Real-Time Coarse-To-Fine Depth Estimation on Stereo Endoscopic Images with Self-Supervised Learning

H. YANG and L.A. KAHRS
University of Toronto Mississauga, Ontario, Canada

IEEE - ISBI 2021
International Symposium on Biomedical Imaging

INTRODUCTION

Fast and accurate depth estimation is an essential task in computer-assisted surgery and robotics, especially for endoscopic and microscopic procedures. We propose a real-time stereo matching model using a staged, coarse-to-fine architecture to estimate disparity from medical stereo camera data with self-supervised learning. Our model processes images with a resolution of 1280x1024 pixels beyond 60 fps, with similar accuracy to the semi-global matching algorithm, and does not require any ground truth depth for training. We evaluated our model on two stereo endoscopic datasets from the literature. A mean absolute error below 1.5 mm and root mean square error below 1.9 mm were identified.

RESEARCH QUESTION

Can we train a depth estimation model for stereo endoscopic images that perform in real-time without using ground truth depth data?

METHOD

Feature Extractor. The feature extractor is designed in a staged, pyramidal architecture to get multi-scaled feature. In particular, our model adopts the U-Net [4] architecture as a feature extractor with shared weight.

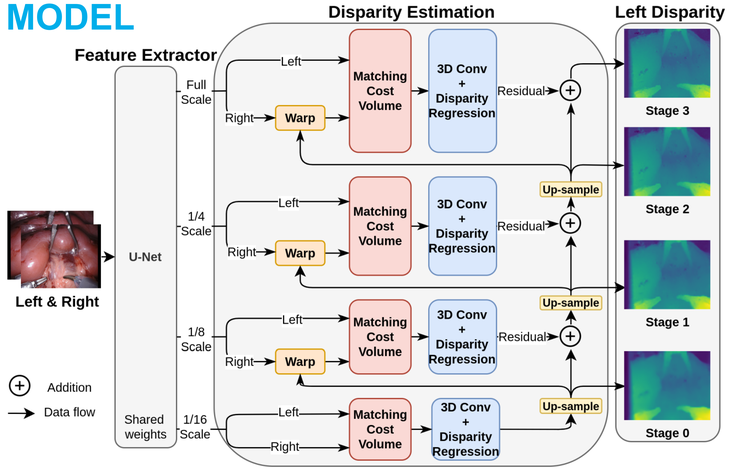
Matching Cost Volume. We use extracted feature maps from the left and right images to build a distance-based matching cost volume.

Regularization & Disparity Regression. The 3D convolution incorporates the context in the initial cost volume and learns to filter and refine the cost volume. Finally, we employ the disparity regression through a differentiable soft *argmin* operation.

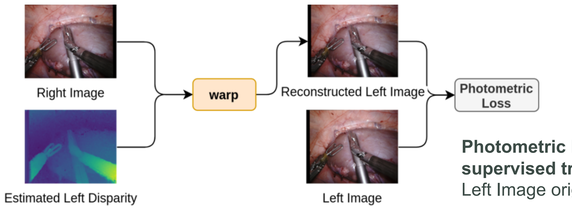
Residual Disparity. For Stage 1, 2, and 3, instead of predicting full disparity, the model corrects the coarse disparity output from the previous stage by estimating residual disparity. This allows us to build a relatively smaller cost volume and use fewer filters in the 3D convolutions, which effectively reduces the computational complexity.

Self-Supervised Objective. The main idea of self-supervised learning for estimating disparity is that, given a pair of rectified stereo images, if the predicted left disparity is correct, then we can generate the reconstructed left image, by sampling the pixel from the right image according to the left disparity. The discrepancy between the original left image and the reconstructed left image can supervises the network to learn to predict an accurate left disparity.

MODEL

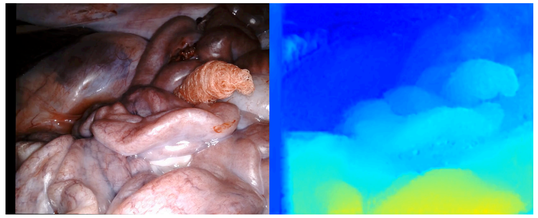


The architecture of our staged model. The input for both training and test phases is a pair of rectified stereo images. The outputs are four left disparity estimations. The input images are from the 2018 Robotic Scene Segmentation (RSS) dataset [3].



