Towards Building the Semantic Map from a Monocular Camera with a Multi-task Network

Lei Fan[1], Yucai Bai[2], Ziyu Pan[1] and Long Chen[1*]

***Abstract*— In many robotic applications, especially for the autonomous driving, understanding the semantic information and the geometric structure of surroundings are both essential. Semantic 3D maps, as a carrier of the environmental knowledge, are then intensively studied for their abilities and applications. However, it is still challenging to produce a dense outdoor semantic map from a monocular image stream. Motivated by this target, in this paper, we propose a method for large-scale 3D reconstruction from consecutive monocular images. First, with the correlation of underlying information between depth and semantic prediction, a novel multi-task Convolutional Neural Network (CNN) is designed for joint prediction. Given a single image, the network learns low-level information with a shared encoder and separately predicts with decoders containing additional Atrous Spatial Pyramid Pooling (ASPP) layers and the residual connection which merits disparities and semantic mutually. To overcome the inconsistency of monocular depth prediction for reconstruction, post-processing steps with the superpixelization and the effective 3D representation approach are obtained to give the final semantic map. Experiments are compared with other methods on both semantic labeling and depth prediction. We also qualitatively demonstrate the map reconstructed from large-scale, difficult monocular image sequences to prove the effectiveness and superiority.**

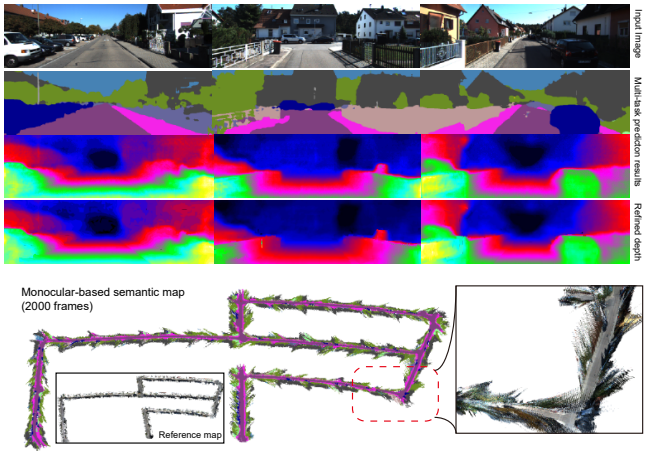


Fig. 1. The 3D semantic reconstruction of proposed method. The first row demonstrates the input image while the following two rows are outputs of our network. The fourth row is the depth after refinement. The semantic map with 2000 monocular images from KITTI odometry dataset [12] is displayed to prove the effectiveness.

I. INTRODUCTION

Urban environment perception, as one of the core issues for autonomous driving and other road scene related applications, provides valuable information for further localization inferring, obstacle avoidance [7], drivable area extraction [20] and etc. To successfully understand the 3D world, various methods have been developed for scene parsing and 3D structure acquiring based on visual sensors. The dense semantic map is a suitable presentation method to provide necessary information. However, compared to stereo vision-based approaches, producing a large-scale dense semantic map remains challenging and computation-consuming. In this paper, we propose a method with a new multi-task semantic and depth prediction model and a superpixel-based refinement to overcome these limitations for monocular semantic mapping.

The current leading scene parsing methods, such as the PSPNet [26] and the Deeplab [5], are designed to assign a semantic label to each pixel. The semantic segmentation methods are mostly based on the encoder-decoder network. The architecture first encodes the input images into concentrated features and then decodes to potentials belonging to each semantic category. For the estimation of depth from only one image, recent works [11], [19], [17], [27] deal with this issue as a parametric learning process, which contains models including conditional random fields, logistic regression, and CNNs. The leading approaches using the CNNs can predict disparities from a single image based on strong fitting abilities and gain significant accuracy.

In particular, the two tasks, i.e., semantic labeling and depth prediction, are strongly interrelated to each other and separate computations bring redundancies. A popular paradigm for predicting different labels is to leverage a multi-task network [24], [10], [16], [22] which combines multiple loss functions to learn diverse objectives. Some specific features that are hard to capture for one task but easy for another can be efficiently obtained within multi-task networks. The risk of overfitting is also reduced with the multi-task training process. In order to explore the potential relationship between depth and semantic knowledge while reducing the computation cost, we design a network with a shared encoder and two decoders to predict depth and semantic information from a single image simultaneously. The network avoids the checkerboard artifacts caused by transposed convolutions during depth prediction by using a specifically designed decoder. The output depth is benefited from the connection between two decoders. We also derive the ASPP module in both decoders to extend the fields-of-view.

