Self-supervised Monocular Depth Estimation with 3D Displacement Module for Laparoscopic Images

Chi Xu[1], Baoru Huang[1], Daniel S. Elson[1]

***Abstract*—We present a novel self-supervised training framework with 3D displacement (3DD) module for accurately estimating per-pixel depth maps from single laparoscopic images. Recently, several self-supervised learning based monocular depth estimation models have achieved good results on the KITTI dataset, under the hypothesis that the camera is dynamic and the objects are stationary, however this hypothesis is often reversed in the surgical setting (laparoscope is stationary, the surgical instruments and tissues are dynamic). Therefore, a 3DD module is proposed to establish the relation between frames instead of ego-motion estimation. In the 3DD module, a convolutional neural network (CNN) analyses source and target frames to predict the 3D displacement of a 3D point cloud from a target frame to a source frame in the coordinates of the camera. Since it is difficult to constrain the depth displacement from two 2D images, a novel depth consistency module is proposed to maintain depth consistency between displacement-updated depth and model-estimated depth to constrain 3D displacement effectively. Our proposed method achieves remarkable performance for monocular depth estimation on the Hamlyn surgical dataset and acquired ground truth depth maps, outperforming monodepth, monodepth2 and packnet models.**

***Index Terms*—Deep learning, self-supervised learning, CNN, 3D displacement, monocular depth estimation.**

I. Introduction

MINIMALLY invasive surgery (MIS) is widely applied in general surgery because of the low trauma for patients [1]–[3]. Compared with traditional open surgery, MIS provides visualization of *in vivo* environments via laparoscopic vision. Since 2D laparoscopic images lack depth information that is available for naked eye 3D human perception and decision making, it may be useful to estimate accurate per-pixel depth maps from these images to reconstruct precise 3D tissues and internal scenes. This will not only provide the surgeon with a realistic surgical experience but also allow other image guidance technologies to be seamlessly incorporated into the procedure. Although depth can also be estimated for images from stereo laparoscopes using various methods, these are only available for certain procedures and locations, with monocular endoscopes remaining more popular [4].

Many monocular depth estimation methods have been proposed, such as monocular feature-based methods (e.g. Structure from Motion (SfM) [5]), supervised learning and self-supervised learning. Monocular feature-based methods utilize conventional feature extractors [6], [7] and feature matching methods to infer ego-motion matrix between frames and depth map, but it is difficult to carry out effective feature matching [4], [8] in some stereo laparoscopic images due to the low number of texture features.

Deep learning algorithms may be more robust for *in vivo* environments, and self-supervised learning models have become popular because it is challenging to acquire ground truth depth maps in real world settings, especially in MIS. Self-supervised learning-based approaches take synthetic target images as the supervisory signal to train the depth estimation model. Stereo training and monocular training are two existing frameworks for self-supervised monocular depth estimation. The monocular training utilizes an additional network to predict the ego-motion matrix between frames to synthesize target images from adjacent frames under the hypothesis of moving camera and static scene [4], [8]–[12]. The stereo training makes use of the geometrical relation between rectified stereo images to infer a dense stereo disparity map whereby the left/right image can be reconstructed by horizontally shifting pixels of the right/left image [13], [14]. In surgery, the position of the laparoscope is often static and the scenes are dynamic (moving surgical instruments and deforming tissues), leading to the identity matrix (ego-motion matrix) and many pixels remaining stationary. Therefore, it is necessary to implement a new module to establish a relationship between frames in monocular training and utilize stereo view synthesis of stereo training for image synthesis.

In this paper, we propose a new self-supervised learning based framework for monocular depth estimation in laparoscopic imaging. Three contributions are achieved: (1) A 3-branch Siamese network was designed to enhance the interaction between adjacent frames during training, improving the performance of the depth estimation model; (2) A 3DD module was formulated to estimate the per-pixel 3D displacement map of 3D point clouds between adjacent frames, establishing a novel relationship between adjacent frames. This module replaces the conventional ego-motion module and matches the surgical scenario well; (3) The depth consistency loss and monocular appearance loss were used to train the 3DD network (3DD-Net).

II. Methods

The overall training framework is depicted in Fig. 1 and several key ideas are introduced in the following sections together with the three loss functions for training the depth estimation model and 3DD-Net.

