НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2024 Том 24 № 1 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2024 Vol. 24 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2024-24-1-118-123

Monocular depth estimation for 2D mapping of simulated environments

Majd Barhoum[1], Anton A. Pyrkin[2]✉

[1,2] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] barhoum.majd213@gmail.com, https://orcid.org/0009-0007-8022-7932
[2] a.pyrkin@gmail.com✉, https://orcid.org/0000-0001-8806-4057

Abstract

This article addresses the problem of constructing maps for 2D simulated environments. An algorithm based on monocular depth estimation is proposed achieving comparable accuracy to methods utilizing expensive sensors such as RGBD cameras and LIDARs. To solve the problem, we employ a multi-stage approach. First, a neural network predicts a relative disparity map from an RGB flow provided by RGBD camera. Using depth measurements from the same camera, two parameters are estimated that connect the relative and absolute displacement maps in the form of a linear regression relation. Based on a simpler RGB camera, by applying a neural network and estimates of scaling parameters, an estimate of the absolute displacement map is formed, which allows to obtain an estimate of the depth map. Thus, a virtual scanner has been designed providing Cartographer SLAM with depth information for environment mapping. The proposed algorithm was evaluated on a ROS 2.0 simulation of a simple mobile robot. It achieves faster depth prediction compared to other depth estimation algorithms. Furthermore, maps generated by our approach demonstrated a high overlap ratio with those obtained using an ideal RGBD camera. The proposed algorithm can find applicability in crucial tasks for mobile robots, like obstacle avoidance, and path planning. Moreover, it can be used to generate accurate cost maps, enhancing safety and adaptability in mobile robot navigation.

Keywords

monocular depth estimation, mapping, linear regression, disparity maps, neural network

Acknowledgements

This paper was supported by the Ministry of Science and Higher Education of the Russian Federation (State Assignement No. 2019-0898).

For citation: Barhoum M., Pyrkin A.A. Monocular depth estimation for 2D mapping of simulated environments. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 1, pp. 118–123. doi: 10.17586/2226-1494-2024-24-1-118-123

УДК 004.896

Использование монокулярной оптики при оценке глубины объектов для двумерного картирования моделируемой среды

Мажд Бархум[1], Антон Александрович Пыркин[2]✉

[1,2] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] barhoum.majd213@gmail.com, https://orcid.org/0009-0007-8022-7932
[2] a.pyrkin@gmail.com✉, https://orcid.org/0000-0001-8806-4057

Аннотация

Введение. Рассмотрена задача построения карты двумерной среды. Предложен алгоритм оценки на основе монокулярной оптики и RGB-изображений. Алгоритм позволяет получать результаты, сопоставимые с подходами на основе дорогостоящих датчиков, таких как RGBD-камеры и лидары. **Метод.** Решение задачи включает нескольких этапов. На начальном этапе выполняется обучение нейронной сети, которая формирует относительную карту несоответствия (смещений) на основе входного потока RGB-изображений от RGBD-камеры. С использованием измерений глубин от той же камеры выполняется оценка двух параметров, связывающих относительную и абсолютную карты смещений в виде линейного регрессионного соотношения. На основе более простой RGB-камеры, путем применения нейронной сети и оценок масштабирующих параметров

формируется оценка абсолютной карты смещений, позволяющей получить оценку карты глубин. Таким образом, синтезирован виртуальный сканер, который формирует данные о глубине для построения карты окружающей среды. **Основные результаты.** Представленный алгоритм апробирован при моделировании движения мобильного робота в среде ROS 2.0. Удалось достичь более быстрого прогнозирования глубины объектов по сравнению с другими алгоритмами оценки глубины. Карты, сгенерированные согласно разработанному алгоритму, продемонстрировали высокую степень совпадения с картами, полученными с помощью идеальной RGBD-камеры. **Обсуждение.** Предложенный алгоритм может найти применение в ключевых задачах управления мобильными роботами, такими как избегание препятствий и планирование пути. Алгоритм может быть использован при разметке карт по областям с различной степенью сложности прохождения, повышая безопасность и адаптивность навигации мобильных роботов.

