Simultaneous Depth Estimation and Surgical Tool Segmentation in Laparoscopic Images

Baoru Huang[1,2], Anh Nguyen[1,5], Siyao Wang[1], Ziyang Wang[3], Erik Mayer[2], David Tuch[4], Kunal Vyas[4], Stamatia Giannarou[1,2], Daniel S. Elson[1,2]

***Abstract*—Surgical instrument segmentation and depth estimation are crucial steps to improve autonomy in robotic surgery. Most recent works treat these problems separately, making the deployment challenging. In this paper, we propose a unified framework for depth estimation and surgical tool segmentation in laparoscopic images. The network has an encoder-decoder architecture and comprises two branches for simultaneously performing depth estimation and segmentation. To train the network end to end, we propose a new multi-task loss function that effectively learns to estimate depth in an unsupervised manner, while requiring only semi-ground truth for surgical tool segmentation. We conducted extensive experiments on different datasets to validate these findings. The results showed that the end-to-end network successfully improved the state-of-the-art for both tasks while reducing the complexity during their deployment.**

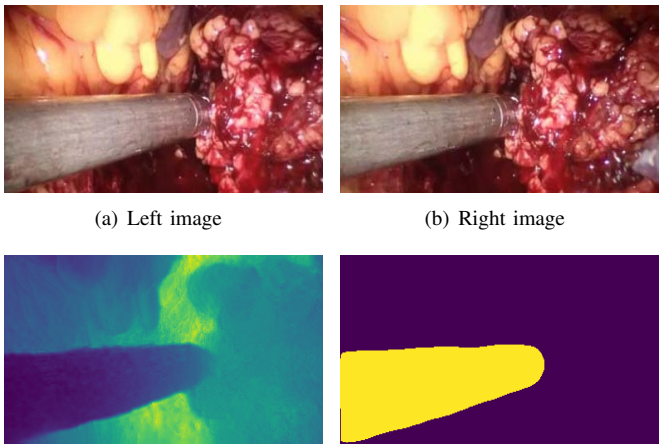
***Index Terms*—Deep learning, Self-supervised depth estimation, Surgical instrument segmentation, Multi-task learning**

I. Introduction

MINIMALLY invasive surgery (MIS), including robot-assisted procedures, provides significant advantages such as reducing operative trauma and the risk of infection. Advanced robotic surgery systems such as the da Vinci surgical platform [1] allow multiple types of information to be integrated together with effective feedback to the surgeon. However, interpreting visual surgical data is complex and involves many tasks such as tissue deformation modeling [2], tool tracking [3], and scene depth estimation [4] [5].

In recent years there has been much work on depth estimation and surgical tool segmentation. Notably, learning-based algorithms have shown excellent prediction capability of the relationship between color images and depth, as well as image segmentation into meaningful regions. These depth-predicting algorithms may use monocular or stereo input data, with either supervised, self-supervised or unsupervised [6] training approaches depending on availability of ground truth labels. Instrument segmentation may also use supervised or unsupervised methods [7]. Knowing the tissue depth and the instrument masks could facilitate tissue scanning [8] or dynamic image overlays [9], which are useful for laparoscopic surgery.

[1]The Hamlyn Centre for Robotic Surgery, Imperial College London, SW7 2AZ, UK Baoru.Huang18@imperial.ac.uk
[2]Department of Surgery & Cancer, Imperial College London, SW7 2AZ, UK
[3]Department of Computer Science, University of Oxford, UK
[4]Lightpoint Medical Ltd.
[5]Department of Computer Science, University of Liverpool, UK



(a) Left image (b) Right image

(c) Depth map (d) Segmentation map

Fig. 1. Simultaneous depth estimation and surgical tool segmentation in laparoscopic images. Given the left and the right image, our method is able to perform depth estimation and surgical tool segmentation simultaneously.

To date, depth estimation and surgical tool segmentation have been mainly treated as separate challenges, requiring time-consuming sequential task completion. In this work, we propose a novel unified framework that can perform simultaneous depth estimation and surgical tool segmentation: SDSNet. Our method does not require manually labeled ground truth, and achieves the state-of-the-art performance for both tasks, as well as reducing the deployment complexity.