Framework Architecture A 3-branch Siamese network - composed of three identical and weight-sharing auto-encoder networks corresponding to target (I_t) and source ($I_{t'}$: I_{t-1} and

[1]{chi.xu20, baoru.huang18, daniel.elson}@imperial.ac.uk
[1]The Hamlyn Centre for Robotic Surgery, Department of Surgery and Cancer, Imperial College London, London SW7 2AZ, UK

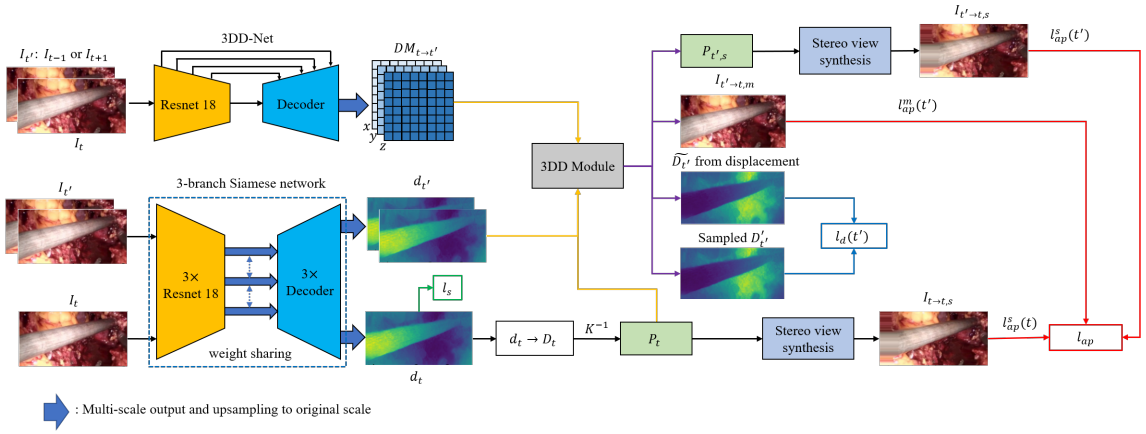

Fig. 1. Framework architecture. The Resnet 18 [15] is pre-trained. The dark blue arrow indicates bilinear interpolation from multi-scale outputs to original scaled outputs. The colored lines are used to indicate correspondence between output data and loss function (red for l_{ap}, blue for l_d, green for l_s).

I_{t+1}) frames respectively - was used to predict dense depth maps (simplified in Fig. 1). It was tested using a single auto-encoder from the 3-branch Siamese network. The 3DD-Net took a stack of I_t and $I_{t'}$ as inputs, with an output in the form of a 3-channel tensor of the same spatial dimension as the input data. The output is called as *3D displacement map*, describing the 3-dimensional displacement (x, y and z directions) of each 3D point cloud between two adjacent frames. To generate a self-supervisory signal, the 3DD module (described in next section) and *view synthesis* were used to describe the 3D displacement of the 3D point cloud between adjacent frames and reconstruct the target frame respectively. Finally, three loss functions were used to train the depth estimation model and 3DD-Net (details below).

3D Displacement Module: The inputs of the 3DD module were the predicted disparity map of the source frame ($d_{t'}$), the 3D point cloud of the target frame (P_t) and the 3DD map ($DM_{t\to t'}$). The 3DD module not only modified the 3D point cloud for stereo view synthesis, but also generated a depth consistency loss and monocular appearance loss to enable the 3DD-Net to learn and limit the 3D displacement. As shown in Fig. 2, the $DM_{t\to t'}$ changed the 3D point cloud from the target frame to the source frame to generate a depth map from displacement ($\widetilde{D_{t'}}$). The sampled depth map ($D'_{t'}$) could be generated by the following formula:

$$D'_{t'} = D_{t'} \langle proj(P_t, DM_{t\to t'}, K) \rangle \tag{1}$$

Here $D_{t'}$ is the predicted depth map of the source frame; $proj()$ is the 2D projecting function to generate 2D sampling coordinates and K is the pre-computed intrinsic matrix of the camera; $\langle\cdot\rangle$ is the sampling operation. $D_{t'}$ and $\widetilde{D_{t'}}$ are compared to limit the displacement in the z direction and maintain the depth consistency between adjacent frames. The 2D sampling coordinates were also applied to monocular view synthesis and the synthetic image limited the 2D displacement in the image plane (x and y dimension) by using the monocular appearance loss.

