Proposal on Monocular Depth Estimation Based on Surgical Video

Member Name:
Mingqian Liao, Kehan Chen
Member SID:
3038745426, 3037507071

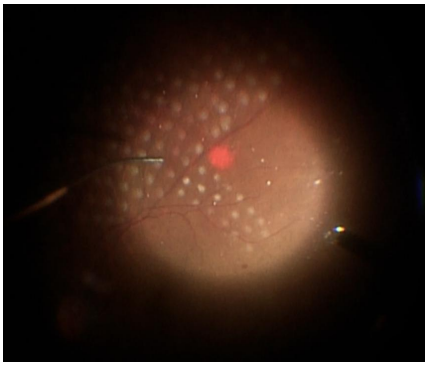
Background

Nowadays, depth estimation is increasingly used, whether in autonomous driving[1][2][2] or 3D scene reconstruction. In the surgical field, the scene where the surgery is performed is inside the eye, and the doctor can only obtain the surgical image through a microscope. The information provided by the microscope image is two-dimensional information, and it is not possible to obtain the depth information in it, i.e., three-dimensional information. In addition, the microscope provides a limited field of view and lacks background control to determine depth. Based on the information obtained from the microscope view alone, it is difficult for the surgeon to determine the relative position of the surgical instruments to the internal tissues of the eye, often only through experience, without having an accurate depth estimate. Therefore, we intend to investigate the topic of monocular video depth estimation in microscopic scenarios.

Dataset

The dataset we intend to use is divided into three parts, first is the **real fundus surgery dataset** from the iMED lab, which will be used for the final test.

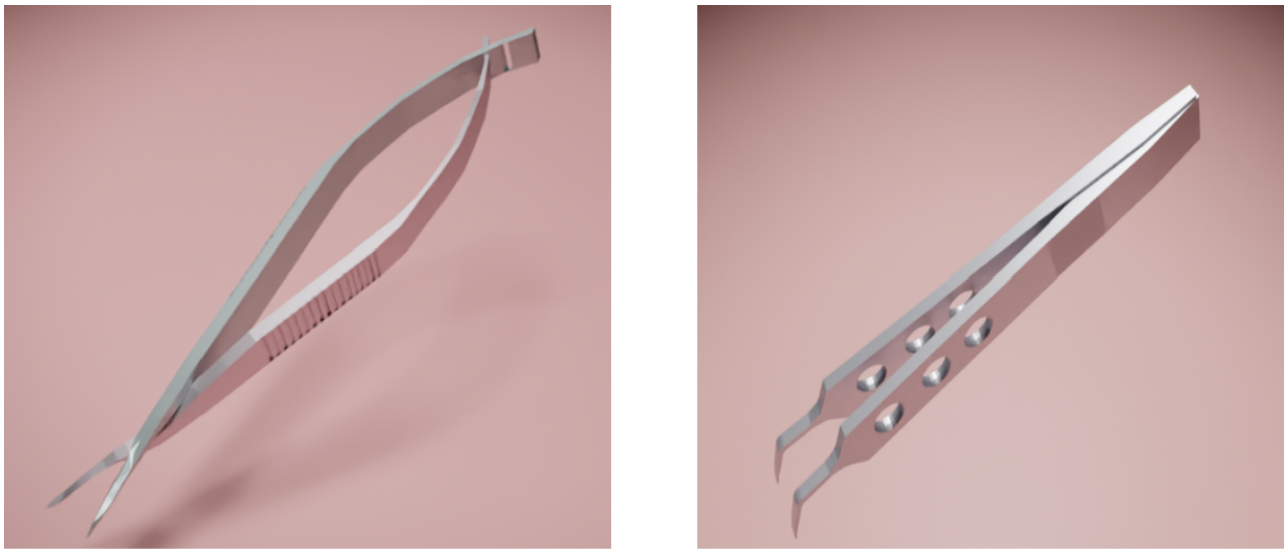
But because of the small number of this dataset and the difficulty of obtaining it, we will also use our own **simulated surgery dataset**, which was taken using a surgical microscope and a simulated fundus in the lab. In order to test the final results of our model, we need a set of metrics to measure the accuracy of our depth estimation results in addition to the visual effects, so we will also use our **virtual dataset** created by reproducing the real surgical scene with



