long** [1] **Fengbo Ren** [2] **Cong Shi** [1]

Abstract

Monocular multi-object detection and localization in 3D space has been proven to be a challenging task. The MoNet3D algorithm is a novel and effective framework that can predict the 3D position of each object in a monocular image and draw a 3D bounding box for each object. The MoNet3D method incorporates prior knowledge of the spatial geometric correlation of neighbouring objects into the deep neural network training process to improve the accuracy of 3D object localization. Experiments on the KITTI dataset show that the accuracy for predicting the depth and horizontal coordinates of objects in 3D space can reach 96.25% and 94.74%, respectively. Moreover, the method can realize the real-time image processing at 27.85 FPS, showing promising potential for embedded advanced driving-assistance system applications. Our code is publicly available at https://github.com/CQUlearningsystemgroup/YicongPeng.

1. Introduction

In recent years, computer vision-based automated driving-assistance technology has made great progress. The rapid development of deep learning-based methods has enabled researchers and engineers to develop accurate and cost-effective advanced driving-assistance systems (ADASs), for which object detection and localization is one of the key functions. Various methods based on convolutional neural networks (CNNs) have been proposed for 2D object detection from monocular video images(Girshick et al., 2014; Redmon et al., 2016; Liu et al., 2016). However, despite its great advantages in terms of efficiency and cost, 3D object detection based on monocular vision is still greatly challenging.

Compared with solutions such as LiDAR and stereo vision, the accuracy of the monocular method is far from sufficient for ADAS applications. For example, when using the KITTI 3D object detection benchmark to detect the category of cars, the average accuracy of the state-of-the-art monocular vision algorithm is 63.02% lower than that of LiDAR-based algorithms(Bao et al., 2019; Shi et al., 2020).

Using monocular and single frames of RGB images for 3D object localization and detection can reduce the hardware cost of ADAS applications, but it also brings great technical challenges. First, the images captured by monocular images lack depth-of-field information, and in principle, it is difficult to achieve 3D object localization. Second, different degrees of vehicle occlusion, lack of image information, inelastic distortion caused by rotating the target object, and distortion caused by lens imaging all make monocular 3D object localization more challenging. To meet these challenges, this paper establishes a neural network called MoNet3D by introducing the geometric relationship of neighbouring objects in 3D space to improve the accuracy of 3D object detection and localization.

Specifically, to cope with the 3D localization problem with severely insufficient constraints, some researchers have recently attempted to use prior knowledge to optimize deep learning methods. For example, 3D-Deepbox uses prior knowledge that the predicted 3D bounding box should closely fit the 2D bounding box(Mousavian et al., 2017). Mono3D_PLiDAR relaxed this constraint, assuming that the 2D projection of a 3D object is globally consistent with the bounding box of the 2D object(Weng & Kitani, 2019). These studies show that the geometric relationship between the 2D and 3D bounding boxes associated with detected objects can help to achieve 3D object localization, but their assumption of global consistency might not be met in the face of various types of noise, such as inelastic distortion, and their experimental results show that the research on monocular 3D positioning is still in an early stage.

To address this challenge, we relax the assumption of *global* geometric consistency. Instead, MoNet3D attempts to incor-

[1]Key Laboratory of Dependable Service Computing in Cyber Physical Society Ministry of Education, College of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China 400044. [2]Arizona State University, Tempe, Arizona, United States. Correspondence to: Xichuan Zhou <zxc@cqu.edu.cn>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

porate prior knowledge of the *local* geometric consistency. Intuitively, the proposed method is based on the observation that, given a pair of objects with similar depths, if they are close to each other in the image, they should also be close to each other in actual 3D space. Therefore, the local geometric relations should be helpful for guiding the prediction of 3D object localization. From a methodological point of view, MoNet3D is an end-to-end deep neural network that consists of three stages. The first stage extracts multi-layer features from the image for object detection and localization. The second stage detects 2D objects from monocular images, and the features of the 2D objects are sent to the third stage for 3D object localization. The local consistency of neighbouring objects is formalized as a regularization term to constrain the prediction of 3D localization. By incorporating prior knowledge of local consistency, MoNet3D can improve the accuracy and convergence speed of the deep network training process.

In summary, the main advantages and contributions of MoNet3D are four-fold.

- *Accurate 3D object localization*: By incorporating prior knowledge of the 3D local consistency, MoNet3D can achieve 95.50% accuracy on average for 3D object localization.
- *More accurate 3D object detection*: MoNet3D achieves 3D object detection accuracy of 72.56% in the KITTI dataset (IoU=0.3), which is competitive with state-of-the-art methods.
- *High efficiency*: MoNet3D can process video images at a speed of 27.85 frames per second for 3D object localization and detection, which makes it promising method for embedded ADAS applications.
- *Open source*: Part of the data and code of MoNet3D will be publicly available on the GitHub website when the paper is published.

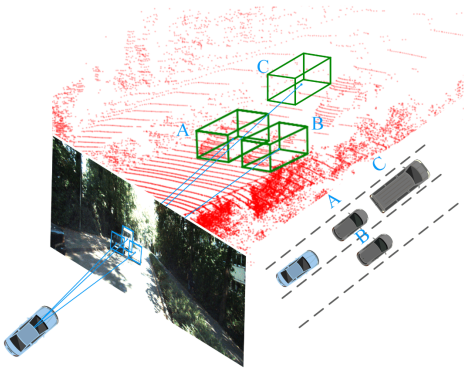

Figure 1. An example applying MoNet3D for 3D object detection using a single RGB image. MoNet3D incorporates the horizontal neighbouring relation between cars A and C in the image, which is important for same-lane determination, to constrain the estimation of 3D localization.