Ключевые слова
монокулярная оценка глубины, картографирование, линейная регрессия, карты несоответствия (смещений), нейронные сети

Благодарности
Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации (госзадание 2019-0898).

Ссылка для цитирования: Бархум М., Пыркин А.А. Использование монокулярной оптики при оценке глубины объектов для двумерного картирования моделируемой среды // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 1. С. 118–123 (на англ. яз.). doi: 10.17586/2226-1494-2024-24-1-118-123

Introduction

In the realm of robotics, an accurate map plays a pivotal role in the robot's interaction with the surrounding world, granting the robot the ability to navigate seamlessly through obstacle filled environments. To that end, Simultaneous Localization and Mapping (SLAM) algorithms are used to enable robots to keep track of their own position and the position of all obstacles around them utilizing the robot's on-board sensors. Visual SLAM algorithms build 3D maps of the environment using multiple cameras or specialized depth sensors. However, considering compact robotic systems used in cost-constrained applications such algorithms may not be practical. On the other hand, Monocular visual SLAM algorithms rely on a single camera to build 3D maps of the environment. Nevertheless, they suffer from scale ambiguity since only a single image with little to no depth information is used for both mapping and localization. 2D SLAM algorithms require less computational demand making them a convenient alternative to visual SLAM, provided a 2D map is sufficient for the problem at hand. But they still require depth information in the form of laser scans to build the surrounding map. Monocular Depth Estimation (MDE) offers the means to retrieve depth information without the dependence on sensors like LIDARS, or RGBD cameras. In the heart of it, the problem of MDE arises from the lack of depth cues in single images rendering conventional estimation approaches useless, but due to the importance of the problem recent advances showed the formidability of neural networks in predicting near perfect depth maps. The problem of MDE can be broken down into two separate sub-problems, namely relative depth estimation and metric depth estimation, the most common segment is the metric one, where the neural network trains to predict a depth map in metric space for situations similar to the datasets it trained on [1–5]. The other segment addresses the problem from a different angle, where instead of direct depth maps, disparity being the inverse depth is predicted, and the disparity of each pixel is only consistent relative to each other. This is achieved by training on a mixture of datasets with different camera models to increase the generalization of the model [6–8].

Integrating MDE with SLAM is an attempt to meet the surging demand for autonomous systems. That combines the power of artificial intelligence and conventional bundle adjustment techniques to improve the visual mapping process. The authors of [9] implemented a pseudo-RGBD SLAM system which used a Convolutional neural network to predict depth maps, and those depth maps are used to improve the accuracy of monocular SLAM. While the authors of [10] implement an unsupervised MDE pipeline that improves the accuracy of monocular SLAM and provides a fail-proof method for when ORB-SLAM does not match enough features, and significantly decreases the initialization time of the SLAM system. In this work, we tackle the integration of MDE as a sensor for 2D mapping using Cartographer SLAM, aiming at the balance between real-time mapping and accuracy we propose a simple yet effective algorithm to map simulated environments.

Monocular Depth Estimation

When choosing a neural network to perform the task of MDE, one needs to take into account the delicate balance between speed and accuracy by testing multiple networks and evaluating the trade-off between frames per second and performance, we reached the conclusion of working with a network with a low parameter count and good performance, even if the accuracy is lower than its peers. In our algorithm we chose one of the networks proposed by the authors of [6] where they trained multiple networks on a mixture of datasets with different input-output resolutions and showed how the improvement varies depending on the model trained and input image resolution. The **Swin2-T** network they trained can promise near real-time operation with acceptable accuracy provided a GPU with adequate capabilities is used when running the network. In their work, the architecture of the network has a smaller image size input 256 × 256 making the network more efficient to deploy. The architecture uses the pre-trained **Swin2-T** as an encoder where the transformer progresses through four

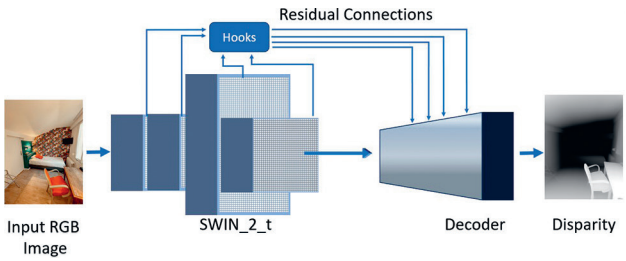

Fig. 1. Architecture of the network with Swin2-T transformer encoder and a decoder that uses the residual connection coming from the encoder stages to decode depth

stages each of which has a number **Swin2** blocks, the total number of these blocks is 12, divided as 2 in each stage except in the third one where there are 6 of them. The output of each stage is passed to the next stage and also hooked to the decoder which consists of 4 convolutional layers to map the encoded data into a depth map similar to input in size except that it has only one channel. The architecture is depicted in Fig. 1.