II. Related Work

Depth Estimation Most existing methods treat depth estimation as a supervised regression problem [10], however, collecting per-pixel ground truth for laparoscopic imaging is challenging. To overcome this limitation, Liu *et al.* [11] introduced a self-supervised algorithm for dense depth estimation in stereo endoscopy. The authors in [12] proposed a geometry-aware network for motion estimation. By enforcing consistency between left and right RGB images, Godard *et al.* [13] produced results that outperformed contemporary supervised methods.

Surgical Tool Segmentation Semantic segmentation of robotic instruments has also attracted a lot of attention in robot-assisted surgery research [14]. Some discriminative models such as Naive Bayesian classifiers [15] and maximum likelihood Gaussian Mixture Models [16] can be trained on color features. More recently, the state of the art has increas-

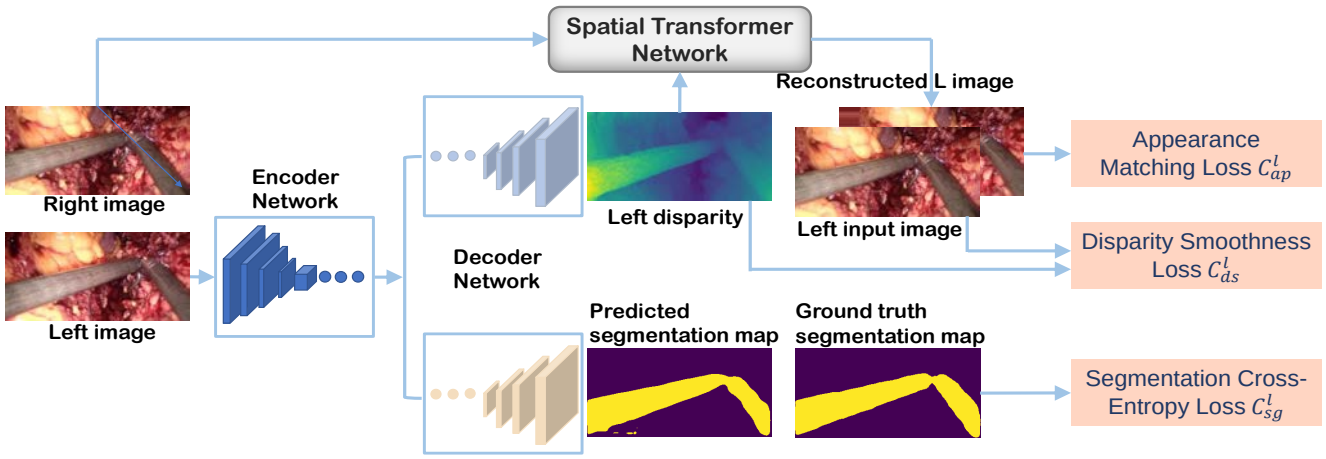


Fig. 2. An overview of our learning framework. The proposed architecture includes two tasks (image depth estimation and segmentation) and loss constraints from both spatial and spectral perspectives across the tasks. The SDSNET generates both depth map and segmented map from the left image, while the Spatial Transformer Network combines the depth map and the right image to reconstruct to left image for unsupervised depth estimation.

ingly focused on fully convolutional neural networks. The authors in [17] used CNN for segmenting robotic tools.

Simultaneous Depth Estimation and Segmentation The task of depth estimation and segmentation are usually tackled separately, with few works unifing both tasks, especially for laparoscopy. Most recent methods used RGB images as the training data, for instance in EdgeStereo [18] the authors incorporated edge detection to accurately estimate depth changes across object boundaries. In medical imaging, self-supervised depth estimation was used to regularize the semantic segmentation during knee arthroscopy [19].

III. METHODOLOGY

An overview of our proposed SDSNET can be found in Fig. 3. We first combined the depth estimation and tool segmentation tasks by sharing an encoder network, where essential geometric features from the input images were extracted. After the encoder, the features flowed separately to two branches (segmentation and depth estimation). By forcing the disparity map to generate a reconstructed input image that is consistent with the original, we could derive an accurate disparity map for depth inference.

Depth Estimation Branch The depth estimation branch was based on the general U-Net architecture [20], *i.e.* an encoder-decoder network with skip connections, which represents local information as well as deep abstract features. The size of the input batch was $b \times 3 \times 192 \times 384$, where b was the batch size, 3 was the number of channels and 192×384 was the size of the input image. A Resnet50 was adopted as the encoder to extract features from the input color image.