2. Related Work

2.1. 3D Object Detection from LiDAR

Most existing studies of 3D object detection are based on LiDAR sensors (Li et al., 2016). More recently, with the development of deep learning methods, Qi proposed using a deep neural network for 3D object detection with point cloud data (Qi et al., 2017a;b; 2018). Later, Zhou divided the point cloud into 3D voxels and converted the set of points in each voxel into a single feature representation through the voxel-feature coding layer(Zhou & Tuzel, 2018). Chen proposed the MV3D method, which combines vision and LiDAR point cloud information.(Chen et al., 2017b). Although these algorithms achieve state-of-the-art results for 3D object detection, they are rarely applied for ADAS applications due to economic reasons.

2.2. 3D Object Detection for a Single Monocular Image

Instead of installing expensive LiDAR-based systems for 3D object detection, many level-three autonomous cars attempt to use computer vision-based approaches for 3D object detection due to their economic advantages. Very recently, Chen proposed applying deep learning in 3D object detection when using a single camera (Chen et al., 2016). Since then, research on monocular-based 3D object detection has attracted increasing attention(Fang et al., 2019; Zhuo et al., 2018; Crivellaro et al., 2017). For example, Roddick proposed OFT-Net, which maps image-based features onto an orthogonal 3D space for 3D object detection (Roddick et al., 2018); Liu proposed measuring the degree of visual fit between the projected 3D region proposal and the 2D object on the image (Liu et al., 2019); Simonelli proposed using the regression loss to make the training process more stable(sim); Li improved the prediction accuracy of the 3D box method by using the fused features of visible surfaces(Li et al., 2019); Qin used both deep and shallow features extracted by a convolutional neural network to improve the prediction accuracy of the centre point(Qin et al., 2019). These studies of monocular 3D object detection are very inspiring. However, thus far, the results are still below the expectation of industrious application, and the state-of-the-art accuracy on the KITTI dataset is generally less than 50% for the category of cars.

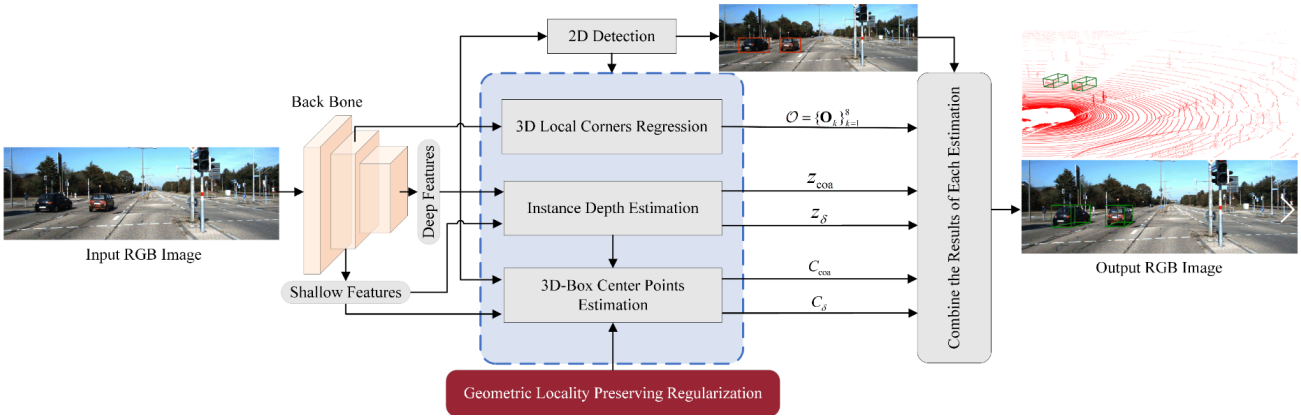

Figure 2. MoNet3D extracts the features from monocular RGB images for 3D object localization. It consists of four modules, including the 2D object detection module, instance depth estimation module, 3D box centre point estimation module, and 3D box corners regression module. Different from previous methods, MoNet3D uses prior knowledge of the geometric locality consistency as a regularization term to constrain the prediction of the centre point of the 3D box.

3. Method

To improve the accuracy of existing monocular-based 3D object detection, we propose the MoNet3D method to use the geometric correlation between neighbouring objects on the image for 3D object localization. Figure 1 briefly illustrates our method. It can be seen for the three cars A, B and C in front of the camera car, that compared to A and B, A and C are closer on the image, which indicates that cars A and C may be closer in the real (3D) world. Based on the observation, we hope to use the horizontal distance relationship of the objects in the picture to constrain the distance of the neighbouring objects in the 3D space, so as to optimize the weight parameters in the training process of the neural network, and then improve the accuracy of 3D object localization. In practice, MoNet3D method may improve the accuracy of lane judgment, which is important for automatic driving application.

3.1. Problem Definition

The MoNet3D method uses a single frame of an RGB image for 3D object detection and localization. Technically, MoNet3D returns the category information, 3D position and size and of the objects of interest in the image in the form of 3D boxes. The 3D box of any object is represented by the 3D centre point $\mathrm{C}_{3\mathrm{d}} = (u^{(3\mathrm{d})}, v^{(3\mathrm{d})}, z^{(3\mathrm{d})})$ and the coordinates of the 8 vertices of the 3D box frame $\mathcal{O} = \{\mathbf{O}_k\}_{k=1}^{8}$.

