REAL-TIME SEMANTIC SLAM WITH DCNN-BASED FEATURE POINT DETECTION, MATCHING AND DENSE POINT CLOUD AGGREGATION

B. Vishnyakov [1], I. Sgibnev , V. Sheverdin, A. Sorokin, P. Masalov, K. Kazakhmedov, S.Arseev

FGUP «State Research Institute of Aviation Systems», Russia, 125319, Moscow, Viktorenko street, 7 - (vishnyakov, sgibnev, sheverdin, ans, masalov, kkirill, sarseev)@gosniias.ru

Commission II, WG II/4

KEY WORDS: multi-sensor platform, autonomous vehicle, SLAM, CNN, dynamic scene analysis, semantic segmentation, data fusion, dynamic scene reconstruction, dense point cloud, feature point, stereo disparity estimation

ABSTRACT:

In this paper we present the semantic SLAM method based on a bundle of deep convolutional neural networks. It provides real-time dense semantic scene reconstruction for the autonomous driving system of an off-road robotic vehicle. Most state-of-the-art neural networks require large computing resources that go beyond the capabilities of many robotic platforms. We propose an architecture for 3D semantic scene reconstruction on top of the recent progress in computer vision by integrating SuperPoint, SuperGlue, Bi3D, DeepLabV3+, RTM3D and additional module with pre-processing, inference and postprocessing operations performed on GPU. We also updated our simulated dataset for semantic segmentation and added disparity images.

1. INTRODUCTION

The task of determining the proper position of a robotic platform is closely related to calculating the coordinates of objects in the surrounding space. The navigation task is to find the position of the vehicle in relation to the 3D scene model (sparse or dense). Usually, navigation and mapping tasks are solved simultaneously using Simultaneous Localization And Mapping (SLAM) algorithms (R. Mur-Artal et al., 2017; J. Engel et al., 2018; Engel et al., 2018; C. Forster et al., 2014; T. Qin et al., 2018; B. Vishnyakov et al., 2020b).

In the last few years, there has been significant progress in solving the data association problem through matching, optical flow calculation, scene reconstruction and segmentation using Deep Convolutional Neural Networks (DCNN).

During the last few decades, a number of methods for identifying, matching, and analyzing image key points have been introduced. Currently, both relatively simple methods based on image gradients and more complex methods of feature analysis using deep convolutional neural networks are widely used. Mapping feature points using neural networks allows us to achieve a more reliable solution, because feature points, calculated using standard methods, group mostly on high contrast objects, such as grass, buildings and trees. Algorithms based on DCNN are trained to detect feature points more wisely, and we get a more uniform distribution of feature points in the image as a result.

Sparse 3D reconstruction can be achieved during SLAM process by calculation and tracking of feature points with inertial sensor prediction. For dense reconstruction you can use a standard mapping pipeline with any classic (H. Hirschmuller, 2005; J. Sun et. al, 2003; V. Kolmogorov et. al, 2001) or DCNN (X. Cheng et. al, 2020; A. Badki et. al, 2020) stereo disparity map calculation method. Neural network stereo disparity is characterized by pretty accurate values for the entire image. In our work, we tried to combine the latest machine learning approaches with state-of-the-art photogrammetry methods.

Finally, there is a problem of making a 3D dense point cloud semantic, so an autonomous platform can distinguish the obstacles. Moreover, if it goes to off-road conditions, grass or tiny bush higher than clearance looks like an impassable obstacle in the point cloud. Marking every pixel in a point cloud with a semantic class label leads to a next level of scene understanding for the autonomous vehicle, changing the passability map.

Feature point detection and descriptor extraction was performed using the SuperPoint architecture (D. DeTone et al., 2018), which is a fully convolutional neural network. Unlike other neural network approaches, SuperPoint takes an image as an input and jointly detects interest points and extracts their descriptors in one pass. This model is trained on different datasets using projective transformations, which allowed authors to obtain state-of-the-art quality assessment results compared to SIFT (D. G. Lowe et al., 2004), ORB (E. Rublee et al., 2011) and other classic methods. This approach can be widely used to solve such problems as SLAM, Structure from Motion (SfM), Multi-view Stereo, etc.