*The corresponding author is Long Chen.
chenl46@mail.sysu.edu.cn
[1]Lei Fan, Ziyu Pan and Long Chen are with School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong, P.R.China.
[2]Yucai Bai is with College of Computer Science, Sichuan University, Chengdu, Sichuang, P.R.China.

Recent visual Simultaneous Localization and Mapping (SLAM) systems could provide a sparse set of 3D points while positioning itself. The density of maps reconstructed from SLAM systems is not qualified for further demanding scene understanding. Stereo matching techniques require an additional camera to yield the disparity map. A recent stereo-based reconstruction framework [3] utilizes instance-level segmentation and moving-object detection approaches to map static scene and dynamic targets. Other reconstruction methods, such as multi-view stereo, could fail to deal with autonomous driving scene due to textureless area and low parallax. To reconstruct a 3D map with semantic information, a monocular camera can offer sufficient data for labeling semantic, where the difficulty mainly lies in the dense disparity calculation. The monocular semantic reconstruction method [18] proposed by Kundu et al. combines a visual SLAM system and scene segmentation results by a 3D Conditional Random Filed (CRF) to model the environment with volumetric representation. Limited by the 3D representation method, the output map loses pixel-level accuracy. In this paper, with the depth and segmentation results from the proposed multi-task network, we further adopt the novel Simple Non-Iterative Clustering (SNIC) superpixelization method [2] to reduce the inconsistency of the depth prediction result, such as the subtle depth fluctuation of planar surfaces from the decoder. The camera pose knowledge can be given by the visual odometry approach [21] or the GPU/IMU device. We concerned the memory requirement of large-scale maps and save the final map by the vertices of superpixel units after polygonal partitioning. An example of our method is shown in Fig. 1.

In the experiment, we quantitatively evaluate the results of both depth prediction and scene segmentation of proposed multi-task network on the Cityscape dataset [8]. The reconstruction of large-scale maps is then conducted in the 3D space with monocular image streams from the KITTI benchmark [12]. In both cases, our method performs better in accuracy as well as in terms of computational efficiency. The rest of the paper is organized as follows. We first review relative works on simultaneous prediction networks and monocular reconstruction in Section II, introduce the proposed method in Section III and demonstrate the experimental result in Section IV. The conclusion is stated in Section V.

II. RELATED WORK

Depth prediction for scene understanding used to heavily rely on stereo vision [15], [25], [6]. Recent studies have been made progress in scene geometric understanding from the monocular camera. An encoder-decoder architecture [19] based on ResNet is proposed by Laina et al., which performs residual learning to predict dense depth maps from a single image. The new effective up-projection structure is adopted to avoid the checkerboard artifacts during the feature map upsampling. Scene segmentation is another active field. The current leading segmentation network [5], called Deeplab v3+, is based on their former work [4] which could extract the boundaries unambiguously referring to the recovered structural information. For semantic reconstruction containing both depth and semantic prediction, separately applying these methods could bring higher computation cost and neglect the shared information between two tasks.

Multi-task learning techniques are designed to use the transfer feature between different tasks by jointly predict labels from a single model. Multi-task networks are adopted in the face attribute estimation, the contour detection, the semantic segmentation [13], etc. Neven et al. propose a network [22] which combines scene segmentation, instance segmentation, and depth prediction into an integrated network based on the ENet. Kendall et al. use the Deeplab v3 as the fundamental architecture [16]. They conduct experiments with uncertainty weight losses demonstrating the superiority and effectiveness of homoscedastic uncertainty for multi-task learning. However, these networks derive the same decoder architecture for each task ignoring the difference of outputs. In the proposed network, specialized decoders are designed to deal with different tasks, and connections between two decoders to transfer available information.

Semantic reconstruction can be basically divided into two categories. The first kind of methods are inheritors of 2D semantic segmentation results [18], [3]. For monocular-based reconstructions, Kundu et al. propose an approach [18] to jointly infer geometric structure and 3D semantic labeling with a CRF model. The experimental results are good but with limitations. Due to the resolution and the ability of volumetric occupancy map, the output misses structural details. The second concerns the need to simultaneously provide semantic and geometric information in the 3D space. The incremental semantic reconstruction approach [23] proposed by Vineet et al. builds the urban environment on a hash-based technique and a mean-field inference algorithm. The traditional stereo matching and visual odometry method are adopted to obtain basic 3D knowledge.

Superpixel segmentation [1], [2] has been applied to promote stereo matching results. The matching algorithm of Yamaguchi et al. [25] called the SPS-st, whose formulation is based on the slanted-plane model with plain-fitting technique. To reduce the memory of large-scale reconstruction while maintaining pixel-level details, the map is stored with the vertices of superpixel after changed into polygons [2]. The advancement of depth refinement and representation method are also conducted in the experiment.