3.2. Overall Network Structure

As shown in Figure 2, the MoNet3D framework first use VGG-16 without the fully connected layer to extract features from a single frame of an RGB image. Similar to (Qin et al., 2019), we combine the shallow features and deep features for further object detection. The MoNet3D framework divides feature processing into four modules:

- The 2D detection module outputs 2D box and object recognition results based on image features and applies this information to subsequent 3D detection.
- The instance-level depth estimation module estimates the depth information of each object and uses it for subsequent 3D box centre point estimation.
- The 3D box centre point estimation module combines the predicted depth information and 2D box information to estimate the centre point coordinates of each 3D box. MoNet3D incorporates prior knowledge of the geometric locality as regularization for training this module.
- The 3D local corner regression module combines the 2D recognition results and image features to regress the coordinate information of the 8 points of the 3D box frame.

It is particularly worth noting that the main challenge of this paper is the estimation of the 3D box centre point, especially the accuracy of the horizontal offset estimation. Its error determines the error in the lane determination, which plays an important role in the control and safety of autonomous driving. To improve the accuracy of the 3D box centre point estimation, MoNet3D adopts the geometric locality preserving regularization method, which is described in detail below.

3.3. Geometric-Locality-Preserving Regularization

To improve the 3D localization accuracy, mathematically, we formalize the assumption of geometric locality consistency as a regularization term. Suppose there are M objects in the training set. The matrix $\mathbf{S} = \{s_{ij}\}$ defines an $\mathrm{M} \times \mathrm{M}$ similarity matrix as follows:

$$s_{ij} = \exp\left[-(u_i^{(2d)} - u_j^{(2d)})^2\right] / \exp\left[(z_i^{(3d)} - z_j^{(3d)})^2 / \lambda\right] \quad (1)$$

where $u_i^{(2d)}$ and $u_j^{(2d)}$ are the horizontal offsets of object i and object j, respectively, in the 2D image and z_i is the ground-truth depth of object i. MoNet3D assumes that when object i and object j have similar 3D depths, these two objects will have larger similarity s_{ij} if their 2D bounding boxes have smaller horizontal offsets. Otherwise, if these two objects have a large 3D depth difference or their horizontal offset in the image is large, their geometric similarity s_{ij} should be small.

To preserve the geometric similarity for predicting 3D localization, MoNet3D applies the similarity relationship defined in Equation 1 to the fully connected layer of the neural network and optimizes the 3D box centre point estimation (Figure 2). Suppose the output of object i in this layer is $\mathbf{y}_i = \mathbf{W}\mathbf{x}_i + \mathbf{b}$, where $\mathbf{y}_i = (u_i^{(3d)}, z_i^{(3d)})$, $\mathbf{x}_i$ represents the input of the fully connected layer, $\mathbf{W}$ is the connection weight, and $\mathbf{b}$ is the deviation vector. Assuming that the training object i and another object j have large similarity values, MoNet3D attempts to optimize the connection weight $\mathbf{W}$ so that objects i and j are close to each other in 3D space. Technically, MoNet3D minimizes the following regularization term $R(\mathbf{W})$ as

$$\arg\min_{\mathbf{W}} \frac{\beta}{2} \sum_{ij} \|\mathbf{W}\mathbf{x}_i - \mathbf{W}\mathbf{x}_j\|_2^2 \, s_{ij} \quad (2)$$

Intuitively speaking, according to the above equation, if the i and j object pairs are nearby with larger s_{ij} values, then s_{ij} would help to reduce the distance between $\mathbf{W}\mathbf{x}_i$ and $\mathbf{W}\mathbf{x}_j$ in the minimization process so that the similarity of object pairs in 2D space can be maintained in 3D space. For more efficient computation, the regularization term $R(\mathbf{W})$ can be equivalently written as

$$\begin{aligned} R(\mathbf{W}) &= \beta \sum_{ij} \mathrm{tr}\left[\mathbf{W}\,(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\mathbf{W}\right] s_{ij} \\ &= \beta \mathrm{tr}\left[\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{D}\mathbf{X}^{\mathrm{T}}\mathbf{W} - \mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{S}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right] \\ &= \beta \mathrm{tr}\left[\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{P}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right] \end{aligned} \quad (3)$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\mathrm{M}}]$ represents the matrix form of the input vectors of the fully connected layer. $\mathbf{D}$ is the diagonal matrix, where the element on the diagonal is d_{ii}: $d_{ii} = \sum_j s_{ij}$, $\mathbf{S} = \{s_{ij}\}$, $\mathbf{P} = \mathbf{D} - \mathbf{S}$. By applying geometric-locality-preserving regularization, MoNet3D can more accurately predict the 3D box centre point associated with each object.

3.4. Loss Functions

In this section, we briefly summarize the loss function of each of the four modules in the MoNet3D neural network.

3.4.1. 2D Estimation

The MoNet3D method first estimates 2D objects in the image after feature extraction and provides region proposals for subsequent 3D object detection and localization. The 2D estimation module is a basic module that predicts and categorizes regions of interest. Here, we use fast regression from YOLO as the main estimation part and add RoIAlign to 2D estimation to improve the accuracy (Redmon et al., 2016; He et al., 2017). By dividing the original image into 32×32 grids (we use g to indicate a specific grid), we let each grid predict two 2D bounding boxes, $b_{2\mathrm{d}}^{\mathrm{g}} = (u^{(2d)}, v^{(2d)}, d, h)$ and the confidence $\mathrm{Pr}_{\mathrm{obj}}$, where $u^{(2d)}, v^{(2d)}, d, h$ are the coordinates of the centre point of the 2D box and the length and width of the 2D box for each cell grid g. The final 2D box is then predicted by NMS and RoIAlign. The loss function for 2D estimation can be expressed as

