SM3D: SIMULTANEOUS MONOCULAR MAPPING AND 3D DETECTION

Runfa Li, Truong Nguyen

Department of Electrical and Computer Engineering, University of California, San Diego

ABSTRACT

Mapping and 3D detection are two major issues in vision-based robotics, and self-driving. While previous works only focus on each task separately, we present an innovative and efficient multi-task deep learning framework (SM3D) for Simultaneous Mapping and 3D Detection by bridging the gap with robust depth estimation and "Pseudo-Lidar" point cloud for the first time. The Mapping module takes consecutive monocular frames to generate depth and pose estimation. In 3D Detection module, the depth estimation is projected into 3D space to generate "Pseudo-Lidar" point cloud, where Lidar-based 3D detector can be leveraged on point cloud for vehicular 3D detection and localization. By end-to-end training of both modules, the proposed mapping and 3D detection method outperforms the state-of-the-art baseline by 10.0% and 13.2% in accuracy, respectively. While achieving better accuracy, our monocular multi-task SM3D is more than 2 times faster than the state of the art pure stereo 3D detector, and 18.3% faster than using two modules separately.

***Index Terms*—** SM3D, Monocular Mapping, Monocular 3D detection, Pseudo-Lidar, Depth Estimation

1. INTRODUCTION

Traditional mapping and visual odometry strategies are mostly based on SLAM [1, 2] (Simultaneous Localization and Mapping) algorithm, which is well-known for simultaneously perceiving surrounding environments and keeping track of the ego motion. However, traditional SLAM requires ubiquitous sensors, expensive depth cameras which are not only high-cost but also computationally expensive. SFM (Structure from motion) is a good alternative to SLAM, using only consecutive image snippets. Using an elegant self-supervised learning style, SFM Learner[3] jointly trains the depth model and pose model with photo-consistency loss between target and warped images. Monodepth2[4] introduces SSIM[5] in SFM to enforce photometric consistency, filtering out lighting changes and imaging noises. Similarly, [6] adds perceptual loss by calculating the pixel-wise CNN features. Since learning of mapping by scene flow assumes that the background is completely rigid, many supervised methods design segmentation mask to eliminate dynamic objects [7, 8], while unsupervised methods rely on "soft mask" by adding optical self-supervision in the training process. DF-Net[9] uses a pretrained optical flow network to segment non-rigid objects, while GeoNet[10] employs a subsequent module to compensate for the final predicted motion estimation.

Existing 3D detection algorithms mostly use 2D-3D prototypes [11, 12, 13], which are based on 2D object detection, where different geometric constraints are imposed to project 2D proposals to 3D. Although these approaches give reasonable 3D proposals, they lack in producing accurate 3D location. Alternatively, Lidar-based methods [14, 15, 16] are far more accurate than state-of-the-art 2D-3D prototypes. Similar to mapping, latest detection works try to eliminate expensive Lidar, Radar, and depth camera, by using only images. The Lidar-based detection prototype could be directly applied to images by depth estimation, which saves the cost on device and maintains high accuracy. Based on the experimental results, we inherit the Lidar-based style and use monocular-based depth for end-to-end training which further improve the performances over state-of-the-art monocular-based models while achieve much higher efficiency compared to stereo-based models [17, 18].

Since the previous works estimate the ego-map and 3D detection separately, they ignore the potential to efficiently integrate a multi-task model. Our key contributions can be briefly summarized as follows:

- We propose a multi-task framework that only takes monocular inputs to estimate ego-map and 3D detection simultaneously by bridging the gap with robust depth estimation and Pseudo-Lidar generation. The proposed framework is validated to be much more efficient than using stereo inputs or dealing with tasks separately.
- We derive and impose a novel pose-consistency constraint in the end-to-end self-supervised training for Mapping Module while keeping the efficiency in testing, which boosts the mapping performance by 10.0%.
- We successfully end-to-end train a 3D detector only using monocular images. The resulting 3D Detection module outperforms the state-of-the-art monocular 3D detectors by 13.2%.

This project is supported in part by the SRIP program at ECE Dept., UCSD.

