Self-Supervised Depth Estimation in Laparoscopic Image using 3D Geometric Consistency

Baoru Huang[1,2], Jian-Qing Zheng[3,4], Anh Nguyen[5], Chi Xu[1,2], Ioannis Gkouzionis[1,2], Kunal Vyas[6], David Tuch[6], Stamatia Giannarou[1,2], and Daniel S. Elson[1,2]

[1] The Hamlyn Centre for Robotic Surgery, Imperial College London, London, UK
Baoru.Huang18@imperial.ac.uk
[2] Department of Surgery & Cancer, Imperial College London, London, UK
[3] The Kennedy Institute of Rheumatology, University of Oxford, Oxford, UK
[4] Big Data Institute, University of Oxford, Oxford, U.K.
[5] Department of Computer Science, University of Liverpool, Liverpool, UK
[6] Lightpoint Medical Ltd., Chesham, UK

Abstract. Depth estimation is a crucial step for image-guided intervention in robotic surgery and laparoscopic imaging system. Since per-pixel depth ground truth is difficult to acquire for laparoscopic image data, it is rarely possible to apply supervised depth estimation to surgical applications. As an alternative, self-supervised methods have been introduced to train depth estimators using only synchronized stereo image pairs. However, most recent work focused on the left-right consistency in 2D and ignored valuable inherent 3D information on the object in real world coordinates, meaning that the left-right 3D geometric structural consistency is not fully utilized. To overcome this limitation, we present M3Depth, a self-supervised depth estimator to leverage 3D geometric structural information hidden in stereo pairs while keeping monocular inference. The method also removes the influence of border regions unseen in at least one of the stereo images via masking, to enhance the correspondences between left and right images in overlapping areas. Intensive experiments show that our method outperforms previous self-supervised approaches on both a public dataset and a newly acquired dataset by a large margin, indicating a good generalization across different samples and laparoscopes.

Keywords: Self-Supervised Monocular Depth Estimation · Laparoscopic Images · 3D Geometric Consistency

1 Introduction

Perception of 3D surgical scenes is a fundamental problem in computer assisted surgery. Accurate perception, tissue tracking, 3D registration between intra- and pre-operative organ models, target localization and augmented reality [15, 8] are

predicated on having access to correct depth information. Range finding sensors such as multi-camera systems or LiDAR that are often employed in autonomous systems and robotics are not convenient for robot-assisted minimally invasive surgery because of the limited port size and requirement of sterilization. Furthermore, strong 'dappled' specular reflections as well as less textured tissues hinder the application of traditional methods [16]. This has led to the exploration of learning-based methods, among which fully convolutional neural networks (CNNs) are particularly successful [1, 22].

Since it is challenging to obtain per-pixel ground truth depth for laparoscopic images, there are far fewer datasets in the surgical domain compared with mainstream computer vision applications [2, 9]. It is also not a trivial task to transfer approaches that are based on supervised learning to laparoscopic applications due to the domain gap. To overcome these limitations, view-synthesis methods are proposed to provide self-supervised learning for depth estimation [10, 15], with no supervision via per-pixel depth data. Strong depth prediction baselines have been established in [3, 4, 11]. However, all of these methodologies employed left-right consistency and smoothness constraints in 2D, *e.g.* [3, 5, 7], and ignored the important 3D geometric structural consistency from the stereo images.

Recently, a self-supervised semantically-guided depth estimation method was proposed to deal with moving objects [13], which made use of mutually beneficial cross-domain training of semantic segmentation. Jung *et al.* [12] extended this work by incorporating semantics-guided local geometry into intermediate depth representations for geometric representation enhancement. However, semantic labels are not common in laparoscopic applications except for surgical tool masks, impeding the extension of this work. Mahjourian *et al.* [17] presented an approach for unsupervised learning of depth by enforcing consistency of ego-motion across consecutive frames to infer 3D geometry of the whole scene. However, in laparoscopic applications, the interaction between the surgical tools and tissue creates a dynamic scene, leading to failure of local photometric and geometric consistency across consecutive frames in both 2D and 3D. Nevertheless, the 3D geometry inferred from left and right synchronized images can be assumed identical, allowing adoption of 3D- as well as 2D-loss.