III. MONOCULAR CAMERA-BASED SEMANTIC RECONSTRUCTION

As shown in Fig. 2, the input to the proposed method is a monocular image stream. The disparity map and the semantic segmentation are precalculated from the proposed multi-task network, whose architecture will be introduced in the following part. We use a superpixel segmentation method with additional depth and semantic knowledge to perform smoothing to the original depth. Further depth refinement is obtained on the planar surface, i.e. the road, under supervised by the semantic segmentation result. Each superpixel is then

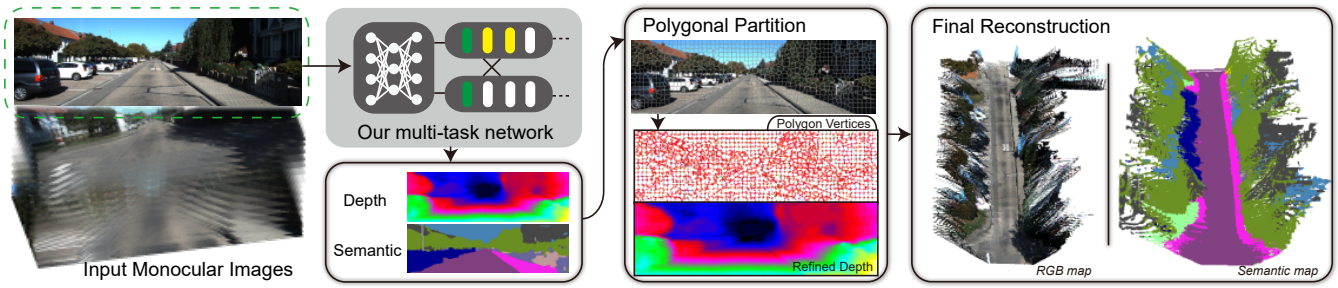


Fig. 2. Overview of our monocular-based 3D reconstruction method. Given an image stream, the proposed multi-task network predicts its depth and semantic simultaneously. The polygonal superpixel segmentation is applied to refine the depth output while reducing the map memory cost. With the camera pose knowledge, we give the final 3D semantic maps rendering with RGB data and semantic information, respectively.

regarded as the basic unit for map storage after transforming into polygons, which saves memory requirement dramatically especially for large-scale structural space.

A. Multi-task Network for Joint Inferring

The proposed network is demonstrated in Fig. 3. The network contains three basic modules which are the shared encoder, the depth decoder, and the semantic decoder. For the shared encoder, we use the ResNet-50 [14] to produce rich and contextual features. The ResNet-50 is a trade-off choice between accuracy and memory requirements compared to the ResNet-101. Both decoders receive the same feature map from ResNet-50 as input while using ASPP modules to promote contextual awareness. The ASPP module contains three dilated convolutions ($rate = 6, 12, 18$), a 1×1 convolution and an image pooling layer. By applying filters of multiple rates to probe input feature maps, the network enlarges its fields-of-view to capture objects or image context of different scales.

The ASPP module in the depth decoder is to capture the discrepancy between nearby features to finally convert into depth. In the following upsampling process, we choose the fast up-projection structure instead of the bilinear interpolation in the original Deeplab v3+ [5], which reduces the blur effect of boundaries during the depth prediction. The fast up-projection also benefits results to avoid checkerboard artifacts compared to the transpose convolution upsampling, especially for the road surface. The fast up-projection calculates group of filters with 3×3, 2×2, 2×3 and 2×2 convolutions on the feature map and interleaves them into a unified output matching the upsampling shape. Compared to deconvolutions, the interleaving mechanism helps to reduce these noises during depth prediction. The detailed difference is discussed in the experiment. We use the scale-invariant error as the training loss for the depth prediction part, which adapts to multi-scale objects. The loss formulation L_d is defined as

$$L_d(\tilde{y}_d, y_d) = \frac{1}{n}\sum_i^n d_i^2 - \frac{1}{2n^2}(\sum_i^n d_i)^2, \tag{1}$$

where $\tilde{y}_d$ and y_d denote the depth prediction and the depth ground truth, respectively, n is the volume of valid pixels and $d_i = \log \tilde{y}_{d,i} - \log y_{d,i}$.