The decoder consisted of five cascaded blocks of multiple scales. Previously, multi-scale depth predictions and image reconstruction used gradient locality of a bilinear sampler [21], which was prone to create 'holes' and texture-copy artifacts in large low-texture regions. In our work, similar to [13], this problem is tackled by decoupling the resolutions of the disparity maps and corresponding color images used to compute the reprojection error. The lower resolution depth maps were first upsampled to the input image resolution and then reprojected and resampled. From the second block, the output of each block was taken by the convolutional layer and followed by a sigmoid activation function, which generated the disparity map at each scale. In total, 4 scales were used with output sizes $b\times1\times24\times48$, $b\times1\times48\times96$, $b\times1\times96\times192$, $b \times 1 \times 192 \times 384$. The largest of these was the final disparity map which was the same size as the input image.

The final disparity map (the sigmoid output) $\hat{D}$ was converted to a depth map by $D = 1/(a\hat{D} + b)$, in which a and b constrained D between 0.1 and 80 units.

Segmentation Branch The shared encoder features were also fed into the segmentation map decoder, which consisted of five convolutional layers and three upsampling layers to interpolate the features to full image resolution. The first layer was $b \times 256 \times 6 \times 12$ and took the $b \times 2048 \times 6 \times 12$ input followed by an ELU activation function. The upsampling layer interpolated the features to four times the input size. After the third convolutional layer, the features were concatenated into the corresponding layer from the depth estimation decoder to perform feature fusion between the two branches. The size of the segmentation subnetwork output was $b \times k \times 192 \times 384$, where $k = 2$ was the number of classes. In practice, the depth decoder features were concatenated with the segmentation branch decoder in the fourth block. We generated the surgical instrument segmentation semi-ground truth by applying the network from [22] pretrained on the EndoVis dataset [23].

Multi-Task Loss The network was trained end-to-end using a multi-task loss function $C_{\mathrm{t}}^{\mathrm{l}}$, which was formed as

$$C_{\mathrm{t}}^{\mathrm{l}} = \alpha_{\mathrm{dp}} C_{\mathrm{dp}}^{\mathrm{l}} + \alpha_{\mathrm{sg}} C_{\mathrm{sg}}^{\mathrm{l}} \tag{1}$$

where C_{dp}^{l} is the loss from the depth estimation branch and C_{sg}^{l} is from semantic segmentation, as described below.

1) Depth Loss: In the depth estimation branch, the depth loss $C_{\mathrm{dp}}^{\mathrm{l}}$ consisted of the appearance matching loss $C_{\mathrm{ap}}^{\mathrm{l}}$ and