Relative to Metric Depth Maps

In their work, the authors of [8] propose a loss function to facilitate training on a mix of multiple datasets in disparity space, instead of the usual depth ground truth in online datasets. The loss function used to produce a prediction of relative disparity which is scale and shift-invariant, where, if $\mathbf{d}$ is the predicted disparity and $\mathbf{d}^*$ is the ground truth disparity, they define:

$$\hat{\mathbf{d}} = s\mathbf{d} + t,\ \hat{\mathbf{d}}^* = \mathbf{d}^*, \tag{1}$$

where, s, t from (1) can be determined as a least-squares problem for each pixel i from M as:

$$(s, t) = \text{argmin}_{s,t} \sum_{i=1}^{M} (s\mathbf{d}_i + t - \mathbf{d}_i^*)^2. \tag{2}$$

And the loss for one image using $\hat{\mathbf{d}}$, $\mathbf{d}^*$ from (1) becomes:

$$\mathcal{L}(\hat{\mathbf{d}}, \hat{\mathbf{d}}^*) = \frac{1}{2M} \sum_{i=1}^{M} \text{mae}(\hat{\mathbf{d}}_i - \hat{\mathbf{d}}_i^*). \tag{3}$$

They add to it the gradient loss to capture edges of objects in the image with the term proposed by [11] as:

$$\mathcal{L}_{reg}(\hat{\mathbf{d}}, \hat{\mathbf{d}}^*) = \frac{1}{M} \sum_{k=1}^{K} \sum_{i=1}^{M} (|\nabla_x R_i^k| + |\nabla_y R_i^k|, \tag{4}$$

where k refers to the index of scale level (they use 4), ∇_y, ∇_x are the gradients on x-axis and y-axis respectively, and $R_i = \hat{\mathbf{d}}_i - \mathbf{d}_i^*$. To adapt these losses to multiple datasets, they combine (3) and (4) to get the final loss function applied to the experiments.

$$\mathcal{L}_l = \frac{1}{N_l} \sum_{n=1}^{N_l} \mathcal{L}(\hat{\mathbf{d}}^n, \hat{\mathbf{d}}^{*n}) + \alpha \mathcal{L}_{reg}(\hat{\mathbf{d}}^n, \hat{\mathbf{d}}^{*n}),$$

where N_l is the size of the mixture of datasets in the training experiment, and α is 0.5.

Notation. Let vectors $\hat{\mathbf{D}} = col(\widehat{d_1}, \widehat{d_2}, \ldots, \widehat{d_M})$ be the predicted disparity map by the network, and $\mathbf{D} = col(d_1, d_2, \ldots, d_M)$ be the inverse of the ground truth depth obtained from the RGBD camera (images reshaped as vectors of the same length), $b \in \mathbb{R}$, $W \in \mathbb{R}$ is the shift between maps, and the scale respectively as an argument of (2), then we can define $W = \bar{s}$, $b = \bar{t}$ as average values of the scale and shift in our simulation, $\mathbf{E}$ is a vector column of ones of length M.