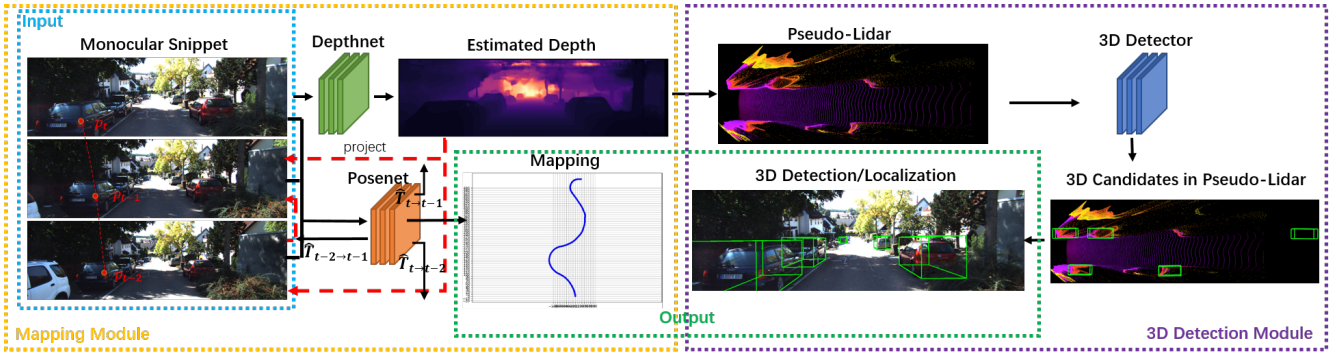


Fig. 1: Overview of our SM3D. Mapping Module: Jointly learning and estimating depth and pose. 3D Detection Module: Jointly learning and estimating depth and 3D detection. Input: Monocular snippet. Output: Mapping/3D Detection & Location.

2. SM3D FRAMEWORK

In this section, we present the proposed SM3D network with Mapping and 3D detection modules (See Fig. 1). The input is real-time monocular snippet of consecutive frames, the outputs are real-time ego-map to the current frame and 3D object detection/localization of the current frame.

2.1. Mapping Module

We design our Mapping Module following the success of the concurrent SFM works, which use a similar prototype based on the baseline SFM Learner[3]. This module takes a snippet of consecutive monocular frames, jointly trains a depth and pose network in a self-supervised manner.

We use the following transformation to project a pixel from the target view p_t to the source view p_s:

$$p_s = K\hat{T}_{t\to s}\hat{D}_t(p_t)K^{-1}p_t \tag{1}$$

where K is the camera intrinsics, $\hat{T}_{t\to s}$ is the pose estimation from target view to source views, $\hat{D}_t(p_t)$ is the estimated depth of pixel p_t from depth network. Bilinear interpolation is then used to populate discrete value for coordinates of the projected pixels. The reconstructed target view can be produced with the inverse transformation, then the photometric consistency loss can be defined as:

$$\mathcal{L}_{vs} = \sum_s \sum_p |I_t(p) - \hat{I}_s(p)| \tag{2}$$

where p is the index of the pixel coordinates, I_t is the target view and I_s are the source views. $\hat{I}_s$ is the reconstructed target view inverse-transformed from the source view. We improve the SFM model in two aspects.

First, to the best of our knowledge, previous SFM works use disparity network. However, if the disparity network is trained to estimate depth, its intrinsic error will be exacerbated for far-away objects [19]. Under such concern, our SM3D initializes with a depth network, which leads to better performance.

Second, to improve the long-term robustness of pose estimation, many previous works, especially SLAM algorithms, randomly choose key frames in longer time range for training. However such an approach ignores the inner connection of a snippet in a given length because some frames are skipped. To take advantage of the inner connection between all frames for the training, we derive a novel pose-consistency constraint on top of the total loss. Different to the key-frame strategy, our image snippet is similar to a "sliding window" with stride one to slide over all frames, and there is no skipping frame. When the snippet length is 3, where frame T is the target view, and frames $T-1, T-2$ as the source views, the proposed skip-time pose consistency constraint can be described as:

$$\hat{T}_{t-1\to t}\hat{T}_{t-2\to t-1} = \hat{T}_{t-2\to t} \tag{3}$$

Here $\hat{T}_{m\to n}$ is the estimated pose from frame m to n. When using longer snippet, it will be extended between all frames, which maximally utilize the inner connection of pose consistency for training.