$$L_{2\mathrm{d}} = L_{\mathrm{conf}} + \alpha L_{\mathrm{b2d}} \quad (4)$$

where $L_{\mathrm{conf}} = \mathcal{E}_{\mathrm{g}}\left[\mathcal{S}\left(\widehat{\mathrm{Pr}}_{\mathrm{obj}}^{\mathrm{g}}\right), \mathrm{Pr}_{\mathrm{obj}}^{\mathrm{g}}\right]$, $L_{\mathrm{b2d}} = \sum_{\mathrm{g}} {}_{\mathrm{g}}^{\mathrm{obj}} \cdot \mathcal{L}_1\left(\widehat{b}_{2\mathrm{d}}^{\mathrm{g}}, b_{2\mathrm{d}}^{\mathrm{g}},\right)$, $\mathrm{Pr}_{\mathrm{obj}}$ refers to the confidence of the ground truth, $\widehat{\mathrm{Pr}}_{\mathrm{obj}}^{\mathrm{g}}$ refers to the confidence of the predictions, $\mathcal{S}(\cdot)$ is expressed as the softmax function, $\mathcal{E}(\cdot)$ is the cross entropy, $\mathcal{L}_1(\cdot)$ is the L1 distance loss function, α is the balance coefficient, and ${}_{\mathrm{g}}^{\mathrm{obj}}$ indicates whether there is an object in cell g (1 if there is, 0 if there is not).

3.4.2. Instance Depth Estimation

We let z^{g} denote the object depth in an arbitrary grid g. Similar to MonoGRNet, MoNet3D combines deep and shallow features to improve the accuracy of the depth estimation network (Qin et al., 2019). MoNet3D first predicts the rough depth $z_{\mathrm{coa}}^{\mathrm{g}}$ from the deep features and then uses shallow features for fine-tuning. The final instance-level depth can be estimated as $z^{\mathrm{g}} = z_{\mathrm{coa}}^{\mathrm{g}} + z_{\delta}^{\mathrm{g}}$, where z_{δ}^{g} is predicted by the shallow features. The loss function for estimating the depth is formalized as

$$L_{\mathrm{z}} = \gamma L_{\mathrm{zcoa}} + L_{\delta \mathrm{z}} \quad (5)$$

where $L_{\mathrm{zcoa}} = \sum_{\mathrm{g}} {}_{\mathrm{g}}^{\mathrm{obj}} \cdot \mathcal{L}_1\left(\widehat{z}_{\mathrm{coa}}^{\mathrm{g}}, z^{\mathrm{g}}\right)$, $L_{\delta \mathrm{z}} = \sum_{\mathrm{g}} {}_{\mathrm{g}}^{\mathrm{obj}} \cdot \mathcal{L}_1\left(\widehat{z}_{\mathrm{coa}}^{\mathrm{g}} + \widehat{z}_{\delta}^{\mathrm{g}}, z^{\mathrm{g}}\right)$, z^{g} is the depth information of the ground truth, $\widehat{Z}_{coa}^{\mathrm{g}}$ is the depth information of the prediction from the deep features, $\widehat{Z}_{\delta}^{\mathrm{g}}$ refers to the object depth information of the shallow feature prediction, and γ refers to the balance coefficient.

3.4.3. 3D-Box Estimation

This module of 3D box estimation predicts the centre point $C_{3\mathrm{d}} = (u^{(3\mathrm{d})}, v^{(3\mathrm{d})}, z^{(3\mathrm{d})})$ and vertices $\mathcal{O} = \{\mathbf{O}_k\}_{k=1}^{8}$ of the 3D bounding box. To obtain the centre point $C_{3\mathrm{d}}$ of the 3D box, we inversely map the 2D box centre $C_{2\mathrm{d}} = (u^{(2\mathrm{d})}, v^{(2\mathrm{d})})$ through the camera calibration file provided by KITTI to obtain the coarse 3D position $C_{\mathrm{coa}}^{\mathrm{g}}$. The 2D to 3D inverse mapping expression is as follows:

$$\begin{cases} u^{(3\mathrm{d})} = (u^{(2\mathrm{d})} - \theta) * \frac{\widehat{z}^{\mathrm{g}}}{f} \\ v^{(3\mathrm{d})} = (v^{(2\mathrm{d})} - \varphi) * \frac{\widehat{z}^{\mathrm{g}}}{f} \end{cases} \tag{6}$$

where f is the focal length of the camera and θ and φ are the main point parameters of the camera. For the coordinates of the 8 vertices of the 3D box $\mathcal{O} = \{\mathbf{O}_k\}_{k=1}^{8}$, the method we choose is to directly use the deep features for regression. Similar to depth estimation, we also used shallow features to regress the offset C_{δ}^{g} of the 3D box centre points $C_{3\mathrm{d}}^{\mathrm{g}}$. The final $C_{3\mathrm{d}}^{\mathrm{g}}$ can be expressed as $C_{3\mathrm{d}}^{\mathrm{g}} = C_{\mathrm{coa}}^{\mathrm{g}} + C_{\delta}^{\mathrm{g}}$. The loss function for $C_{3\mathrm{d}}^{\mathrm{g}}$ and $\mathcal{O}$ can be expressed as

$$L_{3\mathrm{d}} = \sum_{\mathrm{g}} {}_{\mathrm{g}}^{\mathrm{obj}} \cdot \mathcal{L}_1\left(\widehat{C}_{\mathrm{coa}}^{\mathrm{g}} + \widehat{C}_{\delta}^{\mathrm{g}}, C_{3\mathrm{d}}^{\mathrm{g}}\right) + R(\mathbf{W}) \tag{7}$$

$$L_{\mathcal{O}} = \sum_{\mathrm{g}} \sum_{k} {}_{\mathrm{g}}^{\mathrm{obj}} \cdot \mathcal{L}_1\left(\widehat{\mathbf{O}}_k^{\mathrm{g}}, \mathbf{O}_k^{\mathrm{g}}\right) \tag{8}$$

where $C_{3\mathrm{d}}^{\mathrm{g}}$ is the 3D centre point coordinate of the ground truth, $\widehat{C}_{\mathrm{coa}}^{\mathrm{g}}$ is the 3D centre point coordinate of the prediction from the deep features, $\widehat{\mathbf{O}}_k^{\mathrm{g}}$ is the prediction of $\mathbf{O}_k^{\mathrm{g}}$, and $R(\mathbf{W})$ refers to the regularization term that constrains the adjacent relationship of the object pair in 3D.