Retrieving the scale and shift of the disparity map can align it with the metric one as:

$$\mathbf{D} = [\hat{\mathbf{D}}\ \mathbf{E}] \begin{bmatrix} W \\ b \end{bmatrix}. \tag{5}$$

Equation (5) can be treated as a least squares problem then, there exist optimal values for W, b as:

$$[\hat{W}, \hat{b}]^T = \left([\hat{\mathbf{D}}\ \mathbf{E}]^T [\hat{\mathbf{D}}\ \mathbf{E}]\right)^{-1} \left([\hat{\mathbf{D}}\ \mathbf{E}]^T \mathbf{D}\right). \tag{6}$$

Then we reason that we can solve this problem using a linear regression model minimizing the square errors as:

$$L = \sum_{i=1}^{M} \left(\mathbf{D_i} - [\hat{\mathbf{D}}_\mathbf{i}\ \mathbf{E_i}] \begin{bmatrix} \hat{\mathbf{W}} \\ \hat{\mathbf{b}} \end{bmatrix} \right)^2.$$

Leaving the metric predicted depth to be denoted using the optimal values from (6) as:

$$\mathbf{D}^* = \frac{1}{(\hat{W}\hat{\mathbf{D}} + \hat{b})}.$$

Pseudo Laser Scanner

From the pinhole camera model with no distortion, the invertible projection equation can be written as follows:

$$\begin{bmatrix} u \\ v \\ 1 \\ 1/z \end{bmatrix} = \frac{1}{z} \mathbf{K}[\mathbf{R}|\mathbf{t}] \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \tag{7}$$

where u, v are the pixel indices in the image, x, y, z are the 3D coordinates of the point, and $\mathbf{K}[\mathbf{R}|\mathbf{t}]$ are the intrinsic and extrinsic parameters of the camera. Using equation (7) and the estimated depth, we can re-project the pixels in the image onto a 3D point cloud, but for 2D mapping only one row of the image is sufficient, thus we re-project all the pixels of the middle row of the image onto a 3D line; that mimics the work of a laser scanner as was done by the authors of [12]. We also introduce a limit on the range of the point cloud re-projected, since we compromised accuracy for speed when choosing the network. Not to mention that a smaller range corresponding to a smaller search window can prove to be easier to handle when mapping.

$$z = \begin{cases} \mathbf{D}^*, \textit{if } 0 < D^* \leq 1.5 \\ \Phi, \textit{otherwise} \end{cases}.$$

Cartographer SLAM

Cartographer is a popular SLAM system designed to create highly accurate 2D and 3D maps of environments

while localizing a robot within those maps. Cartographer SLAM uses the concept of drawing multiple sub-maps and then stitching them together to form the entire map of the environment, it draws the map as a grid and estimates the pose of the robot using a graph optimization algorithm called the CERES scan matching. The main Steps in this Graph-based SLAM system are the following:

1. Transform points from the scan frame which is the sensor frame into the sub-map frame.

$$T_{\xi P} = \begin{Bmatrix} \cos(\xi_\theta) & -\sin(\xi_\theta) \\ \sin(\xi_\theta) & \cos(\xi_\theta) \end{Bmatrix} P + \begin{Bmatrix} \xi_x \\ \xi_y \end{Bmatrix},$$

   where the pose is described by the position x, y and the orientation θ.
2. Then a hit or miss VOXEL probabilistic filter is used to find the probability of a cell in the grid of the sub-map being occupied or not, based on the observed points using the laser scan. If a point was already observed its odds $odds(p) = \frac{p}{1-p}$ are updated as:

$$M = odds^{-1}(odds(M_{old})odds(p_{hit})).$$

3. CERES scan matching: before adding a scan to a sub-map, the pose of the scan, denoted as ξ, is optimized in relation to the current local sub-map using a scan matcher based on the Ceres library. The scan matcher finds the scan pose that maximizes the probabilities associated with the scan points within the sub-map.

$$\underset{\xi}{\operatorname{argmin}} \sum_{k=1}^{K} (1 - M_{smooth}(T_\xi h_k))^2.$$

4. Loop closure: it is formulated as a nonlinear least squares problem that allows adding residuals to take additional data into account. Every few seconds the CERES scan matching is applied to compute the nonlinear optimization problem.

$$\underset{\Xi^m, \Xi^s}{\arg\min}\ 1/2 \sum_{i,j} \rho(E^2(\xi_i^m, \xi_j^s; \sum, \xi_{ji})),$$

   where the sub-map poses Ξ^m and the scan poses Ξ^s are optimized according to constraints [13].