2.2. 3D Detection Module

3D Detection Module is designed in a Pseudo-Lidar approach that jointly trains depthnet and 3D detector, in an end-to-end approach. We generate Pseudo-Lidar from depth estimation by projecting each 2D pixel to 3D space. Given 2D coordinate of each pixel (u, v) in the depth map, the projection process can be derived as:

$$z = D(u,v),\ x = \frac{(u-c_U)\times z}{f_U},\ y = \frac{(v-c_V)\times z}{f_V} \tag{4}$$

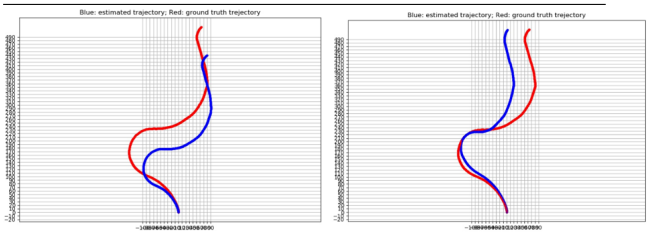
where z, x, y are the depth, width and height of the corresponding projected point in 3D coordinate. (c_U, c_V) is the pixel location w.r.t the camera center and f_V is the vertical focal length of the camera.

Since the Pseudo-Lidar generation process is clearly differentiable w.r.t z in Eq. (4), the end-to-end training is applicable by back-propagating the loss from 3D detector all the way back to the monocular depthnet. Similar to [20], the jointly learning loss is defined as:

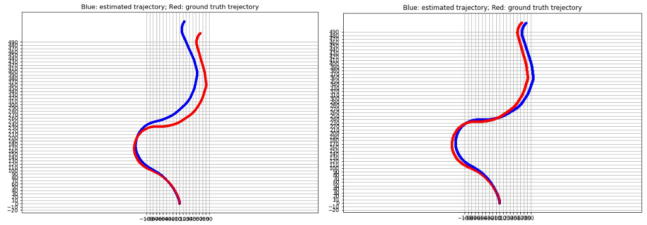
$$\mathcal{L} = \lambda_{det}\mathcal{L}_{det} + \lambda_{depth}\mathcal{L}_{depth} \tag{5}$$

Table 1: Absolute Trajectory Error (ATE) on the KITTI odometry test split averaged over all 3-frame snippets, all models are trained with snippet length 3 on our subset of KITTI odometry training split.

Method	Seq. 09	Seq. 10
SFM Learner	0.0100 ± 0.0063	0.0085 ± 0.0073
SM3D (ours)	$\mathbf{0.0090 \pm 0.0052}$	$\mathbf{0.0084 \pm 0.0067}$



(a) SuperPoint (b) Libviso2



(c) SFM Learner (d) SM3D (Ours)

Fig. 2: Mapping results on the first 700 frames of sequence 09, KITTI test split. The ground truth trajectory is in red, the estimated trajectory is in blue. The estimated trajectory is recovered from the estimated pose.

where $\mathcal{L}_{det}$ and $\mathcal{L}_{depth}$ are the loss of 3D detection and depth estimation, λ_{det} and λ_{depth} are corresponding weights. $\mathcal{L}_{det}$ is a combination of the classification loss for candidate category and the regression loss for the bounding box location. $\mathcal{L}_{depth}$ is the L1 loss of the estimated depth and the ground truth.

2.3. SM3D Model

Finally, we combine the trained Mapping Module and 3D Detection Module together to build our SM3D network. For the utility purpose, since the posenet from Mapping Module can be used independently of the jointly trained depthnet, we use the depthnet jointly trained with 3D detector for SM3D.