Another step in SLAM process is feature point matching. We propose a modified SuperGlue architecture (P. -E. Sarlin et al., 2020), which takes interest points and descriptors from two images as an input. This architecture is based on a graph neural network with attention units that increase the receptive field of the descriptors and ensure their cross-interaction. SuperGlue

[1] Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-399-2021 | © Author(s) 2021. CC BY 4.0 License.

outperforms other approaches in the task of feature point matching on pairs of images in complex indoor and outdoor environments. SuperPoint and SuperGlue can be integrated into modern visual odometry systems.

Two neural network architectures were considered to build a dense reconstruction. In (X. Cheng et. al, 2020), a model designed with Neural Architecture Search (NAS) (B. Zoph et al., 2016) algorithm was proposed to construct a dense reconstruction of a scene using stereo matching. NAS has already been applied to various computational tasks such as classification, detection and semantic segmentation. The basic idea of NAS is simple, namely to find the optimal architecture it is necessary to allow the network to be able to tweak operation parameters (for example, convolutions with different kernel sizes), thus better adapt the model to the task. However, so far NAS has not been applied to the dense reconstruction problem. This is partly due to the fact that modern human-designed stereo matching networks already have a huge number of parameters, thus the direct application of NAS to such massive neural networks is an extremely computationally expensive task. However, extensive experiments have shown that this network obtained using NAS outperforms many modern methods of building dense reconstruction in terms of accuracy in the KITTI Stereo 2015 tests (M. Menze, 2015).

However, for some applications, such as autonomous vehicles, it may be useful to trade off accuracy for lower latency. In (A. Badki et al., 2020), a Bi3D method was presented, which estimates the depth using binary classifications. Unlike classic neural network depth extraction methods, which determine the depth value of each pixel, this approach, based on binary classification, allows to calculate a dense disparity map – determine the pixels closer or farther than the D value. This property is a powerful mechanism for balancing accuracy and latency. Bi3D can detect objects closer than D value in just a few dozens of milliseconds, or estimate depth with quantization, where inference time depends linearly on the number of quantization levels. Bi3D can also produce full range depth estimation and provide quality state-of-the-art neural network methods for dense stereo reconstruction.

Semantic segmentation is a pixel-by-pixel classification of an image, it gives a detailed view on the shape of objects in it. In recent years we can see an increasing number of applications of semantic segmentation, such as autonomous vehicles, robotic systems and virtual reality for which an understanding of the scene is necessary. Image semantic segmentation is crucially important for the automatic control system of modern autonomous vehicles. An accurate understanding of the surrounding scene is important for navigation and decision-making by control system of robotic autonomous platform.

3D object detection is an essential component of scene perception by autonomous vehicle. Currently, most 3D object detectors heavily rely on LIDAR data for obtaining accurate depth information. The 3D detection can be divided into two groups by the type of data: LiDAR and image-based methods (Y. Wang et al., 2018; P. Li et al., 2020). LiDAR-based systems can provide accuracy and reliable point cloud of object surfaces in 3D scene. Therefore, most of the recent 3D object detection use LiDAR data to obtain the state-of-the-art results. However, LiDAR system has some disadvantages: its price, unsustainability to rainy conditions, etc. In (P. Li et al., 2020) authors proposed a model for monocular 3D objects detection using only RGB images, called RTM3D. Authors designed a fully convolutional model to predict object key points, dimensions, and orientation. This model only requires RGB images without additional data such as instance segmentation, disparity image or pseudo-lidar data (Y. Wang et al., 2018). Nevertheless, experiments on the KITTI 3D detection dataset indicate that the RTM3D surpasses many previous state-of-the-art methods in both efficiency and accuracy by a large margin.