In the semantic decoder, we connect the depth feature with a $3 \times 3 \times 48$ convolution to provide useful information. The pixel sharing the same depth value or gradient has higher possibility to belong to the same semantic category. The cross-entropy loss is adopted to learn pixelwise label probabilities by averaging the loss over the pixels in each batch. The loss function L_s for semantic prediction is

$$L_s(\tilde{y}_s, y_s) = \sum_i^n (-\tilde{y}_{s,i} \log y_{s,i} - (1 - \tilde{y}_{s,i}) \log(1 - y_{s,i})), \tag{2}$$

where $\tilde{y}_s$ is the semantic prediction and y_s is the corresponding ground truth. We minimize the combined loss, which is $L = \alpha L_s(\tilde{y}_s, y_s) + (1 - \alpha) L_d(\tilde{y}_d, y_d)$, during our multi-task training process. The α is a balancing factor and is set to 0.75 in our experiments.

Training: The training process is started form the ResNet-50 model pretrained on the ImageNet-1K dataset [9]. Following the training protocol of Deeplab v3+ [5], we employ the same learning rate schedule, i.e., the "poly" policy. The learning rate is set to 0.007 while crop size is 512×1024 for the Cityscape dataset [8] and 192×624 for the KITTI dataset [12]. We use a random image scale and flip for data augmentation. In order to maintain the output of both decoders in the same interval, the possibility of semantic prediction belongs to $[0, 1]$, and we transform the disparity label to log space divided by the maximum value. The proposed model is trained on the fine-annotated Cityscape dataset [8] with 1575 images containing both semantic and depth ground truths. 19 labels are chosen as the training label. To further conduct on the KITTI dataset [12], the network is trained with 200 groups of images provided by the KITTI dataset.

B. Polygonal Superpixel Partitioning

Given the input image $\mathcal{I}$, we adopt the superpixel segmentation method to provide basic partition results $S = \{S_1, \ldots, S_k\}$. The image $\mathcal{I}$ is transformed into the CIELAB color space. After initializing the centroid of pixels belonging to a regular grid, we use the following formulation to measure the distance between kth centroid $C[k]$ and jth

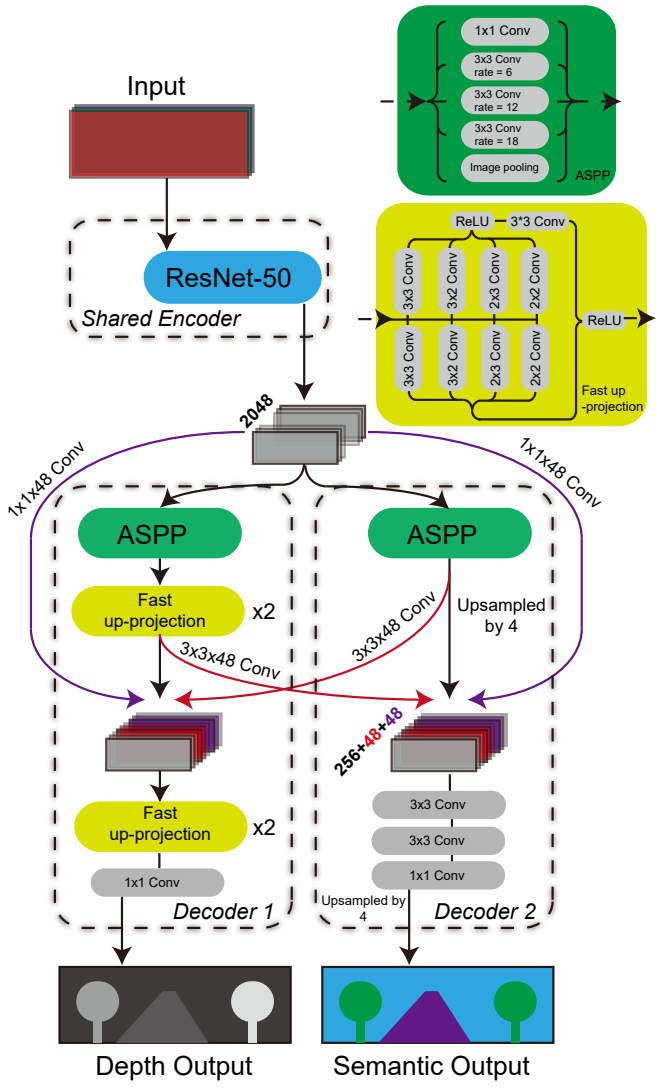


Fig. 3. The proposed network architecture. Given an input RGB image, the network simultaneously produces the depth and the semantic prediction results. The detailed architecture of ASPP and fast up-projection display on the up-right corner.

candidate pixel, which is defined as

$$d_{j,k} = \sqrt{\frac{\|\mathrm{x}_j - \mathrm{x}_k\|_2^2}{s} + \frac{\|\mathrm{c}_j - \mathrm{c}_k\|_2^2}{m} + h(\tilde{y}_{s,j}, \tilde{y}_{s,k})}, \quad (3)$$

where s and m are two normalizing factors for spatial and color distances, and $h(.,.)$ is set to 0 when two pixels sharing the same semantic label and a larger value otherwise. We follow the simple non-iterative clustering [2] to update the boundary of superpixels, which requires less memory and performs faster.