Real Dataset

Simulated Dataset

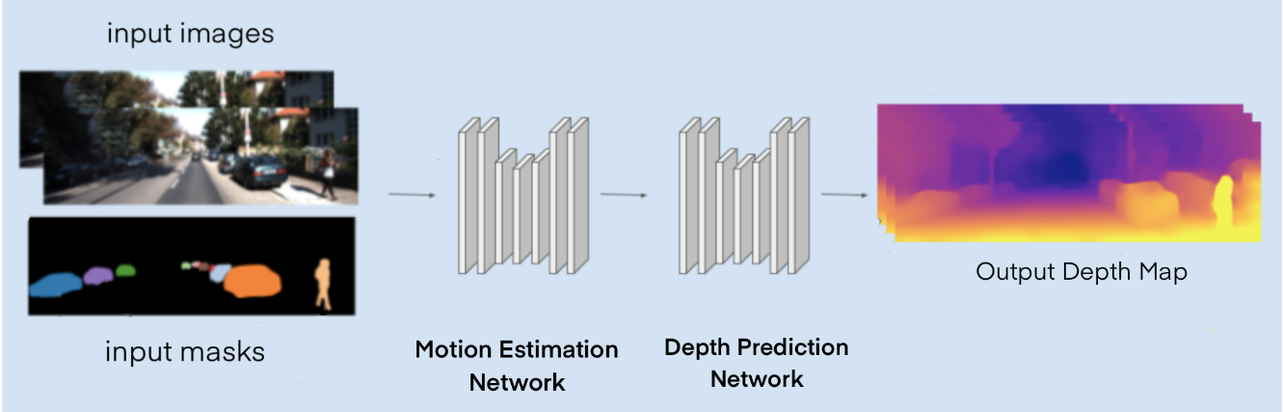
the software blender,
which contains the relative depth information of the instruments and the fundus, which can be used as an evaluation metric for the model results This can be used as an evaluation metric for the model results.



virtual dataset

Method

By referring to the paper Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras[4], we intend to use the motion estimation network as well as the depth estimation network for depth estimation, and we intend to improve the model by pre-processing the data and changing the loss function of the model to better fit our scenario.



Depth in the Wild

Criteria

In addition to visuals that better represent the relative depth relationship between the device and the eye, by reading papers related to depth estimation, we will use the following metrics to evaluate our model[5]:
Absolute Error, we think that it is successful if the error is less than 1;
Square Error, we think that it is successful if the error is less than 50;
Root Mean Square Error(RMSE), we think that it is successful if the error is less than 100.

Preliminary work

We have initially created some virtual datasets and have also acquired some simulated datasets as well as a small number of real surgical scene datasets.
We have also explored some modeling methods for monocular depth estimation to be used in our scenario.

Reference

[1]Li, H., Gordon, A., Zhao, H., Casser, V., & Angelova, A. (2020). Unsupervised monocular depth learning in dynamic scenes. arXiv preprint arXiv:2010.16404.
[2]Perone, C. S., Chapiro, A., & Cohen-Or, D. (2021). M4Depth: Monocular depth estimation for autonomous vehicles in unseen environments. arXiv preprint arXiv:2105.09847v3.
[3]Liang, H., Li, Z., Yang, Y., & Wang, N. (2023). Distilled Multiscale Context Aggregation for Efficient Monocular Depth Estimation. arXiv preprint arXiv:2301.03178.
[4]Gordon, A., Li, H., Jonschkowski, R., & Angelova, A. (2019). Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In

Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 8977-8986).
[5]Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3828-3838).