In this paper we propose a method for solving the problem of visual-inertial dense stereo odometry and SLAM using a hybrid approach based on the advantages of machine learning methods for data association and standard photogrammetry methods for self-position calculation. We implement the indirect semantic SLAM method, where two pre-trained neural networks are used to compare feature points and calculate a dense reconstruction. Using semantic segmentation, we set each pixel in a point cloud an obstacle class label. The proposed approach allows us to achieve a much better quality of spatial position determination and semantic dense reconstruction compared to the classical methods.

2. METHOD

In this paper, we propose a method for computing odometry and dense reconstruction based on our own implementation of the SLAM method, which is an improved version of classic algorithms (C. Forster et al., 2014; R. Mur-Artal et al., 2017; B. Vishnyakov et al., 2020a) in terms of accuracy, speed, and robustness. Our method of dense stereo reconstruction is based on the neural extraction of point features and computation of disparity by means of deep neural networks.

We used SuperPoint neural network architecture for interest points detection and descriptors extraction and by using SuperGlue neural network for singular point matching.
Moreover, we collected and annotated our city and off-road dataset presented in (B. Vishnyakov et al., 2020b) for SuperPoint algorithm training. In particular, we calculated feature points and descriptors, then iteratively found and removed all outliers, and then the model was finetuned. Thus, we were able to significantly improve the quality of the algorithm in outdoor city and off-road scenes.

For feature point matching we trained the SuperGlue neural network architecture on our annotated dataset. For training we used image pairs from left and right cameras of the stereo pair with different (but close) frame timestamps. It made our method much more robust to the environment changes and raised the overall matching quality by gaining the ability to find the corresponding feature points in different parts of images.

Result images of the proposed approaches demonstrated in (Figures 1 and 2), matching score of white points is less than threshold, matching score of red points is greater than threshold.

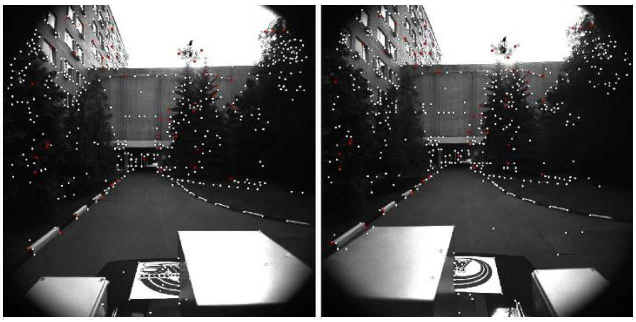


Figure 1. The result of the algorithm for interest points detection and descriptors extraction.

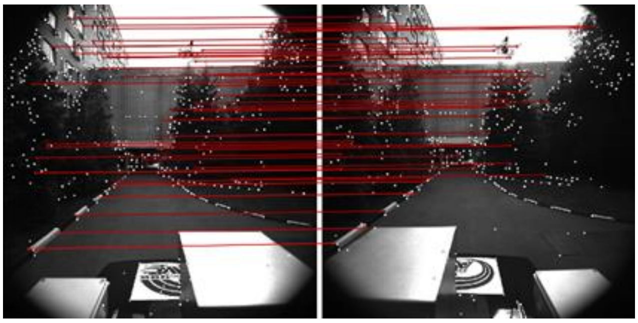


Figure 2. The result of the algorithm for matching interest points.

For dense scene reconstruction we used Bi3D algorithm, based on the data from the stereo pair. The algorithm receives rectified images as an input, and outputs dense disparity. We calculate depth from disparity using camera calibration parameters (Figure 3). We run this algorithm once or twice per second during the SLAM procedure to get dense scene reconstruction while moving the vehicle.