View Synthesis: In our framework, there are two view synthesis processes: monocular and stereo. The monocular view synthesis reconstructs images in the same coordinates as the

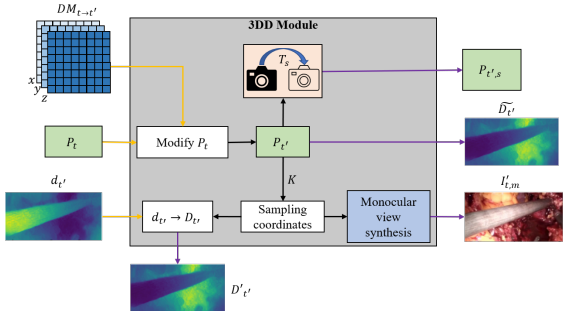

Fig. 2. The 3DD module architecture. The orange and purple lines represent the inputs and outputs respectively.

camera. Such reconstructed images are used to constrain the displacement of x and y dimensions. Due to the laparoscope remaining stationary, most pixels captured by the camera have no disparity between adjacent frames, which makes depth estimation difficult to learn from the appearance loss during training and results in a high number of infinite depth predictions during testing. To solve this, stereo view synthesis sampled pixel values from the other image of the stereo pair for training. Two view synthesis operations are described in formulas (2) and (3).

$$I'_{t'\to t,m} = I_{t',m} \langle proj(P_t, DM_{t\to t'}, K) \rangle \tag{2}$$

$$I'_{t'\to t,s} = I_{t',s} \langle proj(P_t, DM_{t\to t'}, T_s, K) \rangle \tag{3}$$

Where $I'_{t'\to t,m}$ and $I'_{t'\to t,s}$ are target images reconstructed by monocular view synthesis and stereo view synthesis respectively; $I_{t',m}$ and $I_{t',s}$ are the captured source images for monocular and stereo view synthesis respectively; T_s is the pre-computed extrinsic matrix, changing the coordinates of the 3D point cloud for stereo view synthesis (details in Fig. 2).

View-field Mask: For stereo view synthesis, pixels from the leftmost region of the left image were not sampled because they were out of view for the right camera. The appearance loss from such regions caused degradation [10] and must be masked. Further, the view-field mask prevented the 3DD-Net

generating abnormal displacement in the z direction. Therefore, a mask excluded such regions and was generated from 2D sampling coordinates in stereo view synthesis [9], [10], [16]. When the depth estimation model predicted depth maps of input frames (D_t), P_t could be generated by multiplying by K^{-1}. Then, the reference coordinates of P_t were changed from the target frame to the source frame by T_s. Finally, the 2D sampling coordinate could be computed by multiplying by K. In the 2D sampling coordinates, the coordinates of effective pixels were between -1 and 1 and other values represented pixels that are out of the field of view. Therefore, the mask (M) could be generated by the following formulae:

$$\text{coord} = proj(D_t, K^{-1}, T_s, K) \tag{4}$$

$$M = \begin{cases} 1 & \text{if coord}(i,j,:) \in [-1,1] \\ 0 & \text{else} \end{cases} \tag{5}$$

Loss Function In this section, we propose a loss function to retain depth consistency between adjacent frames and efficiently constrain 3D displacement.

Appearance Loss: The appearance loss was composed of a monocular and stereo appearance loss. Both loss functions combined SSIM loss [17] and L1 loss [18] in a specific proportion [13]. The monocular appearance loss was generated by comparing the target image with an image reconstructed by monocular view synthesis, which aimed to constrain the displacement to the x and y dimensions. Inspired by [11], per-pixel minimum loss was applied to the monocular appearance loss to better handle occlusions caused by moving surgical instruments. The stereo appearance loss compared the target and reconstructed images, enabling the depth estimation model to learn depth from frames and stereo pairs. In this framework, the $\widehat{I'}_{t'\to t}$ and $\widehat{I}_t$ were images that only contained overlapping regions of stereo pairs (masked out by M). The α was equal to 0.85.

$$l^s_{ap}(t') = \tfrac{\alpha}{2}(1 - SSIM(\widehat{I'}_{t'\to t,s}, \widehat{I}_t)) + (1-\alpha)\left\|\widehat{I'}_{t'\to t,s} - \widehat{I}_t\right\| \tag{6}$$

$$l^m_{ap} = \min_{t'} \tfrac{\alpha}{2}(1 - SSIM(\widehat{I'}_{t'\to t,m}, \widehat{I}_t)) + (1-\alpha)\left\|\widehat{I'}_{t'\to t,m} - \widehat{I}_t\right\| \tag{7}$$