Results and Discussion

Experiment. The simulation features a simple differential drive robot. The simulation process passes two stages as depicted in Fig. 2. First, the linear regression model is trained on relative depth estimated as an input, and the target will be the ground truth depth from an RGBD camera. When triggered (after linear regression finishes training), the second stage starts, the robot is loaded into a new environment where the trained linear regression model rescales and shifts the relative disparity transforming it to a metric one which is then inversed to achieve metric depth. From this metric depth only the middle row is projected as a 3D point cloud forming a pseudo laser scanner that is used as an input to Cartographer SLAM.

Metric Depth Results. We first present the performance of the aligned metric depth of our algorithm. In the process of evaluating the MDE performance, the testing prediction and ground truth are masked according to cropping criteria corresponding to a specific region of interest in the images of the testing dataset, like the Eigen-crop proposed by [14], or the Garg-crop proposed by [15]. These criteria are typically used to ensure that the evaluation metrics reflect the performance of the model in areas that are most relevant to the task. In our work, the alignment procedure aims to render pre-trained work applicable in an environment of choice; hence, we evaluate the entire image (no cropping) pursuing a better comprehensive assessment of the performance across the entire field of view as shown in Fig. 3. MDE is treated as a regression problem and is evaluated using the following metrics, where M is the number of pixels d_i, d_i^*, are ground truth and predicted depth:

— Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{M} \sum_{i}^{M} |d_i - d_i^*|^2}.$$

— Mean Relative Error (REL):

$$REL = \frac{1}{M} \sum_{i}^{M} \frac{|d_i - d_i^*|}{d_i}.$$

— Threshold accuracy (δ): ratio of predicted pixels having relative errors within a threshold.

$$\max\left(\frac{d^*}{d}, \frac{d}{d^*}\right) < \delta_i,\ \delta_i = 1.25^i.$$

We compare the metric depth performance with the ground truth by evaluating multiple metrics, and the results are shown in Table.

Better performance is highlighted and made bold. Note that the maximum depth in our testing environment is 22 m.

It is worth mentioning that these results are lacking in terms of accuracy, but when projecting to 3D point clouds, we can see that the nearest object seen by the camera is aligned almost perfectly with the ground truth. We figure that limiting the range of the projected depth on the

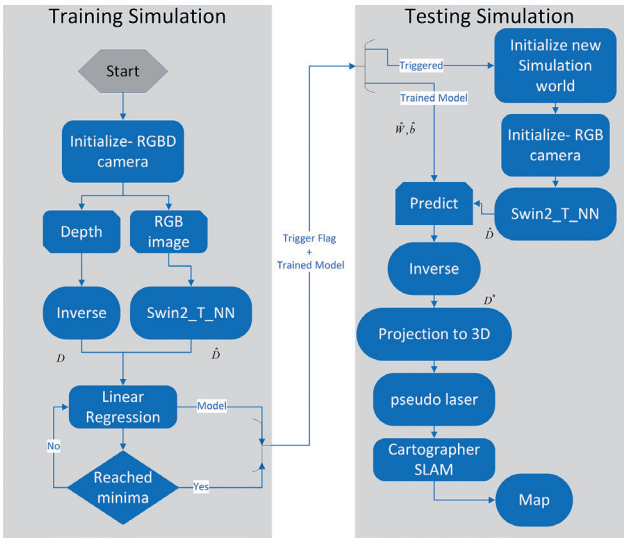

Fig. 2. Simulation algorithm

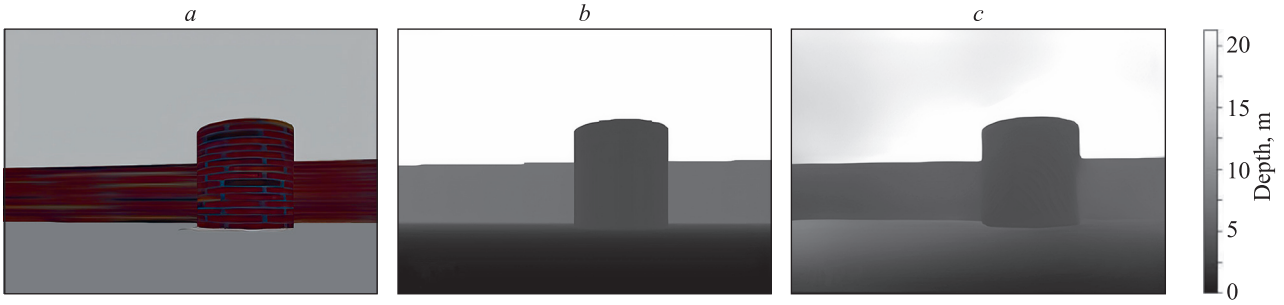

Fig. 3. Input RGB image (*a*), ground truth depth (*b*), aligned predicted depth (*c*), maximum width and height seen in the scene is 12 × 6.5 m