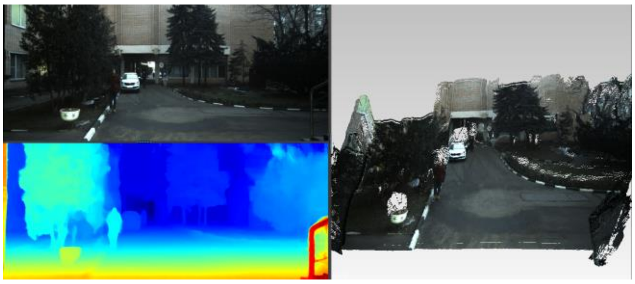
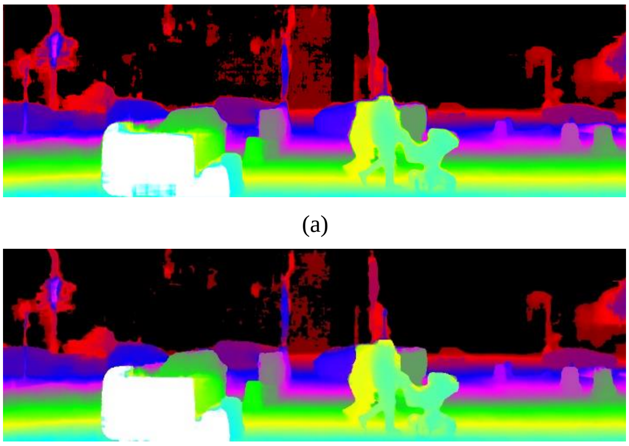


Figure 3. The result of the dense stereo reconstruction algorithm.

The main drawback of the algorithm that generates dense disparity is that any sharp edges on the disparity map (for example, object contours) are being smoothed (Figure 4 (a)). This creates significant errors on such edges, causing noisy trails to appear in corresponding regions of the point cloud. To lower the noise, we processed the disparity map (treated as a grayscale image in this context) with the following steps: use morphological dilation in the disparity map, find edges on the initial disparity map using Canny algorithm with high thresholds, use dilation again with lower radius to turn these edges into the thick lines and then fill the corresponding regions on the disparity map with data from the dilated map image. This algorithm effectively thickens the border of any object, turning continuous transition into a sharp edge (Figure 4 (b)).



(a)

(b)

Figure 4. The result of the dense stereo reconstruction algorithm (a), result after postprocessing (b).

For semantic segmentation we used our algorithm presented in (I. Sgibnev et al., 2020). This algorithm is based on the lightweight architectures as a backbone for real-time solution of semantic segmentation problem for autonomous vehicle. Moreover, we replaced DeepLabV3 (L.-C. Chen et al., 2017) decoder by DeepLabV3+ (L.-C. Chen et al., 2018) decoder and retrained it on our dataset, which improved accuracy of the scene segmentation by 1.9% mIoU.

To make our dense point cloud semantic we use a rectified image from the left camera as an input to the semantic segmentation algorithm. While calculating depth we use information about pixel class for scene reconstruction (Figure 5).

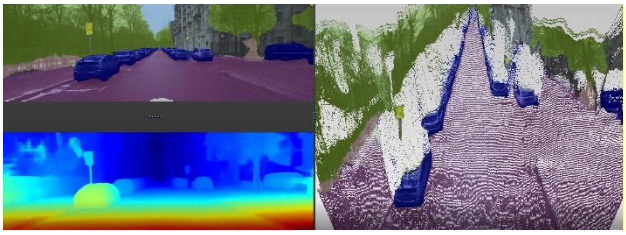


Figure 5. The result of the semantic segmentation and dense stereo reconstruction algorithm.

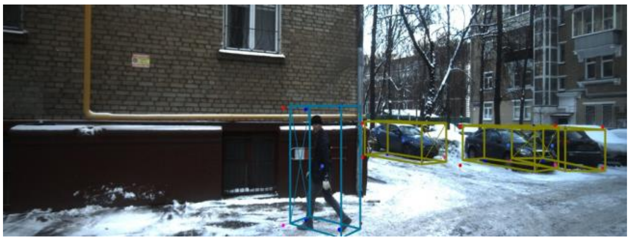


Figure 6. The result of the 3D detection algorithm.