3. EXPERIMENTAL RESULTS

3.1. Setup

We use KITTI dataset to evaluate the algorithm performance. On the Mapping Module, for simplicity, we use a snippet length of 3, and train all models on a subset of KITTI odometry training split. Similar to previous works, we evaluate on sequence 09 and 10 of KITTI odometry test split. For the 3D Detection Module, we train and evaluate with KITTI 3D detection benchmark, with 3712, 3769, 7518 images for training, validation and testing, respectively. We use two Nvidia GTX 1080Ti GPU for training. For end-to-end training of the Mapping Module, we initialize the depthnet from a pretrained BTS network [21], while the posenet is jointly trained from scratch using same network as SFM Learner. The image size is 128×416, the learning rate is set to 2×10^{-4} with batchsize 4, the parameters α and β of Adam optimizer are set to 0.9 and 0.999. For end-to-end training of the 3D Detection Module, we first initialize from pretrained BTS and choose PointRCNN as our 3D detector. Similar to [20], we first fix depthnet to train the 3D detector from scratch. Finally, we jointly train the detector and fine-tune the pretrained BTS depth network, the depth ground truth is produced by projecting Lidar ground truth. The depth ground truth is projected from Lidar data in KITTI. The image size is 352×1216, For testing, we use a single Nvidia GTX 1080 Ti GPU.

3.2. Qualitative Results

Mapping Module. Based on the Absolute Trajectory Error (ATE) reported in Table 1, the proposed SM3D network is 10.0% and 1.2% better than the baseline SFM learner on sequence 09 and 10, respectively. Recovered from the estimated pose, we visualize the mapping trajectory as shown in Fig 2. SuperPoint[22], Libviso2[23], and SFM Learner[3] are chosen for comparison, where all models are trained on our data split. As observed, the proposed SM3D network has the closest trajectory to the ground truth, which outperforms all methods above. Such results further validate our design of the skip-time photometric consistency constraint and the utility of depth rather than disparity. For simplicity we set the snippet length to be 3, the impact of the snippet length will be furthered studied in the future.

3D Detection Module. We report the 3D detection results of car category. As shown in Table 2, we compare our results to state-of-the-art models with monocular frames and stereo pairs input from KITTI, for the average precision (AP) of both bird's eye view (BEV) and 3D with the threshold IoU at 0.5 and 0.7, respectively. First, it is obvious that the later Pseudo-Lidar methods outperform all other "2D-3D" approaches. Inheriting such prototype, the proposed SM3D network outperforms all Pseudo-Lidar strong baselines on the average precision (AP) of both BEV and 3D for IoU threshold of 0.7, implies that SM3D performs well for challenging cases. Compared to state-of-the-art strong baseline Mono3DPLIDAR [27], our SM3D is 13.2% and 11.8% better on AP_{BEV} and AP_{3D}, respectively. More importantly, our one-stream 3D detection module is more efficient compared to the two-stream structure of Mono3DPLIDAR[27], which has an additional 2D instance segmentation subbranch and network. We claim that there is no need to sacrifice the model efficiency by using "2D-3D" joint supervision training approach. Computationally expensive strategies such as instance or semantic segmentation is not needed as the one-

Table 2: Qualitative detection results comparison on KITTI val set. We report the average precision (in %) of car category on bird's eye view and 3D object detection as AP_{BEV} and AP_{3D}. Top two rows are state-of-the-art "2D-3D" methods, middle three rows are concurrent 3D methods, bottom two rows in blue are state-of-the-art concurrent Pseudo-Lidar based 3D methods. The proposed SM3D (in red) outperforms all monocular methods at IoU=0.7. AP at lower IoU threshold 0.5 only reflects the 3D box proposal/candidates, while the higher IoU threshold at 0.7 reflects the precision of the box coordinates, thus the higher AP at IoU=0.7 validates that our SM3D is better for 3D localization, not just for 3D proposal.