Table. Metric Depth results. Downward arrows ↓ mean the lower the value the better, upward ↑ mean higher is better

MDE algorithm	δ_1↑	δ_2↑	δ_3↑	*REL*↓, m	*RMSE*↓, m	Image dimension, pixel	Frames per second, Hz
Ada_bins [1]	0.438	0.588	0.760	1.158	3.886	640 × 480	8.5
Proposed algorithm	0.454	0.499	0.587	1.759	4.452	256 × 256	28.2

synthetic 2D laser scanner proves to be useful in solving depth accuracy problems.

Mapping. In this section, we present the results of mapping using our algorithm in comparison to the map generated by using the RGBD camera. The objective is to assess the accuracy of our approach in simulating the performance of real-world RGBD sensors. To that end, we run two consecutive simulations in the same world where in the first one we drive the robot equipped with the RGBD camera to map the environment, saving the trajectory the robot traversed while mapping as a file containing the action commands (linear, and angular velocities). When done, the map is saved as an image. In the second simulation we let the robot equipped with only an RGB camera use the same saved action commands to traverse the environment again following the same trajectory and mapping using our approach, since Cartographer SLAM is a pose graph optimization algorithm; making the robot pass through the same poses will produce similar grid maps that can be compared. This is done to ensure that acquired maps are aligned and variations between sub-maps are minimal. It is worth mentioning that the robot starts from the same position in the two simulations (near the bottom right wall), the trajectory of choice followed by the robot is a non-complete clockwise lap along the wall. When it reaches the center cylinders, it makes full rotation around its *z*-axis.

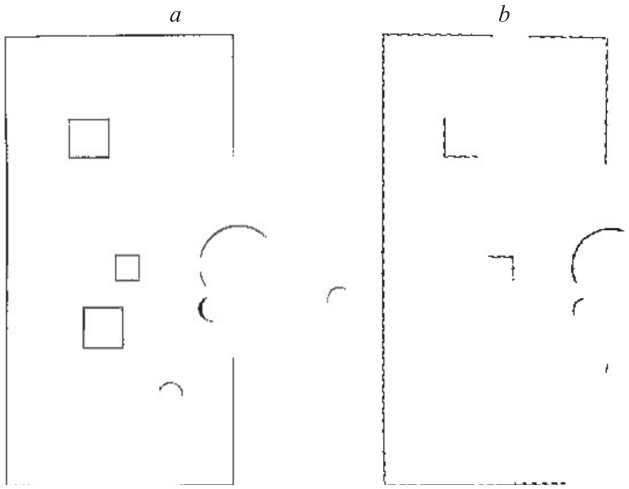

Fig. 4. Mapping results using the ideal RGBD camera (*a*), mapping using our algorithm (*b*)

To evaluate, we first address the accuracy of the mapping in terms of spatial fidelity and spatial consistency where we examine the alignment of key landmarks. To that end, we calculate the overlap between the map generated using our approach and the map saved using the RGBD which amounts to 88.17 %, indicating a high degree of correspondence between the two maps. We address obstacle recognition where it can be seen from Fig. 4 that in certain areas our algorithm fails to recognize the entirety of the object. This is due to the uncertainty in mapping generated by Frame-to-Frame variability since the network is designed to predict the depth in each frame. In some cases, the prediction may exhibit small variations due to the network sensitivity to visual changes and the rabid change between consecutive frames.