RTM3D is a single-stage convolutional neural network for an accurate and efficient 3D object detection using only monocular image. This model focuses on 3D object detection for autonomous driving systems. Inspired by CenterNet (K. Duan, et al., 2019), this model is a fully convolutional architecture to predict 9 key points (8 vertexes and the central point of a 3D bounding box) (Figure 6). A simple and efficient architecture combines the strengths of both CNN and perspective geometry, and also achieves real-time 3D object detection using only monocular RGB images.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-399-2021 | © Author(s) 2021. CC BY 4.0 License.

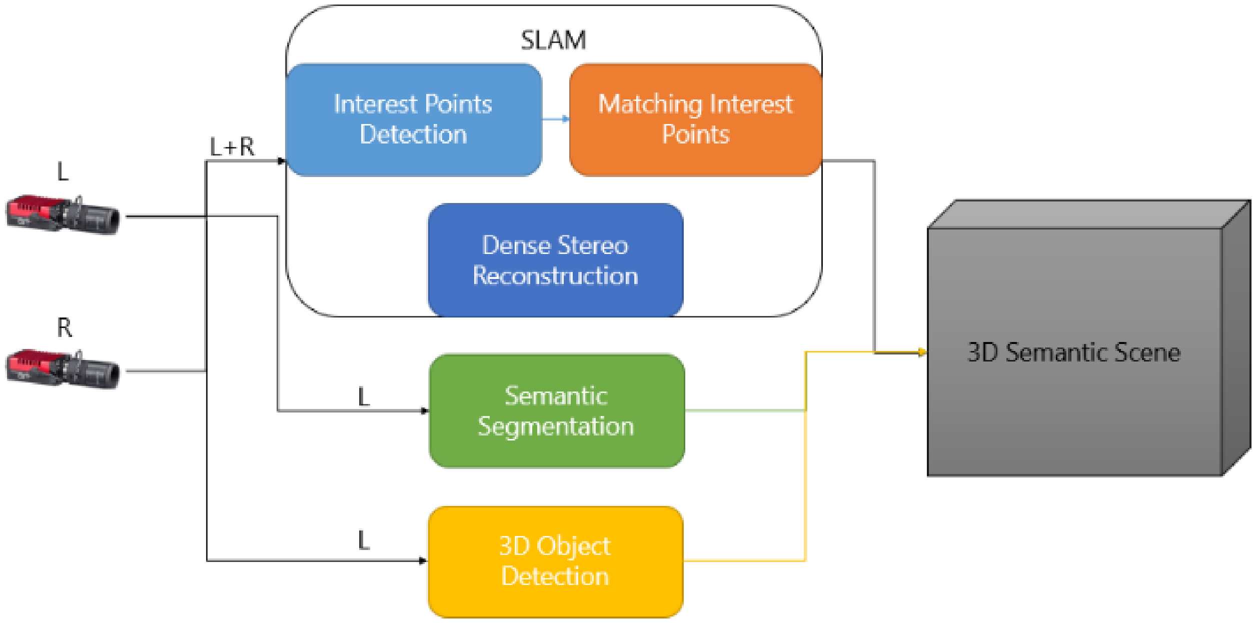


Figure 7. Scheme of constructing 3D semantic scene.

Using SuperPoint, SuperGlue, Bi3D, DeepLabV3+ and RTM3D we can build a semantic three-dimensional model which one of the key elements of an autonomous robotic vehicle. The SLAM component allows you to get an odometry and dense point cloud. The imposition of semantic segmentation and bounding boxes on dense point cloud gives us the class of each point.

Those algorithms, running in parallel (Figure 7), provide a real-time semantic, dense, and dynamic 3D-model of a scene. Using this model, artificial intelligence algorithms, running on an off-road autonomous robotic vehicle, can adjust the patency map and correct the optimal path.