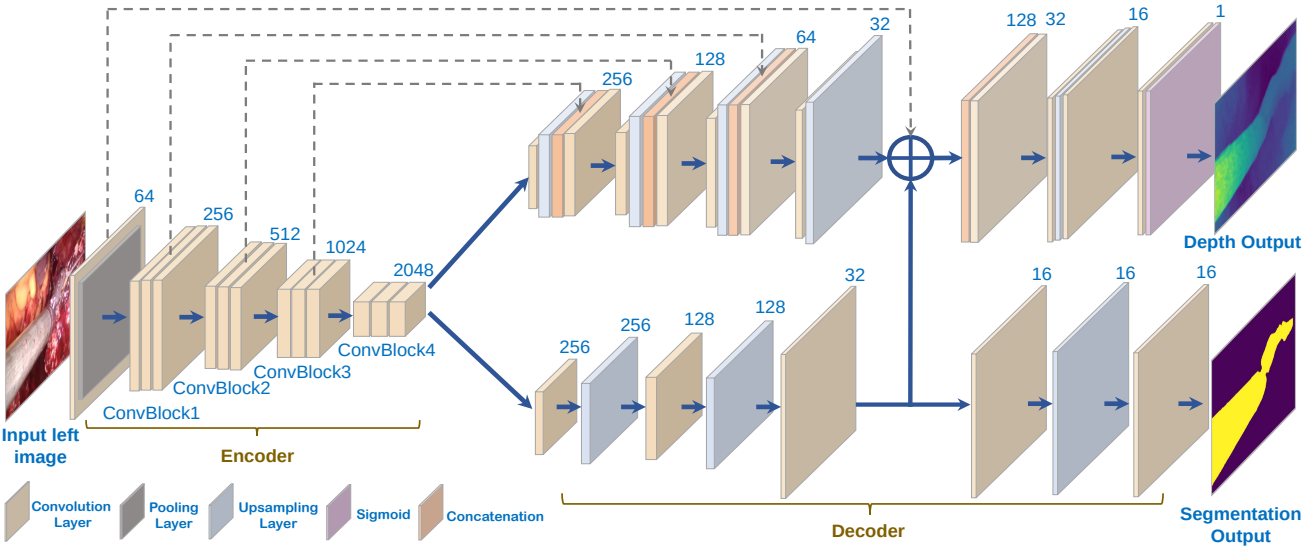


Fig. 3. The detailed architecture of SDSNET. The depth branch and the segmentation branch share the same encoder network. The features of the third convolutional layer in the segmentation branch decoder are fused with the features from the fourth block in the depth branch decoder. We use only one ConvBlock to represent all repeated blocks for better visualization

.

disparity smoothness loss C_{ds}^{l} as:

$$C_{dp}^{\text{l}} = \sum_{s=1}^{4} C_s^{\text{l}} = \sum_{s=1}^{4} (C_{ap}^{\text{l}} + \alpha_{ds} C_{ds}^{\text{l}}) \tag{2}$$

where α_{ds} was set to 0.001.

Appearance Matching Loss The appearance matching loss C_{ap}^{l} forced the reconstructed image to be similar to the corresponding training input and was computed for the higher input resolution. During training, the autoencoder in the depth estimation branch generated a disparity map $\hat{D}_t$ from the input left color image I_t^l. This map was then transformed using an image sampler from the Spatial Transformer Network (STN) [21], along with the right input image I_t^r (the counterpart of I_t^l), to reconstruct the left image I_t^{l*}. This sampler model used bilinear interpolation and the output pixel was the weighted sum of four input pixels. This bilinear sampler was locally fully differentiable and could be seamlessly integrated into the fully convolutional architecture, in contrast to [24]. Hence, there was no need to simplify or approximate the cost function. As in [25], we applied a combination of L_1 loss and structural similarity (SSIM) index as the photometric image reconstruction cost C_{ap}^{l}. Training the depth estimation network then required minimizing the reconstruction loss between the reconstructed image $I^{\text{l}*}$ and the corresponding training input I^{l}, where N denotes the number of pixels.

$$C_{\text{ap}}^{\text{l}} = \frac{1}{N} \sum_{i,j} \frac{\gamma}{2} (1 - \text{SSIM}(I_{ij}^{\text{l}}, I_{ij}^{\text{l}*})) + (1 - \gamma) \| I_{ij}^{\text{l}} - I_{ij}^{\text{l}*} \|_1 \tag{3}$$

Similar to [13], the SSIM was simplified to a 3×3 block filter rather than a Gaussian, and γ was set to 0.85.

Disparity Smoothness Loss Smooth disparities were favored by this loss, and since discontinuities usually occur at image gradients [13], this cost was weighted by an edge-aware term based on the image gradients ∂I.

$$C_{\text{ds}}^{\text{l}} = \frac{1}{N} \sum_{ij} | \partial_x \hat{D}_{ij}^{\text{l}} | e^{-|\partial_x I_{ij}^{\text{l}}|} + | \partial_y \hat{D}_{ij}^{\text{l}} | e^{-|\partial_y I_{ij}^{\text{l}}|} \tag{4}$$

2) Segmentation Loss: The segmentation branch only considers the full resolution image to reduce the computational complexity. For the segmentation subnetwork, given a sequence of input images and the corresponding sequence of semi-ground truth segmentation annotations, we performed end-to-end training by minimizing the normalized pixel-wise cross-entropy loss [26], which is denoted as C_{sg}^{l}.

$$C_{\text{sg}}^{\text{l}} = -\sum_{i=1}^{N} \hat{y}_i * \log(y_i) \tag{5}$$

where y_i, $\hat{y}_i$ are the predicted value, and semi-ground truth.