Conclusion

In conclusion, our approach showcases significant promise in mapping simulated environments. In terms of metric depth estimation, it falls short when compared to other algorithms trained to directly produce metric depth on certain datasets, which is reasonable since we traded the accuracy of the depth map with a faster inference rate. However, this low-accuracy depth map does not hinder the algorithm where in terms of mapping simulated environments, the high map overlap presented in the experimental results reinforces its effectiveness and potential for various robotics and simulation applications, emphasizing the balance between accuracy and real-time performance. It is paramount moving forward to boost the algorithm ability to handle dynamic scenes as well

as enhance depth estimation and improve its accuracy. Conducting real-world testing and further optimizing, the algorithm are also venues that should be addressed in the future in an effort to pave the way for robust and accurate mapping in resource-constrained platforms.

References

1. Bhat S.F., Alhashim I., Wonka P. AdaBins: Depth estimation using adaptive bins. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4009–4018. https://doi.org/10.1109/cvpr46437.2021.00400
2. Li Z., Wang X., Liu X., Jiang J. BinsFormer: Revisiting adaptive bins for monocular depth estimation. *arXiv*, 2022, arXiv:2204.00987. https://doi.org/10.48550/arXiv.2204.00987
3. Zhang S., Yang L., Mi M.B., Zheng X., Yao A. Improving deep regression with ordinal entropy. *arXiv*, 2023, arXiv:2301.08915. https://doi.org/10.48550/arXiv.2301.08915
4. Ranftl R., Bochkovskiy A., Koltun V. Vision transformers for dense prediction. *Proc. of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12179–12188. https://doi.org/10.1109/iccv48922.2021.0119
5. Xie Z., Geng Z., Hu J., Zhang Z., Hu H., Cao Y. Revealing the dark secrets of masked image modeling. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14475–14485. https://doi.org/10.1109/cvpr52729.2023.01391
6. Birkl R., Wofk D., Müller M. MiDaS v3.1 – A Model zoo for robust monocular relative depth estimation. *arXiv*, 2023, arXiv:2307.14460. https://doi.org/10.48550/arXiv.2307.14460
7. Bhat S.F., Birkl R., Wofk D., Wonka P., Müller M. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv*, 2023, arXiv:2302.12288. https://doi.org/10.48550/arXiv.2302.12288
8. Ranftl R., Lasinger K., Hafner D., Schindler K., Koltun V. Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, vol. 44, no. 3, pp. 1623–1637. https://doi.org/10.1109/tpami.2020.3019967
9. Xing X., Cai Y., Lu T., Yang Y., Wen D. Joint self-supervised monocular depth estimation and SLAM. *Proc. of the 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 4030–4036. https://doi.org/10.1109/icpr56361.2022.9956576
10. Geng M., Shang S., Ding B., Wang H., Zhang P. Unsupervised learning-based depth estimation-aided visual slam approach. *Circuits, Systems, and Signal Processing*, 2020, vol. 39, pp. 543–570. https://doi.org/10.1007/s00034-019-01173-3
11. Li Z., Snavely N. MegaDepth: Learning single-view depth prediction from internet photos. *Proc. of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2041–2050. https://doi.org/10.1109/cvpr.2018.00218
12. Tran M., Ly N. Mobile robot planner with low-cost cameras using deep reinforcement learning. *Proc. of the 7th NAFOSTED Conference on Information and Computer Science (NICS)*, 2020, pp. 54–59. https://doi.org/10.1109/nics51282.2020.9335852
13. Hess W., Kohler D., Rapp H., Andor D. Real-time loop closure in 2D LIDAR SLAM. *Proc. of the 2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 1271–1278. https://doi.org/10.1109/icra.2016.7487258
14. Eigen D., Puhrsch C., Fergus R. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, 2014, vol. 27, pp. 2, 5, 6.
15. Garg R., Kumar B.G.V., Carneiro G., Reid I. Unsupervised CNN for single view depth estimation: Geometry to the rescue. *Lecture Notes in Computer Science*, 2016, vol. 9912, pp. 740–756. https://doi.org/10.1007/978-3-319-46484-8_45