In this paper, we propose a new framework for self-supervised laparoscopic image depth estimation called M3Depth, leveraging not only the left-right consistency in 2D but also the inherent geometric structural consistency of real-world objects in 3D (see section 2.2 for the 3D geometric consistency loss), while enhancing the mutual information between stereo pairs. A U-Net architecture [20] was employed as the backbone and the network was fed with only left image as inputs but was trained with the punitive loss formed by stereo image pairs. To cope with the unseen areas at the image edges that were not visible in both cameras, blind masking was applied to suppress and eliminate outliers and give more emphasis to feature correspondences that lay on the shared vision field. Extensive experiments on both a public dataset and a new experimental dataset demonstrated the effectiveness of this approach and a detailed ablation study

indicated the respective positive influence of each proposed novel module on the overall performance.

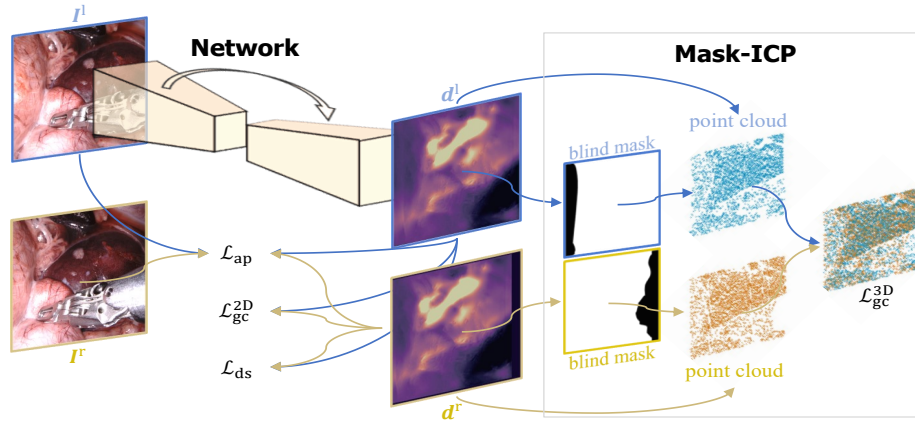## 2   Methodology

2.1   Network Architecture

Fig. 1. Overview of the proposed self-supervised depth estimation network. ResNet18 was adopted as the backbone and received a left image from a stereo image pair as the input. Left and right disparity maps were produced simultaneously and formed 2D losses with the original stereo pair. 3D point clouds were generated by applying the intrinsic parameters of the camera and iterative closest point loss was calculated between them. Blind masks were applied to the 2D disparity maps to remove outliers from areas not visible in both cameras.

Network Architecture The backbone of the M3Depth followed the general U-Net [20] architecture, *i.e.* an encoder-decoder network, in which an encoder was employed to extract image representations while a decoder with convolutional layer and upsampling manipulation was designed to recover disparity maps at the original scale. Skip connections were applied to obtain both deep abstract features and local information. To keep a lightweight network, a ResNet18 [6] was employed as the encoder with only 11 million parameters. To improve the regression ability of the network from intermediate high-dimensional features maps to disparity maps, one more ReLU [18] activation function and a convolutional layer with decreased last latent feature map dimension were added before the final sigmoid disparity prediction. Similar to Monodepth1 [3], in M3Depth, the left image $\boldsymbol{I}^{\mathrm{l}}$ of a stereo image pair $(\boldsymbol{I}^{\mathrm{l}}, \boldsymbol{I}^{\mathrm{r}} \in \mathbb{R}_+^{h \times w \times 3})$ was always the input and the framework generated two distinct left and right disparity maps $\boldsymbol{d}^{\mathrm{l}}$,

$d^{\mathrm{r}} \in \mathbb{R}_+^{h \times w}$ simultaneously, *i.e.* $\mathcal{Z} : I^{\mathrm{l}} \mapsto (d^{\mathrm{l}}, d^{\mathrm{r}})$. Given the camera focal length f and the baseline distance b between the cameras, left and right depth maps $D^{\mathrm{l}}, D^{\mathrm{r}} \in \mathbb{R}_+^{h \times w}$ could then be trivially recovered from the predicted disparity, $(D^{\mathrm{l}}, D^{\mathrm{r}}) = bf / (d^{\mathrm{l}}, d^{\mathrm{r}})$. h and w denote image height and width. Full details of the architecture are presented in the supplementary material.