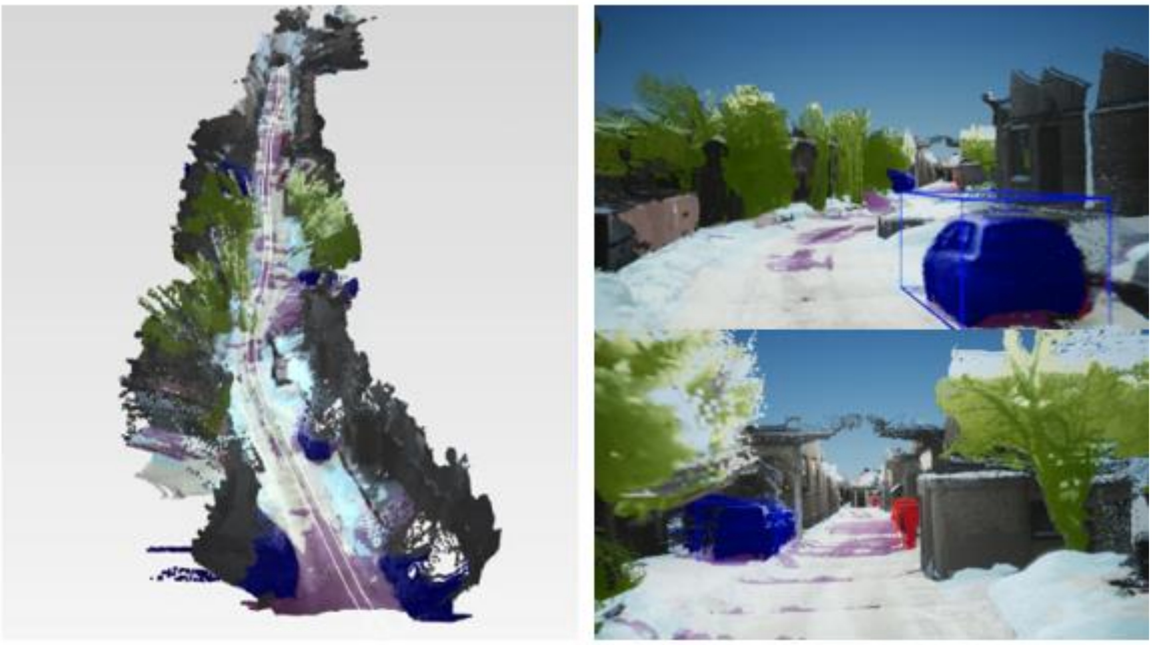


Figure 8. Visual odometry and 3D semantic scene.

3. DATASET

Moreover, we upgrade our simulated dataset consisting of 3,500,000 images.

We used our own software package based on Unreal Engine 4 graphics engine that provides a large set of tools for realistic 3D modelling (Figure 9).