Depth Consistency Loss: Inspired by Godard *et al.*'s left-right consistency [13], we propose the depth consistency loss to constrain the displacement in the z dimension and maintain the depth consistency between frames. To balance the contributions of depth consistency loss and appearance loss, the normalized loss was used [8].

$$l_d(t') = \frac{\sum(M \cdot (D'_{t'} - \widetilde{D_{t'}})^2)}{\sum(M \cdot (D'^2_{t'} + \widetilde{D^2_{t'}}))} \tag{8}$$

Edge-aware Smoothness Loss: The edge-aware smoothness loss is widely used in loss function for depth estimation to reduce the noisy depth values, except values at edges.

$$l_s = \left|\partial_x d^*_t\right| e^{-|\partial_x I_t|} + \left|\partial_y d^*_t\right| e^{-|\partial_y I_t|} \tag{9}$$

Overall Loss: The overall loss is shown in equation (10). The F is the set of source frames: $\{t-1, t+1, s\}$. Considering that the depth consistency loss was added to the loss function, the weight λ for smoothness loss was increased to 0.002 to retain the smoothness contribution.

$$L = \frac{\sum_{t'\in F} l^s_{ap}(t') + l^m_{ap}}{4} + \frac{\sum_{t'\in\{t-1,t+1\}} l_d(t')}{2} + \lambda \cdot l_s \tag{10}$$

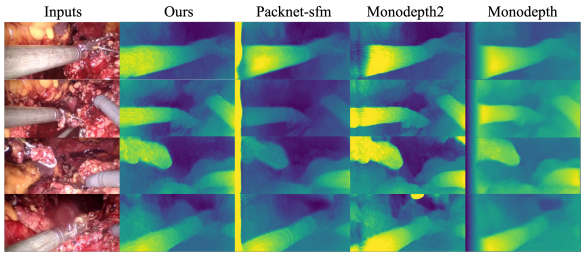

Fig. 3. Qualitative result comparison between our method, packnet [12], monodepth2 [11], monodepth [13]. The first column contains example test images. The other columns are the corresponding disparity maps.

III. Experimental Setup and Results

Experiment Setup The Hamlyn surgical dataset [19] was used to train and evaluate the models, containing 192 × 382 rectified stereo laparoscopic image pairs, 34240 pairs for training and 7191 pairs for validation. The stereo camera had the same intrinsic matrix for all laparoscopic images. The stereo pairs were rectified, therefore the extrinsic matrix between stereo camera was a horizontal translation with a fixed distance. Furthermore, to make the experiment more convincing, 100 laparoscopic images with ground truth depth maps acquired by projected gray-code structured lighting and were collected and analysed, as shown in Fig. 4 [20].

The model was implemented in Pytorch [21]. For training, optimizer Adam was used in 15 epochs with a batch size of 12 and a learning rate set to 0.0001. The training took 12 hours with a single 16GB NVIDIA Tesla P100. The monodepth [13], monodepth2 [11] and packnet [12] models were implemented for comparison. Considering that monodepth2 achieved the same performance in both monocular and stereo training, only stereo training was applied for monodepth2 [11] and packnet [12] to make the comparison fair.

Comparison Study In this section, we implemented three self-supervised models for comparison. The SSIM based on the Hamlyn surgical dataset [19] and the metrics based on acquired ground truth depth maps were taken as criteria to evaluate the predicted depth maps, as shown in Table I. Further, the qualitative comparison was also conducted between models, as depicted in Fig. 3.

For testing, we selected one branch of the three-branch Siamese network as the depth estimation model. Compared with the other models, our model performed better and had the fewest parameters. The qualitative result comparison shows that the artifacts were significantly reduced in our model, including border artifacts (appearing at regions of source frames not visible in both images) and texture-copy artifacts (caused by incorrect translation from input images). The border artifacts were mainly removed by view-field masking (in Fig. 5) and the texture-copy artifacts were reduced by the proposed depth consistency loss (in Fig. 3).