Image Reconstruction Loss in 2D With the predicted disparity maps and the original stereo image pair, left and right images could then be reconstructed by warping the counter-part RGB image with the disparity map mimicking optical flow [3] [14]. Similar to Monodepth1 [3], an appearance matching loss $\mathcal{L}_{\mathrm{ap}}$, disparity smoothness loss $\mathcal{L}_{\mathrm{ds}}$ and left-right disparity consistency loss $\mathcal{L}_{\mathrm{lr}}^{2D}$ were used to encourage coherence between the original input and reconstructed images $(I^{\mathrm{l}*}, I^{\mathrm{r}*})$ as well as consistency between left and right disparities while forcing disparity maps to be locally smooth.

$$\mathcal{L}_{ap}^{\mathrm{r}} = \frac{1}{N} \sum_{i,j} \frac{\gamma}{2}(1 - \mathrm{SSIM}(I_{ij}^{\mathrm{r}}, I_{ij}^{r*})) + (1-\gamma)\|I_{ij}^{\mathrm{r}} - I_{ij}^{r*}\|_1 \tag{1}$$

$$\mathcal{L}_{\mathrm{lr}}^{2\mathrm{D(r)}} = \frac{1}{N} \sum_{ij} |\mathbf{d}_{ij}^{\mathrm{r}} + \mathbf{d}_{ij+\mathbf{d}_{ij}^{\mathrm{r}}}^{\mathrm{l}}| \tag{2}$$

$$\mathcal{L}_{\mathrm{ds}}^{\mathrm{r}} = \frac{1}{N} \sum_{ij} |\partial_x(\mathbf{d}_{ij}^{\mathrm{r}})|e^{-|\partial_x I_{ij}^{\mathrm{r}}|} + |\partial_y(\mathbf{d}_{ij}^{\mathrm{r}})|e^{-|\partial_y I_{ij}^{\mathrm{r}}|} \tag{3}$$

where N is the number of pixels and γ was set to 0.85. Note that 2D losses were applied on both left and right images but only equations for the right image are presented here.

2.2   Learning 3D Geometric Consistency

Instead of using the inferred left and right disparities only to establish a mapping between stereo coordinates and generate reconstructed original RGB input images, a loss function was also constructed that registered and compared left and right point clouds directly to enforce the 3D geometric consistency of the whole scene. The disparity maps of the left and right images were first converted to depth maps and then backprojected to 3D coordinates to obtain left and right surgical scene point clouds $(P^{\mathrm{l}}, P^{\mathrm{r}} \in \mathbb{R}^{hw \times 3})$ by multiplying the depth maps with the intrinsic parameter matrix (K). The 3D consistency loss employed Iterative Closest Point (ICP) [17, 21], a classic rigid registration method that derives a transformation matrix between two point clouds by iteratively minimizing point-to-point distances between correspondences.

From an initial alignment, ICP alternately computed corresponding points between two input point clouds using a closest point heuristic and then recomputed a more accurate transformation based on the given correspondences. The final residual registration error after ICP minimization was output as one of the returned values. More specifically, to explicitly explore global 3D loss, the ICP

loss at the original input image scale was calculated with only 1000 randomly selected points to reduce the computational workload.

2.3   Blind Masking

Some parts of the left scene were not visible in the right view and vice versa, leading to non-overlapping generated point clouds. These areas are mainly located at the left edge of the left image and right edge of the right image after rectification. Depth and image pixels in those area had no useful information for learning, either in 2D and 3D. Our experiments indicated that retaining the contribution to the loss functions for such pixels and voxels degraded the overall performance. Many previous approaches solved this problem by padding these areas with zeros [3] or values from the border [4], but this can lead to edge artifacts in depth images [17].

To tackle this problem, we present a blind masking module $\mathcal{M}^{l,r}$ that suppressed and eliminated these outliers and gave more emphasis to correspondences between the left and right views. First, a meshgrid was built with the original left image pixel coordinates in both horizontal $\mathcal{X}_{\text{grid}}$ and vertical $\mathcal{Y}_{\text{grid}}$ directions. Then the $\mathcal{X}_{\text{grid}}$ was shifted along the horizontal direction using the right disparity map $\boldsymbol{d}^r$ to a get a new grid $\mathcal{X}'_{\text{grid}}$, which was then stacked with $\mathcal{Y}_{\text{grid}}$ to form a new meshgrid. Finally, grid sampling was employed on the new meshgrid with the help of the original left image coordinates, from which the pixels that were not covered by the right view for the current synchronized image pair were obtained. By applying the blind masking on the depth maps for the stereo 3D point cloud generation, a 3D alignment loss was obtained as follows.