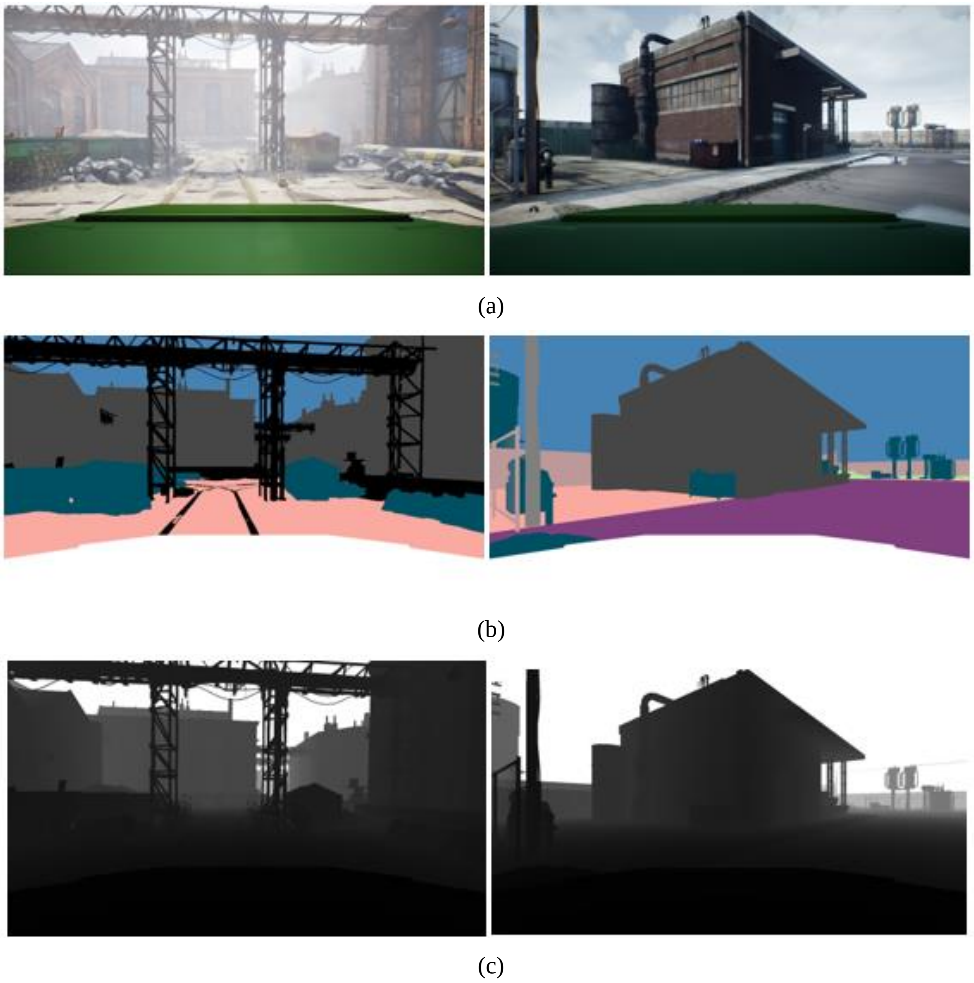


Figure 9. Input images from our datasets (a), semantic segmentation masks (b), depth mask (c).

4. IMPLEMENTATION

The peculiarity of the implementation of our method is the adaptive use of computing resources. All of the neural network models were converted to NVIDIA TensorRT format, which significantly reduced inference time and allowed these models to run in real time.

We compared implementation of these models using NVIDIA TensorRT and PyTorch libraries. NVIDIA TensorRT versions requires about thrice less time to process in comparison with PyTorch. You can find performance test results (including preprocessing and postprocessing operations) in Table 1.

Method	Input size	Time(ms) on PyTorch	Time(ms) on TensorRT (fp16)
SuperPoint	1024×1024	43	12
SuperGlue	1024×1024	54	20
Bi3D	384×1248	370	125
DeepLabV3+	384×1248	85	27
RTM3D	384×1248	97	40

Table 1. Inference time models on PyTorch and NVIDIA TensorRT.

Inference time in Table 1 was measured on an industrial PC having Intel Core i7 gen7 and Nvidia Geforce RTX 2080 onboard.

Interest point detection and matching, depth calculation, semantic segmentation and 3D object detection algorithms can be run in parallel. Eventually, we get less than 150ms per pair of frames from left and right cameras, which is about 7 FPS.

5. CONCLUSIONS AND FUTURE WORK

Analyzing the results, we can conclude that using deep neural networks for the SLAM problem and dense stereo semantic reconstruction allows to achieve better results in terms of reliability, accuracy of the solution and scene understanding.

In near future we want to apply our lidar and camera calibration approach, described in (B. Vishnyakov et al., 2020b), to our dense scene reconstruction method. We are quite sure that fusion of lidar data and depth map can improve the precision of calculated distance to obstacles.