4. Experiments

We performed experiments on the KITTI dataset to verify and evaluate the effectiveness of our algorithm. Figure 3 shows the visualization results of MoNet3D on the KITTI dataset. Pictures of three typical test scenarios are shown here, including high-speed roads, town roads, and neighbourhood roads. The pictures in lines 2 to 4 show the comparison between our proposed method and the latest object localization methods (MonoGNet (Qin et al., 2019), M3D (Xu & Chen, 2018), and MonoPSR (Ku et al., 2019)) and the 3D object detection results with the real detection results. In general, the MoNet3D method can effectively identify cars in 3D scenes, although in high-speed road scenes and town road scenes, some vehicle images have incomplete objects. Further observation reveals that M3D and MonoPSR have errors in long-distance object localization. In town road scenes, due to the consideration of the geometric similarity of adjacent objects, the MoNet3D method can better identify distant objects.

4.1. Experiment Setup

Most of the researches on 3D object detection of monocular cameras is verified on the KITTI dataset, so we also carry out experiments on the challenging dataset from KITTI to verify the effectiveness of the MoNet3D algorithm. We used the same method as Chen to split KITTI data sets into 3712 training images and 3769 testing images(Chen et al., 2015). The KITTI dataset contains three types of objects: easy (the bounding box height is greater than 40 pixels, all the objects are visible and truncated by no more than 15%), moderate (the bounding box height is greater than 25 pixels, most objects are visible and truncated by no more than 30%), hard (the bounding box height is greater than 25 pixels, and most of the objects are invisible and not truncated by more than 50%).

Similar to other 3D object estimations, we used the localization and detection accuracy of automotive objects to verify the effectiveness of our method. In terms of object localization, the experiments calculated the relative accuracy of $u^{(3\mathrm{d})}, v^{(3\mathrm{d})}, z^{(3\mathrm{d})}$ as indicators; in terms of 3D object detection, the experiment used the average 3D accuracy rate and bird's-eye view average accuracy as indicators. For the car category, we compared the average accuracy of 3D object detection by the intersection over union (IOU) measure for different object types under two thresholds: 0.5 and 0.7.

We compared the experimental results of the proposed method on the KITTI dataset with state-of-the-art methods. The comparison methods include methods for extracting 3D object regional proposals, such as MF3D, ROI-10D, MonoPSR, and other latest methods for 3D object detection based on neural networks, such as Mono3D, Deep3Dbox, OFT-net, MF3D, ShiftNET, GS3D, and SS3D. We also compared MoNet3D with 3DOP, a 3D object detection method based on a binocular camera.

The experimental hyperparameter settings referred to MonGRNet. We initialized the model with random parameters. In the experiment, the similarity hyperparameter of Equation 1 was set to 100.00, and α, β and γ were all set to 10.00. Model training uses tensorflow's SGD algorithm with momentum, batchsize is set to 2 and learning rate is set to 10^{-5}. A total of 800000 iterations were trained on the KITTI dataset. Numerical experiments were performed on a computer equipped with an InterCorei7-6900K CPU, 32GB of memory, and an NVIDIA GeForce GTX 1080 Ti