$$\boldsymbol{M}^{l,r}_{ij} = \begin{cases} 1, (\mathbf{d}^{l,r}_{ij} + \mathcal{X}_{ij}) \in \{\mathcal{X}\} \\ 0, (\mathbf{d}^{l,r}_{ij} + \mathcal{X}_{ij}) \notin \{\mathcal{X}\} \end{cases} \tag{4}$$

$$\boldsymbol{P}^{l,r} = \text{backproj}(\mathbf{d}^{l,r}, \boldsymbol{K}, \boldsymbol{M}^{l,r}) \tag{5}$$

$$\mathcal{L}^{3D}_{\text{gc}} = \text{ICP}(\boldsymbol{P}^l, \boldsymbol{P}^r) \tag{6}$$

2.4   Training Loss

Pixel-wise, gradient-based 2D losses and point cloud-based 3D losses were applied to force the reconstructed image to be identical to the original input while encouraging the left-right consistency in both 2D and 3D to derive accurate disparity maps for depth inference. Finally, an optimization loss used a combination of these, written as:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= (\mathcal{L}^r_{2D} + \mathcal{L}^l_{2D}) + \mathcal{L}^{3D}_{lr} \\ &= \alpha_{\text{ap}}(\mathcal{L}^r_{\text{ap}} + \mathcal{L}^l_{\text{ap}}) + \alpha_{\text{ds}}(\mathcal{L}^r_{\text{ds}} + \mathcal{L}^l_{\text{ds}}) + \alpha^{2D}_{lr}(\mathcal{L}^{2D(r)}_{lr} + \mathcal{L}^{2D(l)}_{lr}) + \beta\mathcal{L}^{3D}_{\text{gc}} \end{aligned} \tag{7}$$

where α_* and β balanced the loss magnitude of the 2D and 3D parts to stabilize the training. More specifically, α_{ap}, α_{ds}, α^{2D}_{lr} and β were experimentally set to 1.0, 0.5, 1.0 and 0.001.

3   Experiments

3.1   Dataset

M3Depth was evaluated on two datasets. The first was the *SCARED* dataset [1] released at the MICCAI Endovis challenge 2019. As only the ground truth depth map of key frames in each dataset was provided (from structured light), the other depth maps were created by reprojection and interpolation of the key frame depth maps using the kinematic information from the *da Vinci* robot, causing a misalignment between the ground truth and the RGB data. Hence, only key-frame ground truth depth maps were used in the test dataset while the remainder of the RGB data formed the training set but with the similar adjacent frames removed.

To overcome the *SCARED* dataset misalignment and improve the validation, an additional laparoscopic image dataset (namely *LATTE*) was experimentally collected, including RGB laparoscopic images and corresponding ground truth depth maps calculated from a custom-built structured lighting pattern projection. More specifically, the gray-code detection and decoder algorithm [23] were used with both original and inverse patterns. To remove the uncertainty brought by occlusions and uneven illumination conditions, we used a more advanced 3-phase detection module, in which sine waves were shifted by $\pi/3$ and $2\pi/3$ and the modulation depth $\mathcal{T}$ was calculated for every pixel. Pixels with modulation depth under $\mathcal{T}$ were defined as uncertain pixels, and the equation for calculating the modulation is written as Eq. 8. This provided 739 extra image pairs for training and 100 pairs for validation and testing.

$$\mathcal{T} = \frac{2\sqrt{2}}{3} \times \sqrt{(\mathcal{I}_1 - \mathcal{I}_2)^2 + (\mathcal{I}_2 - \mathcal{I}_3)^2 + (\mathcal{I}_1 - \mathcal{I}_3)^2} \tag{8}$$

where $\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3$ denotes original modulation and modulations after shifts.

Table 1. Quantitative results on the *SCARED* dataset. Metrics labeled with blue headings mean *lower is better* while those labeled with red mean *higher is better*.

Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Mono2 [4]	1.100	74.408	56.548	0.717	0.102	0.284	0.476
PackNet3D [5]	0.733	37.690	32.579	0.649	0.288	0.538	0.722
Mono1 [3]	0.257	20.649	33.796	0.404	0.696	0.837	0.877
M3Depth	**0.116**	**1.822**	**9.274**	**0.139**	**0.865**	**0.983**	**0.997**