Also, we are going to modify the Bi3D architecture to decrease its computational costs. We will use depth loss instead of disparity loss for a dense scene reconstruction model, which may help to solve problems of sharp edges.

ACKNOWLEDGEMENTS

The reported study was funded by RFBR project № 19-07-01248 A.

REFERENCES

A. Badki, A. Troccoli, K. Kim, J. Kautz, P. Sen and O. Gallo, 2020. "Bi3D: Stereo Depth Estimation via Binary Classifications," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 1597-1605.

L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv:1706.05587v3 [cs.CV].

L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. arXiv:1802.02611v3 [cs.CV].

X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, T. Drummond, H. Li, Z. Ge, 2020. Hierarchical Neural Architecture Search for Deep Stereo Matching. arXiv:2010.13501 [cs.CV].

D. DeTone, T. Malisiewicz and A. Rabinovich, 2018. SuperPoint: Self-Supervised Interest Point Detection and Description, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, pp. 337-33712.

K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang and Q. Tian, CenterNet: Keypoint Triplets for Object Detection, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6568-6577.

J. Engel, V. Koltun and D. Cremers, 2018. Direct Sparse Odometry, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 3, pp. 611-625.

C. Forster, M. Pizzoli and D. Scaramuzza, 2014. SVO: Fast semi-direct monocular visual odometry, 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, pp. 15-22.

H. Hirschmuller, 2005. Accurate and efficient stereo processing by semi-global matching and mutual information, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, pp. 807-814 vol. 2.

V. Kolmogorov and R. Zabih, 2001. Computing visual correspondence with occlusions using graph cuts, Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, Vancouver, BC, Canada, pp. 508-515 vol.2.

P. Li, H. Zhao, P. Liu and F. Cao, 2020. Monocular 3D Detection with Geometric Constraints Embedding and Semi-supervised Training. arXiv:2001.03343 [cs.CV]

D. G. Lowe, 2004. Distinctive Image Features from Scale-Invariant
Keypoints, International Journal of Computer Vision, vol.50, No. 2, pp.91-110.

M. Menze and A. Geiger, 2015. Object scene flow for autonomous vehicles, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, pp. 3061-3070.

R. Mur-Artal and J. D. Tardós, 2017. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. IEEE Transactions on Robotics, vol. 33, no. 5, pp. 1255-1262.

T. Qin, P. Li and S. Shen, 2018. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator, in IEEE Transactions on Robotics, vol. 34, no. 4, pp. 1004-1020, Aug. 2018.

E. Rublee, V. Rabaud, K. Konolige and G. Bradski, 2011. ORB: An efficient alternative to SIFT or SURF, 2011 International Conference on Computer Vision, Barcelona, pp. 2564-2571.

P. -E. Sarlin, D. DeTone, T. Malisiewicz and A. Rabinovich, 2020. SuperGlue: Learning Feature Matching With Graph Neural Networks, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 4937-4946.

I. Sgibnev, A. Sorokin, B. Vishnyakov and Y. Vizilter. 2020. Deep semantic segmentation for the off-road autonomous driving, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLIII-B2-2020, 617–622.

J. Sun, N. Zheng and H. -Y. Shum, 2003. Stereo matching using belief propagation, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 7, pp. 787-800.

B. Vishnyakov and V. Sheverdin, 2020a. Real-time SLAM for the off-road autonomous driving, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLIII-B2-2020, 631–635.

B. Vishnyakov, Y. Blokhinov, I. Sgibnev, V. Sheverdin, A. Sorokin, A. Nikanorov, P. Masalov, K. Kazakhmedov, S. Brianskiy, E. Andrienko and Y. Vizilter, 2020b. Semantic scene understanding for the autonomous platform, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLIII-B2-2020, 637–644.

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, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8437-8445.

B. Zoph, Q. V. Le, 2016. Neural Architecture Search with Reinforcement Learning, arXiv:1611.01578 [cs.LG].