With the superpixel segmentation result, we perform boundary tracing to each superpixel and transform superpixels into polygons. The vertices are then stored to represent each superpixel which saves memory compared to all boundary pixels. For the outdoor scene, especially for the driving environment, it is mainly composed of structural elements, such as the building and the road. Changing into polygons has an acceptable influence on the accuracy of boundaries.

TABLE I

THE RESULTS OF OUR METHOD ON THE CITYSCAPE BENCHMARK [8]. WE EVALUATE DIFFERENT DECODER ARCHITECTURES. THE TERM **DECONV.**, **BI. INTERP.**, AND **FAST UP-PROJ.** DENOTE DECONVOLUTION LAYERS, THE BILINEAR INTERPOLATION AND THE FAST UP-PROJECTION MODULE, RESPECTIVELY, AND **RC** IS THE RESIDUAL CONNECTION BETWEEN DECODERS.

Loss		Depth Estimation	
		Mean Error[px]	RMS Error[px]
Decoder	***deconv.***	2.103	4.322
	bi. interp.	2.391	5.132
	fast up-proj.	1.913	4.001
	fast up-proj.+RC	1.793	3.851
	fast up-proj.+RC+refinement	1.611	3.423

C. Depth Refinement

The well-adopted assumption is made that pixels located in the same superpixel are belonging to the same depth plane surface. With the depth prediction $\tilde{y}_d$, the plane-fitting is performed to each superpixel to eliminate noises or fluctuations of the initial prediction. For the superpixel S_i, the plane parameter is defined as (a_i, b_i, c_i), which means the final depth for pixel $p(u, v)$ in S_i is calculated as $a_i u + b_i v + c_i$. After refinement, we further smooth the road surface with random sample consensus algorithm. Note the depth refinement can be performed on other network prediction results. The improvement of depth is demonstrated in the experiment.

D. Large-scale Reconstruction

After producing the polygonal partition results with a plane function assigned to each polygon, we project polygons into the world coordinate. The projection parameter follows the corresponding training data, and we derive the camera pose from the visual odometry method [21] or the available ground truth. Storing vertices requires much less memory compared to point clouds while maintaining comparable pixel-level accuracy. During reconstruction, we give each polygon a semantic label with the multi-task network prediction result.

IV. EXPERIMENTS

The experiment is conducted in two parts. In the first part, we compare the proposed multi-task model performance with other networks [27], [17], [5], [26], [16], [22] in both semantic and depth prediction. The influence of adding connections between decoders and up-projection module is also demonstrated. The depth results after refinement are also shown in this part. We evaluate our model on the Cityscape dataset [8], which aims at road scene understanding tasks. The Cityscape dataset contains images with both the fine labeled semantic ground truth and the depth ground truth. The semantic segmentation contains 34 categories, and the depth ground truth is generated with the popular SGM stereo matching method [15]. The Cityscape dataset is collected with a camera from several cities in different weather and