Method	Input	AP_{BEV}/AP_{3D} (in %), IoU = 0.5			AP_{BEV}/AP_{3D} (in %), IoU = 0.7		
		Easy	Moderate	Hard	Easy	Moderate	Hard
Mono3D[11]	Monocular	30.5/25.2	22.4/18.2	19.2/15.5	5.2/2.5	5.2/2.3	4.1/2.3
Deep3DBox[12]	Monocular	30.0/27.0	23.8/20.6	18.8/15.9	10.0/5.6	7.7/4.1	5.3/3.8
MLF-MONO[24]	Monocular	55.0/47.9	36.7/29.5	31.3/26.4	22.0/10.5	13.6/5.7	11.6/5.4
ROI-10D[25]	Monocular	46.9/37.6	34.1/25.1	30.5/21.8	14.5/9.6	9.9/6.6	8.7/6.3
MonoGRNet[26]	Monocular	-/50.5	-/37.0	-/30.8	-/13.9	-/10.2	-/7.6
PL-MONO-FP[18]	Monocular	70.8/66.3	49.4/42.3	42.7/38.5	40.6/28.2	26.3/18.5	22.9/16.4
Mono3DPLiDAR[27]	Monocular	72.1/68.4	53.1/48.3	44.6/43.0	41.9/31.5	28.3/21.0	24.5/17.5
Naive SM3D (Ours)	Monocular	70.7/45.1	52.5/31.2	47.3/22.6	38.0/5.4	29.6/4.6	26.2/4.5
SM3D (Ours)	Monocular	68.1/65.9	49.1/47.2	41.3/40.0	**45.8/33.1**	**32.8/23.9**	**28.0/20.4**
PL-STEREO-FP[18]	Stereo	89.8/89.5	77.6/75.5	68.2/66.3	72.8/59.4	51.8/39.8	44.0/33.5

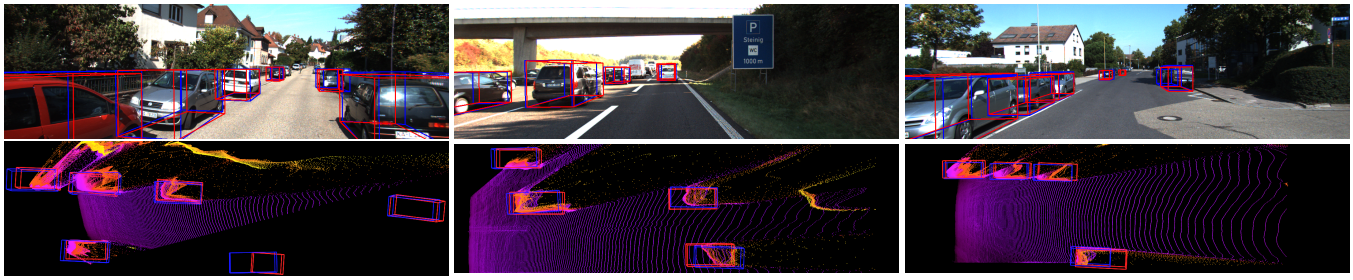


Fig. 3: Qualitative results of the proposed SM3D network on KITTI val set. We visualize our 3D bounding box estimate (in blue) and ground truth (in red) on the frontal images (1st row) and Pseudo-LiDAR point cloud (2nd row).

Table 3: Inference time of each module and model.

Module	Mapping	Detection	SM3D	PL-STEREO[18]
time (ms)	82	267	295	602

stream end-to-end training strategy achieves equal and better accuracy while keeping an efficient structure design.

In our ablation study, we compare AP of end-to-end trained model (SM3D in Table 2), to non-end-to-end trained model (Naive SM3D in Table 2). For end-to-end training, SM3D is 12.7% better on AP_{BEV}. For AP_{3D}, once being trained end-to-end, we achieve a large improvement of 27.7%, which further validates the significance of our end-to-end training strategy. From Table 2, it is clear that large AP performance gap exists between monocular and stereo pairs input. However, compared to other state-of-the-art monocular models, the proposed SM3D further narrows the performance gap to the state-of-the-art stereo model PL-STEREO-FP[18]. More importantly, from Table 3, our Detection Module is 2.3 times faster than PL-STEREO-FP (295 ms/frame vs 602 ms/frame).