Training As there was no per-pixel depth ground truth label available, the depth estimation relied on the image reconstruction similarity, trained in self-supervised mode. For depth estimation, the data augmentation was performed by flipping 50% the input images horizontally. For segmentation, the semi-ground truth was provided for supervised training. The whole SDSNET was trained end-to-end with the combination of losses from each branch that involved the generation of a depth map and segmentation map.

IV. EXPERIMENTS

Experimental Setup We evaluated our SDSNET on two datasets: $\mathcal{D}^{sia}$ [27] and $\mathcal{D}^{por}$ [28]. For the depth estimation branch, similar to [29], we used the SSIM index to evaluate the unsupervised depth estimation. To evaluate the result of the segmentation branch, we manually labeled 400 images with the surgical tool ground truth and the segmentation performance was assessed by the Jaccard index and the Dice Score [20].

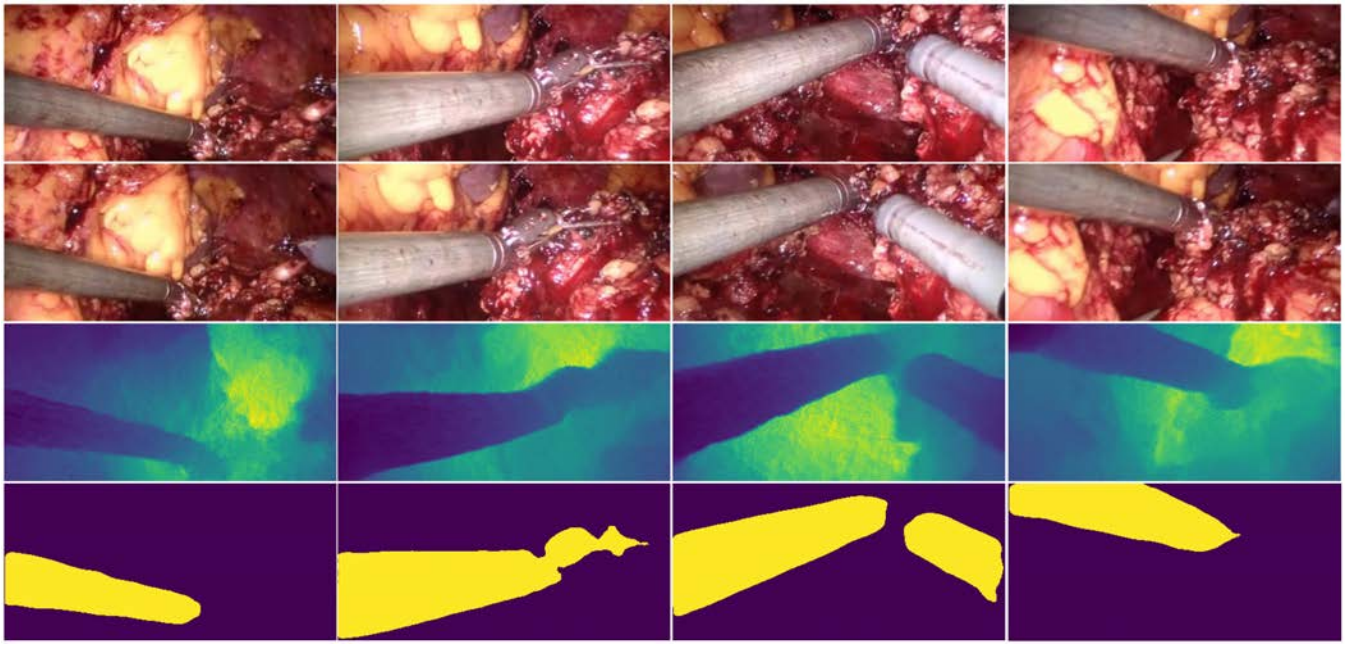


Fig. 4. Qualitative results on stereo pairs. **From top to bottom**: the left input image, the corresponding right image in the stereo pair, the predicted depth map, and the segmentation result from (SDSNET).