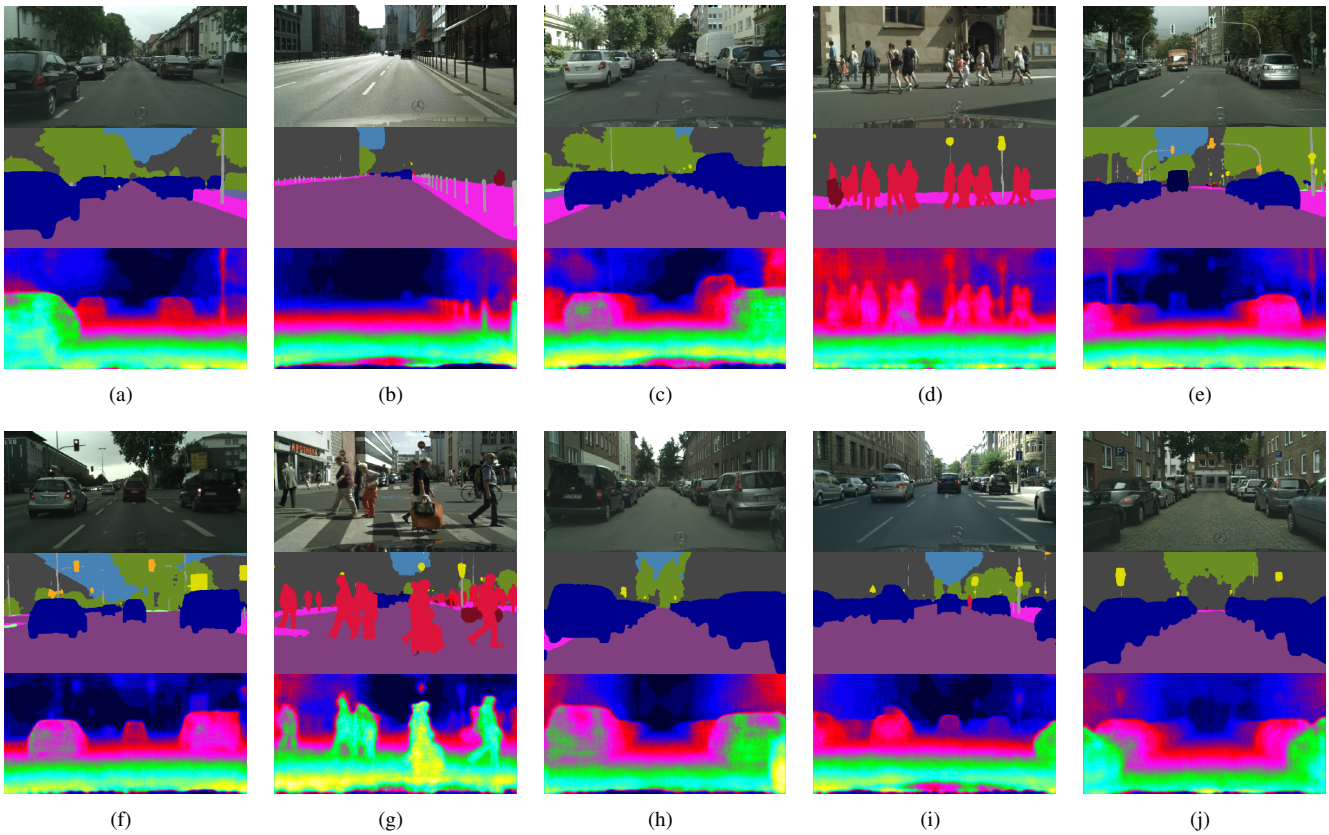


Fig. 4. The prediction result on the Cityscape validation dataset [8]. Each group of result contains the input image, the semantic prediction and the depth prediction from top to bottom.

TABLE II

THE RESULTS ON THE CITYSCAPE BENCHMARK [8]. WE SEPARATE METHODS INTO THREE CATEGORIES BY THEIR OUTPUTS. NOTE THE COMPARISON IS NOT ENTIRELY FAIR, AS MANY METHODS USE ENSEMBLES OF DIFFERENT TRAINING DATASETS AND TRAINING IMAGE SIZES.

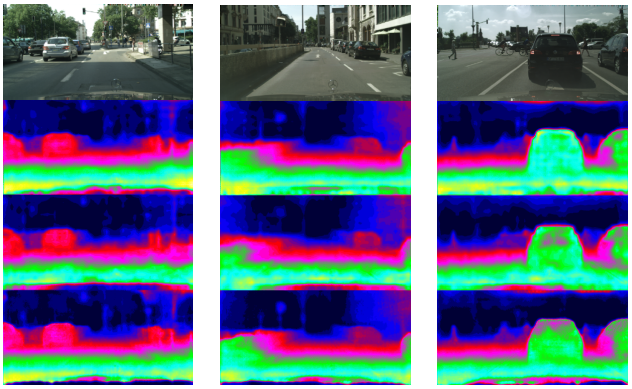
Loss	Segmentation		Depth Estimation						
	IoU Cla.	IoU Cat.	Mean Error$[px]$	Abs Rel	RMS Error$[px]$	Sq Rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monocular depth prediction only methods									
Zhou et al. [27]	-	-	-	0.267	7.85	2.686	0.577	0.840	0.937
Kumar et al. [17]	-	-	-	0.393	10.50	4.683	0.352	0.689	0.905
Semantic segmentation only methods									
Deeplab v3+ (ResNet-101) [5]	0.821	0.920	-	-	-	-	-	-	-
PSPNet [26]	0.802	0.906	-	-	-	-	-	-	-
Joint depth and semantic prediction methods									
Kendall et al. [16]	0.785	0.899	2.920	-	5.880	-	-	-	-
Neven et al. [22]	0.593	0.804	-	-	-	-	-	-	-
Our Method (ResNet-50)	0.708	0.879	1.793	0.220	3.851	2.310	0.655	0.851	0.933

illumination conditions. Qualitative results on the large-scale dataset, namely the KITTI odometry dataset [12], are displayed in the second part to show the effectiveness of proposed monocular reconstruction method. The urban scene provided by the KITTI benchmark is quite challenging, as it covers a large area and contains significant depth variation. The proposed method shows its advancement in dealing with images of large man-made environment.