3.2   Evaluation Metrics, Baseline, and Implementation Details

Evaluation Metrics To evaluate depth errors, seven criteria were adopted that are commonly used for monocular depth estimation tasks [3][4]: mean absolute error (Abs Rel), squared error (Sq Rel), root mean squared error (RMSE), root
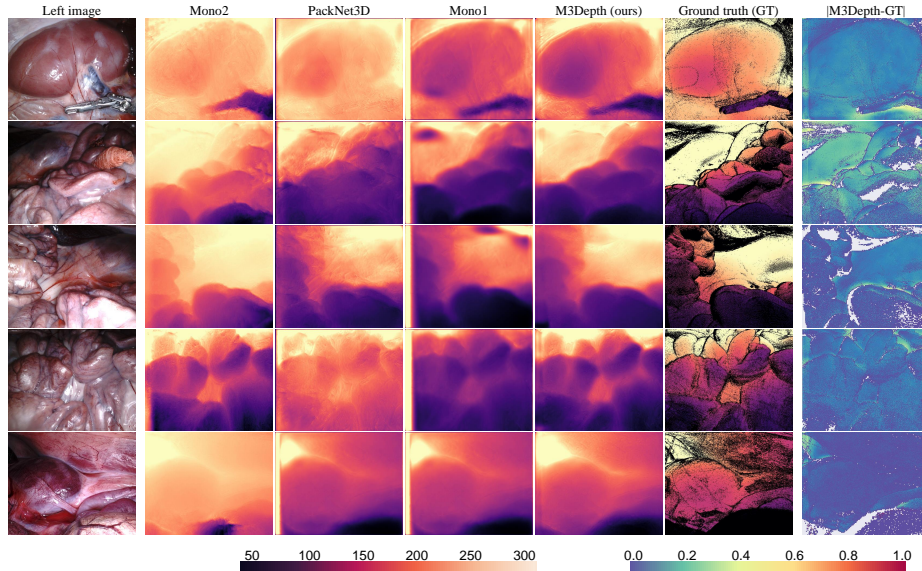
Fig. 2. Qualitative results on the *SCARED* dataset with error map of M3Depth. The depth predictions are all for the left input image. M3Depth generated depth maps with high contrast between the foreground and background and performed better at distinguishing different parts of the scene, reflecting the superior quantitative results in Table 1.

mean squared logarithmic error (RMSE log), and the ratio between ground truth and prediction values, for which the threshold was denoted as δ.

Baseline The M3Depth model was compared with several recent deep learning methods including Monodepth [3], Monodepth2 [4], and PackNet [5], and both quantitative and qualitative results were generated and reported for comparison. To further study the importance of each M3Depth component, the various components of M3Depth were removed in turn.

Implementation Details M3Depth was implemented in PyTorch [19], with an input/output resolution of 256×320 and a batch size of 18. The learning rate was initially set to 10^{-5} for the first 30 epochs and was then halved until the end. The model was trained for 50 epochs using the Adam optimizer which took about 65 hours on two NVIDIA 2080 Ti GPUs.

4   Results and Discussion

The M3Depth and other state-of-the-art results on the *SCARED* and *LATTE* dataset are shown in Table 1 and Table 2 using the seven criteria evaluation

Table 2. Quantitative results on *LATTE* dataset. Metrics labeled with blue headings mean *lower is better* while labeled by red mean *higher is better*.

Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Mono2 [4]	1.601	306.823	87.694	0.913	0.169	0.384	0.620
PackNet3D [5]	0.960	357.023	259.627	0.669	0.135	0.383	0.624
Mono1 [3]	0.389	57.513	99.020	0.424	0.268	0.709	0.934
M3Depth	**0.236**	**21.839**	**57.739**	**0.245**	**0.665**	**0.893**	**0.969**

Table 3. Ablation study results on the *SCARED* dataset. Metrics labeled with blue headings mean *lower is better* while those labeled with red mean *higher is better*.

Method	3GC	Blind masking	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Mono1 [3]	✗	✗	0.257	20.649	33.796	0.404	0.696	0.837	0.877
M3Depth w/ mask	✓	✗	0.150	3.069	13.671	0.249	0.754	0.910	0.956
M3Depth	✓	✓	**0.116**	**1.822**	**9.274**	**0.139**	**0.865**	**0.983**	**0.997**

metrics. M3Depth outperformed all other methods by a large margin on all seven criteria, which shows that taking the 3D structure of the world into consideration benefited the overall performance of the depth estimation task. In particular, the M3Depth model had 0.141, 18.827, 24.522, and 0.265 units error lower than Monodepth1[3] in Abs Rel, Sq Rel, RMSE and RMSE log, and 0.169, 0.146, and 0.12 units higher than Monodepth1 [3] in three different threshold criteria. Furthermore, the average inference time of M3Depth was 105 frames per second, satisfying the real-time depth map generation requirements.