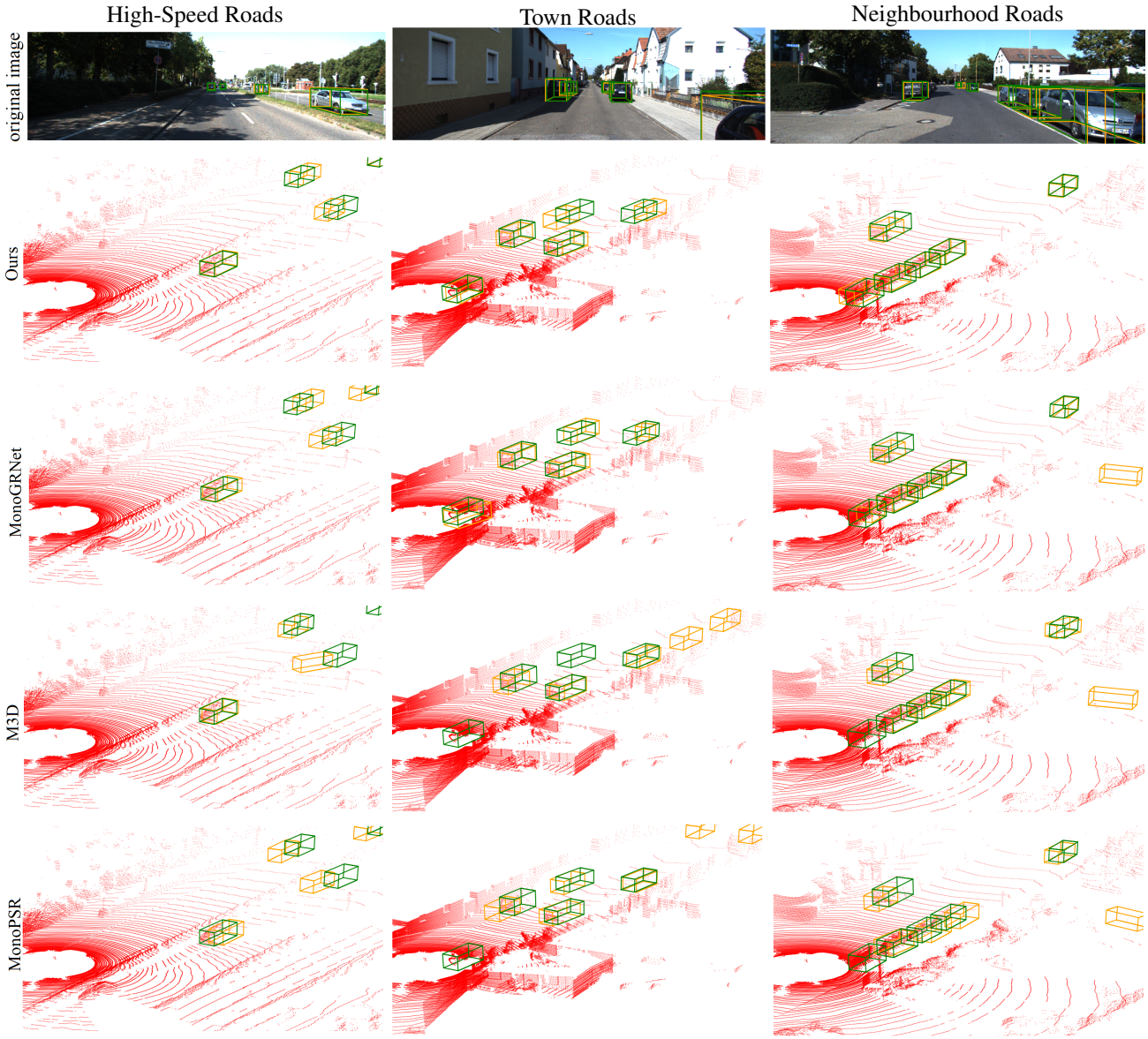

Figure 3. The 3D detection results of MoNet3D for the different scenes in the KITTI benchmark dataset. For all the pictures, the green 3D boxes are the ground truth, the orange 3D boxes are the predictions, and the camera centres are in the bottom-left corner.

graphics card.

4.2. Result of 3D Object Localization and Detection

In terms of object localization, we calculated the accuracy of the horizontal and height predictions estimated by the MoNet3D method and performed a comparison with the classic M3D. The experimental results showed that in the horizontal estimation direction ($u^{(3d)}$), M3D achieved a 90.59% accuracy. Overall, in these three directions, our proposed method achieved an average accuracy of 96.07%, where $u^{(3d)}$ is 94.74%, $v^{(3d)}$ is 97.21%, and $z^{(3d)}$ is 96.25%. The experiments showed that because the horizontal regular optimization method was used, the proposed method was better than the recently proposed M3D in terms of the positioning accuracy of the horizontal direction.

Considering that most of the research on image depth estimation now was pixel-level depth estimation, we compared the instance-level depth estimation we invented with them. According to the latest research on pixel-level depth estimation(Fu et al., 2018; Ren et al., 2019; Liebel & Körner, 2019), for example, the relative absolute error of the depth estimation of DORN on the KITTI dataset was 8.78%(Fu et al., 2018), and the relative absolute error of the depth estimation of MultiDepth on the same dataset was 13.82%(Liebel & Körner, 2019). Compared with pixel-

Table 1. **3D Detection**: Comparisons with the state-of-the-art methods in terms of the average precision for 3D object detection for the car category in the KITTI validation dataset with different IoUs

Method	AP3D (**IoU=0.5**)			AP3D (**IoU=0.7**)			FPS[a]
	Easy	Moderate	Hard	Easy	Moderate	Hard	
3DOP(Chen et al., 2017a)	46.04	34.63	30.09	6.55	5.07	4.10	0.23
Mono3D(Chen et al., 2016)	25.19	18.20	15.22	2.53	2.31	2.31	0.33
OFT-Net(Roddick et al., 2018)	-	-	-	4.07	3.27	3.29	-
FQNet(Liu et al., 2019)	28.16	21.02	19.91	5.98	5.50	4.75	2.00
ROI-10D(Manhardt et al., 2019)	-	-	-	10.25	6.39	6.18	-
MF3D(Novak, 2017)	47.88	29.48	26.44	10.53	5.69	5.39	8.33
MonoDIS(sim)	-	-	-	11.06	7.60	6.37	-
MonoPSR(Ku et al., 2019)	49.65	41.71	29.95	12.75	11.48	8.59	5.00
ShiftNet(nai)	-	-	-	13.84	11.29	11.08	-
GS3D(Li et al., 2019)	30.60	26.40	22.89	11.63	10.51	10.51	0.43
SS3D(Jörgensen et al., 2019)	-	-	-	14.52	13.15	11.85	20.00
M3D-RPN(Brazil & Liu, 2019)	-	-	-	20.27	17.06	15.21	-
Ours	**55.64**±0.45	**34.10**±0.14	**34.10**±0.07	**22.73**±0.30	**16.73**±0.27	**15.55**±0.24	**27.85**

[a] FPS means frames per second and the FPS here refers to the FPS running on the computer.