A. *Multi-task Network*

The performance of proposed network is evaluated in two different aspects. We first change the upsampling component in our depth decoder into the deconvolution, the bilinear interpolation, and the fast up-projection module utilized in this paper. The three modules upsample the feature map into the output with target sizes. As shown in Fig. 5, the blur effect heavily occurs in the output with the decoder of deconvolution. Checkerboard effects is easy to notice especially for planar surfaces, such as the road. Due to the mechanism of bilinear interpolation which is widely used in semantic prediction networks, the uneven disparity happens. The accuracy of different decoder architectures are displayed in Tab. I. Using the residual connection and further depth refinement with superpixels also improves the performance. Examples of our results on the Cityscape validation set [8] is shown in Fig. 4.

We then compare to a number of related works [27], [17],



(a) (b) (c)

Fig. 5. Comparison after modifying the decoder architecture. For each group of depth prediction, we separately use the deconvolution layer, the bilinear interpolation, and the fast up-projection from top to bottom for recovering to the original size.

[5], [26], [16], [22] in Tab. II. Three categories of methods are demonstrated. In terms of semantic segmentation, our model lags behind the Deeplab v3+. The reason is partly due to the encoder difference, i.e., the ResNet-50 contains much fewer parameters and is shallower than the ResNet-101. Also, the proposed model is pretrained on limited dataset due to the lack of ground truths of both depth and semantic. For depth prediction, the proposed method clearly outperforms other models. The detailed definition of loss metric Abs Rel, Sq Rel, and etc. follows [11]. The higher of $\delta < 1.25, 1.25^2, 1.25^3$ denote better performance, and the other metrics are on the contrary.

B. Reconstruction on the KITTI Dataset

The qualitative experiments of outdoor scenes are based on the KITTI odometry dataset [12]. In this part, we intend to show the 3D semantic reconstructions on long monocular image sequences covering a large urban area. The camera pose follows the provided ground truth. As demonstrated in Fig. 6, we choose the sequence 0, 3 and 6 from KITTI odometry dataset with 1000 frames, 800 and 1000 frames, respectively. In the first example (Fig. 6 (a)), the vehicle located on the roadside is clearly presented due to correct semantic segmentation. With the depth information, the position of vehicles in the 3D world can be obtained. The key aspect of the proposed 3D map is to convey the height information which could assist road area extraction and obstacle avoidance, which is well-displayed in the RGB textured map. The large-scale map with semantic information is displayed with polygons in the 3D space, which saves the required memory space, while the RGB textured map is presented based on point clouds for details. Another map of a countryside road with two RGB textured aerial views is shown in Fig. 6 (b). We provide groups of semantic and depth prediction on the left side with the depth after refinement displayed at the last row. The superpixel segmentation helps extract the boundary of objects, which eliminates the potential blur effect in the depth of the network output. The final example Fig. 6 (c) depicts