TABLE I
SSIM SCORES ON THE $\mathcal{D}^{sia}$ TEST SET

	Mean SSIM	Std.SSIM
ELAS [31]	47.3	0.079
SPS [30]	54.7	0.092
V-Basic [27]	55.5	0.106
V-Siamese [27]	60.4	**0.066**
Monodepth [13]	58.4	0.114
Monodepth2 [29]	71.2	0.075
SDSNET without fusion (ours)	71.9	0.079
SDSNET with fusion (ours)	**72.8**	0.073

TABLE II
SEGMENTATION RESULTS ON THE $\mathcal{D}^{sia}$ TEST SET

	IoU	Dice	Time
U-Net [20]	71.16	80.90	0.22
SDSNET (segmentation only) (ours)	73.34	84.13	0.30
SDSNET without fusion (ours)	73.44	84.59	0.38
SDSNET with fusion (ours)	**74.92**	**85.63**	0.35

Baseline For depth estimation, we compared the results to those from the Basic and Siamese architectures [27], Monodepth2 [29], and two non-learning methods, SPS [30] and ELAS [31]. For surgical instrument segmentation, results from the SDSNET were compared with the popular U-Net [20] architecture.

Implementation The SDSNET model was implemented in PyTorch [32], with a batch size of 16 and an input/output resolution of 192 x 384. The learning rate was set to 10^{-4} for the first 15 epochs and dropped to 10^{-5} for the remainder. The hyperparameters α_{dp} and α_{sg} in Equation (1) were empirically set to 10 and 1, respectively. The network was trained for 20 epochs using Adam optimizer [33] and the training took about 8 hours on a single NVIDIA 2080 Ti GPU.

V. RESULTS

Table I summarizes the SDSNET results as well as other depth estimation methods, using the mean and standard deviation (std.) of the SSIM index. The SDSNET outperforms the other methods. More specifically, it is 1.6% higher than Monodepth2 [29] and 12.4% higher than the Siamese architecture [27]. This is a significant improvement and interestingly, we achieve the best result when both the depth estimation branch and the segmentation branch were fused together, with the added benefit of surgical instrument segmentation included.

Table II summarizes the segmentation results using the IoU and Dice index. It can be seen that SDSNET is not only computationally efficient but also produces superior segmentation results 5.28% higher than U-Net [20] for IoU and 5.85% for Dice index. Table II also confirms that the use of a fusion operation when performing depth estimation and segmentation simultaneously can improve the segmentation result. Example qualitative results are presented in Fig 4, showing that SDSNET provides consistent depth estimation and accurate segmentation simultaneously.

Generalization To validate the generalization of our network, an additional experiment used the model trained on the $\mathcal{D}^{sia}$ dataset but tested directly on the $\mathcal{D}^{por}$ dataset, without retraining the whole network. Table III represents the results of SDSNET and Monodepth2 in this experiment using the SSIM index. Overall, SDSNET with fusion from both segmentation and depth estimation branch achieved higher SSIM index, confirming that the SDSNET generalizes well across different datasets, while still achieving competitive

TABLE III
SSIM SCORES ON THE $\mathcal{D}^{por}$ TEST SET

	Mean SSIM	Std.SSIM
Monodepth2 [29]	76.67	0.047
SDSNET with fusion (ours)	**77.53**	**0.041**

performance compared to the recent state-of-the-art methods.

VI. CONCLUSIONS

In this work we have presented SDSNET, a joint learning network that can simultaneously segment surgical tools and estimate the depth for each pixel. The proposed fusion network achieved state-of-the-art performance in both tasks. Besides, the framework does not require any depth labels and segmentation ground truth, and thus allows superior applicability on large-scale *in vivo* video processing where ground truth for per-pixel depth maps and manual segmentation labels are not easy to obtain.

ACKNOWLEDGMENT

This work was supported by the UK National Institute for Health Research (NIHR) Invention for Innovation Award NIHR200035, the Cancer Research UK Imperial Centre, the Royal Society (UF140290) and the NIHR Imperial Biomedical Research Centre.

REFERENCES

[1] D. Zhang, B. Xiao, B. Huang, L. Zhang, J. Liu, and G.-Z. Yang, "A self-adaptive motion scaling framework for surgical robot remote control," *RAL*, 2018.

[2] S. Giannarou, M. Ye, G. Gras, K. Leibrandt, H. J. Marcus, and G.-Z. Yang, "Vision-based deformation recovery for intraoperative force estimation of tool–tissue interaction for neurosurgery," *International journal of computer assisted radiology and surgery*, 2016.