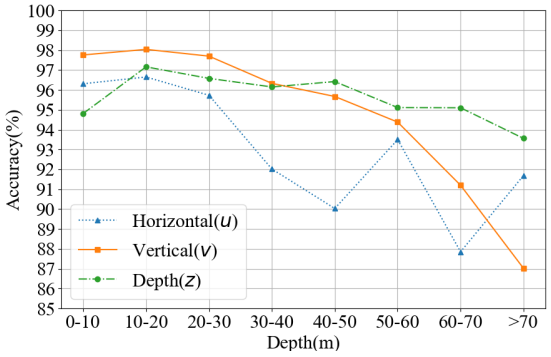

Figure 4. Relative accuracy of the 3D box centre coordinates at different depths. The depth is divided into 5 groups of 10 metres, and the accuracy of $u^{(3d)}, v^{(3d)}, z^{(3d)}$ is calculated at different depths, where the blue line is the relative accuracy of the 3D box centre horizontal coordinate $u^{(3d)}$, the orange line is the relative accuracy of the 3D box centre vertical coordinate $v^{(3d)}$, and the green line is the relative accuracy of the 3D box centre depth coordinate $z^{(3d)}$.

level depth estimation method, the MoNet3D was coarser instance-level estimation method, and the depth estimation average accuracy was 96.25%, which was significantly higher than the pixel-level depth estimation method.

To explore the effect of depth on the localization results, we grouped the test samples into groups of 10 metres (since the maximum distance of the car object in the KITTI dataset is 83 metres, and there are few objects with a depth of 80 metres or more, so we grouped 70-80 metres and 80-90 metres into one group) and evaluate the average accuracy of the MoNet3D method in $u^{(3d)}, v^{(3d)}, z^{(3d)}$. As shown in Figure 4, the average accuracy of the 3D box centre in the three directions is the largest when the depth is between 10 and 20 metres, and the accuracy in all directions decreases as the depth increases. However, even for objects 40 metres away, our proposed method still had a relative accuracy of more than 90% in 3D object localization.

In addition to object localization, our proposed method can also perform 3D object detection, which is a very challenging task. Existing monocular 3D object detection methods have not achieved the accuracy of target recognition (see Table 1), but our research found that MoNet3D can still achieve better 3D recognition results under close-range conditions. When the IoU threshold was set to 0.3 and the depth was 0 to 10 metres and 10 to 20 metres, the accuracy of the proposed method in 3D object detection was 75.40%-80.99%. However, as the depth increased, as there was no other information, it was very difficult to predict the depth using only pictures whose depth information was severely compressed, and the prediction error also increased sharply. This experiment showed that our proposed method achieved good results in 3D object detection, but the current monocular 3D object detection method can only be applied to low-power ADASs (advanced driving-assistance systems) and other low-power embedded systems.

5. Comparison with the State-of-the-Art Methods

To compare with other methods, we compared MoNet3D with recent monocular 3D object detection methods based on the KITTI dataset. The evaluation results are shown in

Table 2. **Bird's-Eye-View 3D Detection**: Comparisons with the state-of-the-art methods in terms of the 3D BEV(Bird's-Eye-View) and the inference time per image for the KITTI validation dataset with different IoUs.

Method	APBEV (**IoU=0.5**)			APBEV (**IoU=0.7**)			FPS[a]
	Easy	Moderate	Hard	Easy	Moderate	Hard	
Mono3D(Chen et al., 2016)	30.50	22.39	19.16	5.22	5.19	4.13	0.33
FQNet(Liu et al., 2019)	32.57	24.60	21.25	9.50	8.02	7.71	-
OFT-Net(Roddick et al., 2018)	-	-	-	11.06	8.79	8.91	2.00
3DOP(Chen et al., 2017a)	55.04	41.25	34.55	12.63	9.49	7.59	0.23
ROI-10D(Manhardt et al., 2019)	46.9	34.1	30.5	14.50	9.9	8.7	5.00
ShiftNet(nai)	-	-	-	18.61	14.71	13.57	-
MonoPSR(Ku et al., 2019)	56.97	43.39	36.00	20.63	18.67	14.45	5.00
MF3D(Novak, 2017)	-	-	-	22.03	13.63	11.60	-
Ours	**59.15**±0.20	**43.26**±0.11	**36.00**±0.06	**27.48**±1.14	**21.80**±0.29	**17.86**±0.26	**27.85**

[a] FPS means frames per second and the FPS here refers to the FPS running on the computer.

Table 1. Overall, our proposed method reached the state-of-the-art level. It is clear from Table 1 that our method significantly outperforms other methods in 3D object detection. When IoU = 0.3, the highest accuracy of our proposed method reached 72.56%.

In addition to 3D object detection, we also evaluate the accuracy of our method in bird's-eye view (BEV) with IoU thresholds of 0.7 and 0.5 and compare it with recent monocular 3D object detection methods. The evaluation results are shown in Table 2. In general, the proposed method also surpasses other new methods recently proposed for BEV. Specifically, when IoU = 0.7, in easy mode, compared with other methods, the leading range is from 5.45% to 22.26%.

Embedded devices are often used in the field of autonomous driving, so there are very high requirements for energy efficiency. Due to the use of the highly efficient neural network, real-time image processing speed can be achieved. Compared with other methods, our regularization method does not bring speed loss with computational efficiency, which provides conditions for the application of embedded devices.

6. System Performance for Embedded ADAS Applications

To better explore the performance indicators of our method on embedded devices, such as the accuracy, real-time performance, and power, we apply the proposed method to an embedded NVIDIA Jetson AGX Xavier system. NVIDIA Jetson AGX Xavier mainly includes an 8-core NVIDIA Carmel ARMv8.2 64-bit CPU, a 512-core Volta architecture GPU consisting of 8 stream multiprocessors, 16 GB of memory, and an FP16 (computing power) of 11 TFLOPS (tera floating-point operations per second). Compared with the computer platform, Jetson AGX Xavier's storage capacity, computing power, and power consumption are far inferior, but our proposed method can be implemented in Jetson AGX Xavier with the same accuracy. The performance is over 5 frames per second(FPS), and the power at this time is 26.13 W.