TABLE I
EVALUATION BASED ON SSIM AND GROUND TRUTH

	μ_{ssim}	σ_{ssim}	μ_{ABE}	σ_{ABE}	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Params
Monodepth	0.5843	0.1140	5.741	3.631	0.295	95.195	56.887	0.320	0.589	0.875	0.962	20M
PackNet	0.7380	0.0698	3.893	1.584	0.194	3.857	14.507	0.311	0.660	0.922	0.975	120M
Monodepth2	0.7199	0.0826	3.152	1.009	0.160	2.281	11.345	0.206	0.730	0.976	0.997	**14M**
Ours	**0.7421**	**0.0641**	**2.684**	**0.913**	**0.136**	**1.758**	**9.829**	**0.165**	**0.818**	**0.991**	**1.000**	**14M**

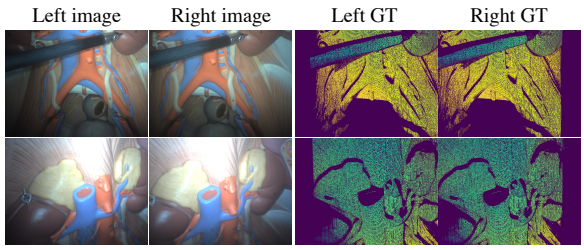

Fig. 4. The acquired ground truth depth maps via da Vinci (Intuitive Inc.) stereo laparoscope and projected gray-code structured light pattern [20].

TABLE II
ABLATION STUDY

	3DD Module	DCL	Abs Rel	Sq Rel	RMSE
Baseline			0.160	2.281	11.345
Baseline	✓		0.154	2.137	10.992
Siamese-net	✓	✓	**0.136**	**1.758**	**9.829**

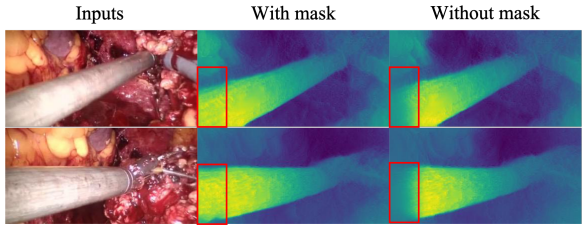

Fig. 5. The effect of view-field masking is shown in red boxes.

Ablation Study In order to study the contribution of 3DD module and depth consistency loss (DCL), an ablation study with acquire ground truth depth maps was conducted. The monodepth2 was set as the baseline model. As shown in Table II, the baseline model trained with 3DD module had improved performance. When the baseline model was replaced by 3-branch Siamese network with DCL, the improvement was significant.

IV. CONCLUSION

We proposed a novel self-supervised framework for monocular depth estimation, achieving state-of-the-art performance not only for the Hamlyn surgical dataset [19], but also for a newly acquired dataset with ground truth. We modified the conventional single auto-encoder network with a 3-branch Siamese network for training, enforcing the interaction between adjacent frames. The 3DD module also significantly improved the model performance via depth consistency and monocular appearance losses.

ACKNOWLEDGMENTS

This work was carried out with support from the UK National Institute for Health Research (NIHR) Invention for Innovation Award NIHR200035, the Cancer Research UK Imperial Centre and the NIHR Imperial Biomedical Research Centre.

REFERENCES

[1] K. Fuchs, "Minimally invasive surgery," *Endoscopy*, 2002.
[2] V. Desiato, M. Melis, B. Amato, T. Bianco, A. Rocca, M. Amato, G. Quarto, and G. Benassai, "Minimally invasive radioguided parathyroid surgery: A literature review," *IJS*, 2016.
[3] E. P. Westebring-van der Putten, R. H. Goossens, J. J. Jakimowicz, and J. Dankelman, "Haptics in minimally invasive surgery–a review," *Minimally Invasive Therapy & Allied Technologies*, 2008.
[4] L. Li, X. Li, S. Yang, S. Ding, A. Jolfaei, and X. Zheng, "Unsupervised learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery," *TII*, 2020.
[5] S. Leonard, A. Sinha, A. Reiter, M. Ishii, G. L. Gallia, R. H. Taylor, and G. D. Hager, "Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data," *TMI*, 2018.
[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.
[7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*. Ieee, 2011.
[8] 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.
[9] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, 2017, pp. 1851–1858.
[10] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *CVPR*, 2018.
[11] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *ICCV*, 2019.
[12] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *CVPR*, 2020.
[13] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017.
[14] 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.
[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[16] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *ECCV*, 2016.
[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *TIP*, 2004.
[18] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Is l2 a good loss function for neural networks for image processing? arxiv preprint," *arXiv preprint arXiv:1511.08861*, 2015.
[19] 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.
[20] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *CVPR*, 2003.
[21] 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.