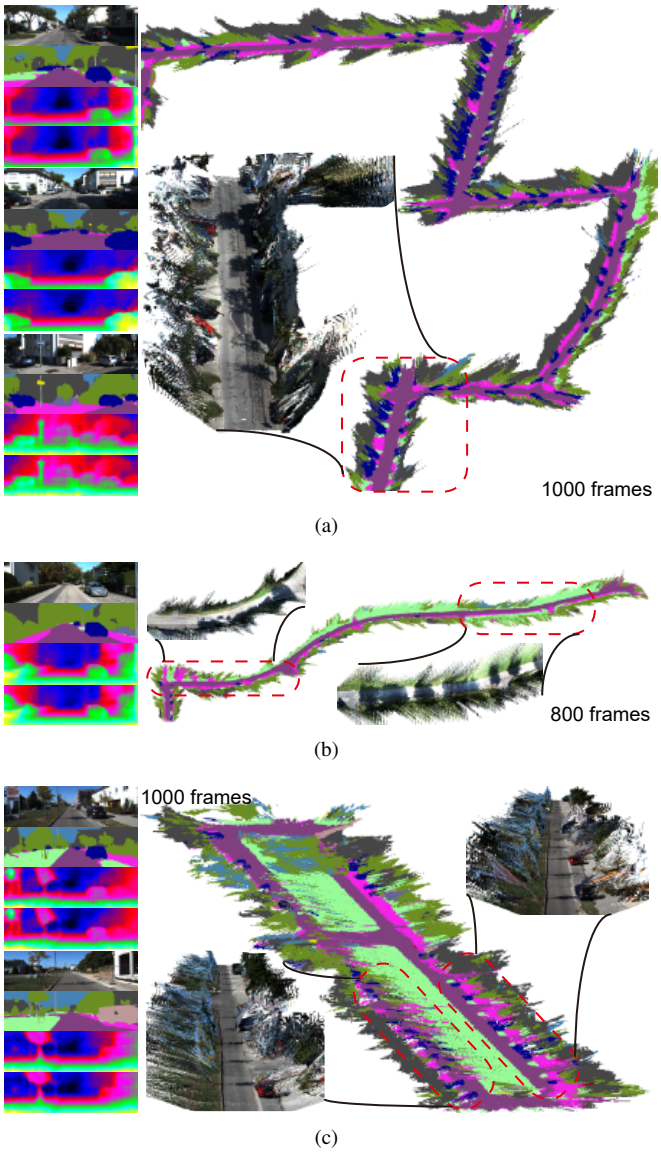


Fig. 6. Three demonstrations of our 3D semantic map on long outdoor image sequences. We also zoom in and textured with RGB data to illustrate details. The example group of semantic and depth prediction are displayed on the left side. The depth after refinement is at the last row of each group.

the road area with a loop.

V. CONCLUSIONS

In this paper, we presented a method to reconstruct dense large-scale semantic maps based on a monocular camera. The semantic and depth are first predicted jointly by a novel multi-task network. The proposed network improves results with ASPP modules enlarging the fields-of-view and the fast up-projection in decoder reducing checkerboard artifacts. The depth prediction also yields available data from the semantic decoder with the residual connection. The superpixels are extracted with a depth- and semantic-aware manner and further changed into polygons. Inconsistencies, i.e., the fluctuation in the depth prediction, are reduced after performing plane-fitting to each segment. The semantic 3D

representation method benefits in both memory and calculation especially in a structural space, such as the urban driving scene. The experimental results show the effectiveness.

REFERENCES

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, Sabine Süsstrunk, et al. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[2] Radhakrishna Achanta and Sabine Süsstrunk. Superpixels and polygons using simple non-iterative clustering. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4895–4904. Ieee, 2017.

[3] Ioan Andrei Bârsan, Peidong Liu, Marc Pollefeys, and Andreas Geiger. Robust dense mapping for large-scale dynamic environments. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*, 2018.

[6] Long Chen, Lei Fan, Jianda Chen, Dongpu Cao, and Feiyue Wang. A full density stereo matching system based on the combination of cnns and slanted-planes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017.

[7] Long Chen, Lei Fan, Guodong Xie, Kai Huang, and Andreas Nüchter. Moving-object detection from consecutive stereo pairs using slanted plane smoothing. *IEEE Transactions on Intelligent Transportation Systems*, 18(11):3093–3102, 2017.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[10] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012.

[13] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 564–571, 2013.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 807–814. IEEE, 2005.

[16] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv preprint arXiv:1705.07115*, 3, 2017.

[17] Arun CS Kumar, Suchendra M Bhandarkar, and Mukta Prasad. Monocular depth prediction using generative adversarial networks. In *1st International Workshop on Deep Learning for Visual SLAM,(CVPR)*, volume 3, page 7, 2018.

[18] Abhijit Kundu, Yin Li, Frank Dellaert, Fuxin Li, and James M Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In *European Conference on Computer Vision*, pages 703–718. Springer, 2014.

[19] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.

[20] Qingquan Li, Long Chen, Ming Li, Shih-Lung Shaw, and Andreas Nüchter. A sensor-fusion drivable-region and lane-detection system for autonomous vehicle navigation in challenging road scenarios. *IEEE Transactions on Vehicular Technology*, 63(2):540–555, 2014.

[21] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[22] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Fast scene understanding for autonomous driving. *arXiv preprint arXiv:1708.02550*, 2017.

[23] Vibhav Vineet, Ondrej Miksik, Morten Lidegaard, Matthias Nießner, Stuart Golodetz, Victor A Prisacariu, Olaf Kähler, David W Murray, Shahram Izadi, Patrick Pérez, et al. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 75–82. IEEE, 2015.

[24] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. *arXiv preprint arXiv:1805.04409*, 2018.

[25] Koichiro Yamaguchi, David McAllester, and Raquel Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *European Conference on Computer Vision*, pages 756–771. Springer, 2014.

[26] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.

[27] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017.