7. Discussion and Conclusion

In this paper, we propose a novel monocular 3D object detection network. Using the proposed regularization term to optimize the corresponding loss function greatly enhances the ability of the original network in 3D object detection and pose estimation and improves the corresponding accuracy. It has excellent performance in 3D object detection, localization and attitude estimation. Moreover, the proposed method also has good migration ability, which can be applied to other networks to improve the accuracy of object detection of corresponding networks. In terms of the real-time performance, our method reaches 27.85 FPS, which is a fairly good real-time performance.

Of course, MoNet3D still has many limitations. First, object detection and localization is difficult using a monocular camera, so MoNet3D still has a series of limitations like other monocular camera object detection methods. Compared with methods based on radar and binocular cameras, 3D object detection methods using monocular cameras have lower accuracies. Current research finds that although 3D object detection methods for monocular cameras have the advantage of low cost, they are only suitable for object localization and 3D object detection at short distances at low speeds. In high-speed scenarios, LiDAR must be combined to increase the accuracy to an acceptable level. Second, the current method is the same as ROI-1D, ShiftNet, MonoPSR, SS3D and other methods, and its accuracy is affected by 2D object detection. In the future, research on new methods of 2D object detection will improve the accuracy of 2D detection; on the other hand, using the information of 3D object detection to improve the accuracy of 2D object detection in turn will be considered.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Contract 61971072.

References

Bao, W., Xu, B., and Chen, Z. Monofenet: Monocular 3d object detection with feature enhancement networks. *IEEE Transactions on Image Processing*, 2019.

Brazil, G. and Liu, X. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9287–9296, 2019.

Chen, X., Kundu, K., Zhu, Y., Berneshawi, A. G., Ma, H., Fidler, S., and Urtasun, R. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems*, pp. 424–432, 2015.

Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., and Urtasun, R. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2147–2156, 2016.

Chen, X., Kundu, K., Zhu, Y., Ma, H., Fidler, S., and Urtasun, R. 3d object proposals using stereo imagery for accurate object class detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1259–1272, 2017a.

Chen, X., Ma, H., Wan, J., Li, B., and Xia, T. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1907–1915, 2017b.

Crivellaro, A., Rad, M., Verdie, Y., Yi, K. M., Fua, P., and Lepetit, V. Robust 3d object tracking from monocular images using stable parts. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1465–1479, 2017.

Fang, J., Zhou, L., and Liu, G. 3d bounding box estimation for autonomous vehicles by cascaded geometric constraints and depurated 2d detections using 3d results. *arXiv preprint arXiv:1909.01867*, 2019.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., and Tao, D. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2002–2011, 2018.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Jörgensen, E., Zach, C., and Kahl, F. Monocular 3d object detection and box fitting trained end-to-end using intersection-over-union loss. *arXiv preprint arXiv:1906.08070*, 2019.

Ku, J., Pon, A. D., and Waslander, S. L. Monocular 3d object detection leveraging accurate proposals and shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11867–11876, 2019.

Li, B., Zhang, T., and Xia, T. Vehicle detection from 3d lidar using fully convolutional network. *arXiv preprint arXiv:1608.07916*, 2016.

Li, B., Ouyang, W., Sheng, L., Zeng, X., and Wang, X. Gs3d: An efficient 3d object detection framework for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1019–1028, 2019.

Liebel, L. and Körner, M. Multidepth: Single-image depth estimation via multi-task regression and classification. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 1440–1447. IEEE, 2019.

Liu, L., Lu, J., Xu, C., Tian, Q., and Zhou, J. Deep fitting degree scoring network for monocular 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1057–1066, 2019.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. Ssd: Single shot multibox detector. In *European conference on computer vision*, pp. 21–37. Springer, 2016.

Manhardt, F., Kehl, W., and Gaidon, A. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2069–2078, 2019.

Mousavian, A., Anguelov, D., Flynn, J., and Kosecka, J. 3d bounding box estimation using deep learning and geometry. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Novak, L. *Vehicle detection and pose estimation for autonomous driving*. PhD thesis, Master's thesis, Czech Technical University in Prague, 2017. Cited on, 2017.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 652–660, 2017a.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pp. 5099–5108, 2017b.

Qi, C. R., Liu, W., Wu, C., Su, H., and Guibas, L. J. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 918–927, 2018.

Qin, Z., Wang, J., and Lu, Y. Monogrnet: A geometric reasoning network for monocular 3d object localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8851–8858, 2019.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Ren, H., El-Khamy, M., and Lee, J. Deep robust single image depth estimation neural network using scene understanding. In *CVPR Workshops*, pp. 37–45, 2019.

Roddick, T., Kendall, A., and Cipolla, R. Orthographic feature transform for monocular 3d object detection. *arXiv preprint arXiv:1811.08188*, 2018.

Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., and Li, H. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10529–10538, 2020.

Weng, X. and Kitani, K. Monocular 3d object detection with pseudo-lidar point cloud. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

Xu, B. and Chen, Z. Multi-level fusion based 3d object detection from monocular images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2345–2353, 2018.

Zhou, Y. and Tuzel, O. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4490–4499, 2018.

Zhuo, W., Salzmann, M., He, X., and Liu, M. 3d box proposals from a single monocular image of an indoor scene. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.