SM3D. We efficiently integrate the two modules to build the proposed SM3D network. Although SM3D is a multi-task model, rather than a pure 3D detector as PL-STEREO-FP, it is still more efficient (more than 2 times faster). Considering the accuracy-efficiency trade-off, the proposed SM3D is more suitable for real-time application than stereo models. Table 3 shows the inference time of each module and the final model of SM3D. As the result of jointly usage of data, SM3D is 18.3% faster than a linear summation of implementing Mapping and Detection module independently, which validates our efficient model design and potential impact in real-time application.

4. CONCLUSION

In this work, we present an efficient multi-task framework SM3D, taking monocular snippets as input to simultaneously estimate map of ego-motion and 3D detection/location of surrounding objects. Using extensive improvements on framework design and novel loss function, we end-to-end train both the Mapping Module and the 3D Detection Module. Extensive results of both mapping and detection show that each module of SM3D network outperforms their state-of-the-arts baselines in accuracy. More importantly, our monocular multi-task SM3D is more efficient than single-task stereo 3D detector. The inference time is significantly faster comparing to method with separate module computation. In the future, we will strengthen the inner connection and collaboration of two modules to improve the performance in each task.

5. REFERENCES

[1] Randall Smith, Matthew Self, and Peter Cheeseman, *Estimating Uncertain Spatial Relationships in Robotics*, Springer New York, New York, NY, 1990.

[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, 2015.

[3] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, 2017.

[4] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow, "Digging into self-supervised monocular depth prediction," in *ICCV*, 2019.

[5] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, 2004.

[6] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, 2016.

[7] Zhe Cao, Abhishek Kar, Christian Hane, and Jitendra Malik, "Learning independent object motion from unlabelled stereoscopic videos," in *CVPR*, 2019.

[8] Zhaoyang Lv, Kihwan Kim, Alejandro Troccoli, Deqing Sun, James Rehg, and Jan Kautz, "Learning rigidity in dynamic scenes with a moving camera for 3d motion field estimation," in *ECCV*, 2018.

[9] Yuliang Zou, Zelun Luo, and Jia-Bin Huang, "Df-net: Unsupervised joint learning of depth and flow using cross-task consistency," in *ECCV*, 2018.

[10] Zhichao Yin and Jianping Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *CVPR*, 2018.

[11] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *CVPR*, 2016.

[12] A. Mousavian, D. Anguelov, J. Flynn, and J. Košecká, "3d bounding box estimation using deep learning and geometry," in *CVPR*, 2017.

[13] A. Naiden, V. Paunescu, G. Kim, B. Jeon, and M. Leordeanu, "Shift r-cnn: Deep monocular 3d object detection with closed-form geometric constraints," in *ICIP*, 2019.

[14] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *CVPR*, 2019.

[15] Y. Chen, S. Liu, X. Shen, and J. Jia, "Fast point r-cnn," in *ICCV*, 2019.

[16] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals using stereo imagery for accurate object class detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[18] Y. Wang, W. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *CVPR*, 2019.

[19] 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," in *ICLR*, 2020.

[20] Rui Qian, Divyansh Garg, Yan Wang, Yurong You, Serge Belongie, Bharath Hariharan, Mark Campbell, Kilian Q. Weinberger, and Wei-Lun Chao, "End-to-end pseudo-lidar for image-based 3d object detection," in *CVPR*, 2020.

[21] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *CoRR*, 2019.

[22] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *CVPR*, 2018.

[23] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *2011 IEEE Intelligent Vehicles Symposium (IV)*, 2011.

[24] B. Xu and Z. Chen, "Multi-level fusion based 3d object detection from monocular images," in *CVPR*, 2018.

[25] F. Manhardt, W. Kehl, and A. Gaidon, "Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape," in *CVPR*, 2019.

[26] Z. Qin, J. Wang, and Y. Lu, "Monogrnet: A geometric reasoning network for monocular 3d object localization," in *AAAI*, 2019.

[27] X. Weng and K. Kitani, "Monocular 3d object detection with pseudo-lidar point cloud," in *ICCVW*, 2019.