Литература

1. Bhat S.F., Alhashim I., Wonka P. AdaBins: Depth estimation using adaptive bins // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. P. 4009–4018. https://doi.org/10.1109/cvpr46437.2021.00400
2. Li Z., Wang X., Liu X., Jiang J. BinsFormer: Revisiting adaptive bins for monocular depth estimation // arXiv. 2022. arXiv:2204.00987. https://doi.org/10.48550/arXiv.2204.00987
3. Zhang S., Yang L., Mi M.B., Zheng X., Yao A. Improving deep regression with ordinal entropy // arXiv. 2023. arXiv:2301.08915. https://doi.org/10.48550/arXiv.2301.08915
4. Ranftl R., Bochkovskiy A., Koltun V. Vision transformers for dense prediction // Proc. of the IEEE/CVF International Conference on Computer Vision. 2021. P. 12179–12188. https://doi.org/10.1109/iccv48922.2021.01196
5. Xie Z., Geng Z., Hu J., Zhang Z., Hu H., Cao Y. Revealing the dark secrets of masked image modeling // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. P. 14475–14485. https://doi.org/10.1109/cvpr52729.2023.01391
6. Birkl R., Wofk D., Müller M. MiDaS v3.1 – A Model zoo for robust monocular relative depth estimation // arXiv. 2023. arXiv:2307.14460. https://doi.org/10.48550/arXiv.2307.14460
7. Bhat S.F., Birkl R., Wofk D., Wonka P., Müller M. ZoeDepth: Zero-shot transfer by combining relative and metric depth // arXiv. 2023. arXiv:2302.12288. https://doi.org/10.48550/arXiv.2302.12288
8. Ranftl R., Lasinger K., Hafner D., Schindler K., Koltun V. Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022. V. 44. N 3. P. 1623–1637. https://doi.org/10.1109/tpami.2020.3019967
9. Xing X., Cai Y., Lu T., Yang Y., Wen D. Joint self-supervised monocular depth estimation and SLAM // Proc. of the 26th International Conference on Pattern Recognition (ICPR). 2022. P. 4030–4036. https://doi.org/10.1109/icpr56361.2022.9956576
10. Geng M., Shang S., Ding B., Wang H., Zhang P. Unsupervised learning-based depth estimation-aided visual slam approach // Circuits, Systems, and Signal Processing. 2020. V. 39. P. 543–570. https://doi.org/10.1007/s00034-019-01173-3
11. Li Z., Snavely N. MegaDepth: Learning single-view depth prediction from internet photos // Proc. of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. P. 2041–2050. https://doi.org/10.1109/cvpr.2018.00218
12. Tran M., Ly N. Mobile robot planner with low-cost cameras using deep reinforcement learning // Proc. of the 7th NAFOSTED Conference on Information and Computer Science (NICS). 2020. P. 54–59. https://doi.org/10.1109/nics51282.2020.9335852
13. Hess W., Kohler D., Rapp H., Andor D. Real-time loop closure in 2D LIDAR SLAM // Proc. of the 2016 IEEE International Conference on Robotics and Automation (ICRA). 2016. P. 1271–1278. https://doi.org/10.1109/icra.2016.7487258
14. Eigen D., Puhrsch C., Fergus R. Depth map prediction from a single image using a multi-scale deep network // Advances in Neural Information Processing Systems. 2014. V. 27. P. 2, 5, 6.
15. Garg R., Kumar B.G.V., Carneiro G., Reid I. Unsupervised CNN for single view depth estimation: Geometry to the rescue // Lecture Notes in Computer Science. 2016. V. 9912. P. 740–756. https://doi.org/10.1007/978-3-319-46484-8_45

Authors

Majd Barhoum — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0007-8022-7932, barhoum.majd213@gmail.com

Anton A. Pyrkin — D.Sc., Full Professor, ITMO University, Saint Petersburg, 197101, Russian Federation, sc 26656070700, https://orcid.org/0000-0001-8806-4057, a.pyrkin@gmail.com

Авторы

Бархум Мажд — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0007-8022-7932, barhoum.majd213@gmail.com

Пыркин Антон Александрович — доктор технических наук, профессор, профессор, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, sc 26656070700, https://orcid.org/0000-0001-8806-4057, a.pyrkin@gmail.com

Received 09.11.2023
Approved after reviewing 18.12.2023
Accepted 20.01.2024

Статья поступила в редакцию 09.11.2023
Одобрена после рецензирования 18.12.2023
Принята к печати 20.01.2024.