[3] B. Huang, Y.-Y. Tsai, J. Cartucho, K. Vyas, D. Tuch, S. Giannarou, and D. S. Elson, "Tracking and visualization of the sensing area for a tethered laparoscopic gamma probe," *IJCARS*, 2020.

[4] B. Huang, J.-q. Zheng, A. Nguyen, S. Giannarou, D. Tuch, K. Vyas, and D. S. Elson, "Self-supervised generative adversarial network for depth estimation in laparoscopic images," in *MICCAI*, 2021.

[5] B. Huang, J.-Q. Zheng, S. Giannarou, and D. S. Elson, "H-net: Unsupervised attention-based stereo depth estimation leveraging epipolar geometry," *arXiv preprint arXiv:2104.11288*, 2021.

[6] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, 2017.

[7] D. Liu, Y. Wei, T. Jiang, Y. Wang, R. Miao, F. Shan, and Z. Li, "Unsupervised surgical instrument segmentation via anchor generation and semantic diffusion," in *MICCAI*, 2020.

[8] J. Zhan, J. Cartucho, and S. Giannarou, "Autonomous tissue scanning under free-form motion for intraoperative tissue characterisation," in *ICRA*, 2020.

[9] N. Zevallos, R. A. Srivatsan, H. Salman, L. Li, J. Qian, S. Saxena, M. Xu, K. Patath, and H. Choset, "A surgical system for automatic registration, stiffness mapping and dynamic image overlay," in *ISMR*, 2018.

[10] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *WACV*, 2019.

[11] X. Liu, A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath, "Dense depth estimation in monocular endoscopy with self-supervised learning methods," *TMI*, 2019.

[12] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, "Sfm-net: Learning of structure and motion from video," *arXiv preprint arXiv:1704.07804*, 2017.

[13] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017.

[14] A. Nguyen, D. Kundrat, G. Dagnino, W. Chi, M. E. Abdelaziz, Y. Guo, Y. Ma, T. M. Kwok, C. Riga, and G.-Z. Yang, "End-to-end real-time catheter segmentation with optical flow-guided warping during endovascular intervention," in *ICRA*, 2020.

[15] S. Speidel, M. Delles, C. Gutt, and R. Dillmann, "Tracking of instruments in minimally invasive surgery for surgical skill analysis," in *International Workshop on Medical Imaging and Virtual Reality*, 2006.

[16] Z. Pezzementi, S. Voros, and G. D. Hager, "Articulated object tracking by rendering consistent appearance parts," in *ICRA*, 2009.

[17] L. C. García-Peraza-Herrera, W. Li, C. Gruijthuijsen, A. Devreker, G. Attilakos, J. Deprest, E. Vander Poorten, D. Stoyanov, T. Vercauteren, and S. Ourselin, "Real-time segmentation of non-rigid surgical tools based on deep learning and tracking," in *International Workshop on Computer-Assisted and Robotic Endoscopy*, 2016.

[18] X. Song, X. Zhao, H. Hu, and L. Fang, "Edgestereo: A context integrated residual pyramid network for stereo matching," in *ACCV*, 2018.

[19] F. Liu, Y. Jonmohamadi, G. Maicas, A. K. Pandey, and G. Carneiro, "Self-supervised depth estimation to regularise semantic segmentation in knee arthroscopy," in *MICCAI*, 2020.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[21] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *arXiv preprint arXiv:1506.02025*, 2015.

[22] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *ICMLA*, 2018.

[23] "Miccai 2017 endoscopic vision challenge: Robotic instrument segmentation sub-challenge," 2017. [Online]. Available: https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/Data/

[24] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *ECCV*, 2016.

[25] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for neural networks for image processing," *arXiv preprint arXiv:1511.08861*, 2015.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[27] M. Ye, E. Johns, A. Handa, L. Zhang, P. Pratt, and G.-Z. Yang, "Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery," *arXiv preprint arXiv:1705.08260*, 2017.

[28] P. Mountney, D. Stoyanov, and G.-Z. Yang, "Three-dimensional tissue deformation recovery and tracking," *SPM*, 2010.

[29] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *ICCV*, 2019.

[30] K. Yamaguchi, D. McAllester, and R. Urtasun, "Efficient joint segmentation, occlusion labeling, stereo and flow estimation," in *ECCV*, 2014.

[31] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *ACCV*, 2010.

[32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.