Photometric loss for self-supervised training. Right and Left Image originate from [3].

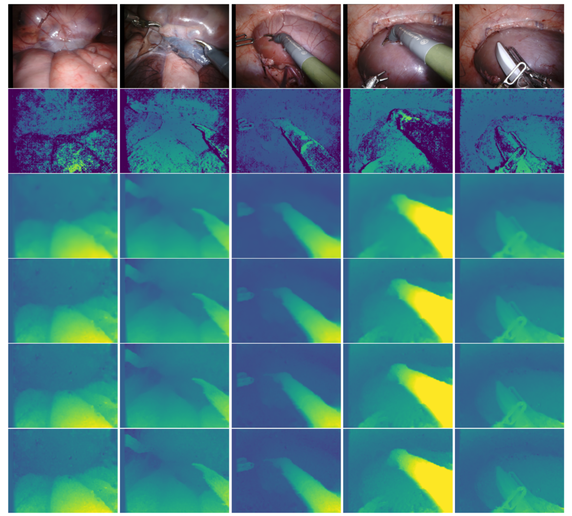
RESULT



Qualitative result on SCARED dataset [6]. Left is the left camera view, right is the disparity. No images from this dataset are used in training.

Methods	Heart 2 (Frame 2) MAE (mm)	Heart 2 (Frame 2) RMSE (mm)	2018 RSS Low Res. (s)	2018 RSS High Res. (s)
SGM	1.81	2.34	**0.0013**	0.0150
Stage 0	1.82	2.19	0.0085	**0.0088**
Stage 1	1.98	2.38	0.0105	0.0104
Stage 2	1.51	**1.88**	0.0122	0.0130
Stage 3	**1.49**	1.94	0.0146	0.0146
Stage 3*	2.28	3.02	0.0146	0.0146

Quantitative comparison for the investigated datasets [3][5]. Comparison for our model and SGM. Best result in each category are in **bold**. ∗ indicates the model trained with high-resolution (1280 × 1024 pixels) data.



Five qualitative comparisons on 2018 RSS dataset [3]. From top to bottom: Input images, SGM, output from Stage 0, 1, 2, and 3.

Exemplary camera image and resulting depth predictions. Left to right: The left input image from RSS [3], output of our model Stage 0 and Stage 3 (red RoI).

CONCLUSIONS

This paper presents our stereo matching model with self-supervised learning for estimating disparity from endoscopic stereo image pairs. Our model uses the coarse-to-fine architecture from AnyNet [2] to enable real-time performance, especially for images with high resolution. No ground truth depth or human labeling is needed for training. We conducted experiments and showed our model can outperform the SGM [1] algorithm in single frames of the silicone heart phantom dataset in terms of accuracy.

ACKNOWLEDGEMENTS

The work was supported by a UTM Undergraduate Research Grant for Haotian Yang. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), RGPIN-2020-05833.

CONTACT INFORMATION

Haotian Yang: haotian.yang@mail.utoronto.ca
Lueder A. Kahrs: lakahrs@cs.toronto.edu

REFERENCES

[1] H. Hirschmuller, "Stereo Processing by Semiglobal Matching and Mutual Information," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 2, pp. 328–341, Feb. 2008.
[2] Y. Wang et al., "Anytime Stereo Image Depth Estimation on Mobile Devices," in 2019 International Conference on Robotics and Automation, May 2019, pp. 5893–5900.
[3] M. Allan et al., "2018 Robotic Scene Segmentation Challenge," arXiv:2001.11190, Aug. 2020.
[4] O. Ronneberger et al., "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Medical Image Computing and Computer-Assisted Intervention 2015, pp. 234–241.
[5] D. Stoyanov et al., "Real-Time Stereo Reconstruction in Robotically Assisted Minimally Invasive Surgery," in Medical Image Computing and Computer-Assisted Intervention 2010, pp. 275–282.
[6] M. Allan et al., "Stereo Correspondence and Reconstruction of Endoscopic Data Challenge," arXiv:2101.01133. Jan 2021.