Detailed ablation study results on the *SCARED* dataset are shown in Table 3 and the impact of the proposed modules, 3D geometric consistency (3GC) and blind masking were evaluated. The evaluation measures steadily improved when the various components were added. More specifically, the addition of the blind masking further boosted the 3GC term, which shows the importance and necessity of removing invalid information from areas that are not visible to both cameras. We note that more quantitative results can be found in our supplementary material.

Qualitative results comparing our depth estimation results against prior work using the *SCARED* dataset are presented in Fig 2. As the sample images shows, the application of temporal consistency encouraged by the 3D geometric consistency loss can reduce the errors caused by subsurface features, and better recover the real 3D surface shape of the tissue. Furthermore, depth outputs from M3Depth show better results along the boundaries of objects, indicating the effectiveness of the proposed 3GC and blind masking modules.

5   Conclusion

A novel framework for self-supervised monocular laparoscopic images depth estimation was presented. By combining the 2D image-based losses and 3D geometry-based losses from an inferred 3D point cloud of the whole scene, the

global consistency and small local neighborhoods were both explicitly taken into consideration. Incorporation of blind masking avoided penalizing areas where no useful information exists. The modules proposed can easily be plugged into any similar depth estimation network, monocular and stereo, while the use of a lightweight ResNet18 backbone enabled real time depth map generation in laparoscopic applications. Extensive experiments on both public and newly acquired datasets demonstrated good generalization across different laparoscopes, illumination condition and samples, indicating the capability to large scale data acquisition where precise ground truth depth cannot be easily collected.

References

1. Allan, M., Mcleod, J., Wang, C.C., Rosenthal, J.C., Fu, K.X., Zeffiro, T., Xia, W., Zhanshi, Z., Luo, H., Zhang, X., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv:2101.01133 (2021)
2. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
3. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017)
4. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3828–3838 (2019)
5. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2485–2494 (2020)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Huang, B., Nguyen, A., Wang, S., Wang, Z., Mayer, E., Tuch, D., Vyas, K., Giannarou, S., Elson, D.S.: Simultaneous depth estimation and surgical tool segmentation in laparoscopic images. IEEE Transactions on Medical Robotics and Bionics **4**(2), 335–338 (2022)
8. Huang, B., Tsai, Y.Y., Cartucho, J., Vyas, K., Tuch, D., Giannarou, S., Elson, D.S.: Tracking and visualization of the sensing area for a tethered laparoscopic gamma probe. International Journal of Computer Assisted Radiology and Surgery **15**(8), 1389–1397 (2020)
9. Huang, B., Zheng, J.Q., Giannarou, S., Elson, D.S.: H-net: Unsupervised attention-based stereo depth estimation leveraging epipolar geometry. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4460–4467 (2022)
10. Huang, B., Zheng, J.Q., Nguyen, A., Tuch, D., Vyas, K., Giannarou, S., Elson, D.S.: Self-supervised generative adversarial network for depth estimation in laparoscopic images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 227–237. Springer (2021)
11. Johnston, A., Carneiro, G.: Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4756–4765 (2020)

12. Jung, H., Park, E., Yoo, S.: Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12642–12652 (2021)
13. Klingner, M., Termöhlen, J.A., Mikolajczyk, J., Fingscheidt, T.: Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In: European Conference on Computer Vision. pp. 582–600. Springer (2020)
14. Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International Conference on 3D Vision (3DV). pp. 218–227. IEEE (2021)
15. Liu, X., Sinha, A., Ishii, M., Hager, G.D., Reiter, A., Taylor, R.H., Unberath, M.: Dense depth estimation in monocular endoscopy with self-supervised learning methods. IEEE transactions on medical imaging **39**(5), 1438–1447 (2019)
16. Luo, W., Schwing, A.G., Urtasun, R.: Efficient deep learning for stereo matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5695–5703 (2016)
17. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and egomotion from monocular video using 3d geometric constraints. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5667–5675 (2018)
18. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Icml (2010)
19. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
21. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: Proceedings third international conference on 3-D digital imaging and modeling. pp. 145–152. IEEE (2001)
22. Tran, M.Q., Do, T., Tran, H., Tjiputra, E., Tran, Q.D., Nguyen, A.: Light-weight deformable registration using adversarial learning with distilling knowledge. IEEE Transactions on Medical Imaging (2022)
23. Xu, Y., Aliaga, D.G.: Robust pixel classification for 3d modeling with structured light. In: Proceedings of Graphics Interface 2007. pp. 233–240 (2007)