Unsupervised learning of endoscopy video frames' correspondences from global and local transformation

Mohamad Ali Armin[1,2], Nick Barnes[1,4], Salman Khan[1],Miaomiao Liu[1],
Florian Grimpen[3],Olivier Salvado[2]

[1]CSIRO (Data61), Canberra, Australia
m.a.armin@gmail.com
[2] Biomedical Informatics Group, Brisbane, Australia
[3]Department of Gastroenterology and Hepatology, Royal Brisbane and Women's Hospital
[4]College of Engineering and Computer Science (ANU)
Olivier.Salvado@csiro.au

Abstract. Inferring the correspondences between consecutive video frames with high accuracy is essential for many medical image processing and computer vision tasks (e.g. image mosaicking, 3D scene reconstruction). Image correspondences can be computed by feature extraction and matching algorithms, which are computationally expensive and are challenged by low texture frames. Convolutional neural networks (CNN) can estimate dense image correspondences with high accuracy, but lack of labeled data especially in medical imaging does not allow end-to-end supervised training. In this paper, we present an unsupervised learning method to estimate dense image correspondences (DIC) between endoscopy frames by developing a new CNN model, called the EndoRegNet. Our proposed network has three distinguishing aspects: a local DIC estimator, a polynomial image transformer which regularizes local correspondences and a visibility mask which refines image correspondences. The EndoRegNet was trained on a mix of simulated and real endoscopy video frames, while its performance was evaluated on real endoscopy frames. We compared the results of EndoRegNet with traditional feature-based image registration. Our results show that EndoRegNet can provide faster and more accurate image correspondences estimation. It can also effectively deal with deformations and occlusions which are common in endoscopy video frames without requiring any labeled data.

Keywords: convolutional neural network, unsupervised learning, image correspondences, registration

1 Introduction

Estimating image correspondences is the base of many medical image processing and computer vision algorithms. Traditional methods such as SIFT [1] or KLT [2] have shown remarkable results in estimating image correspondences and registering endoscopy frames [3, 4], yet they are computational expensive, may fail for frames with sparse textures, and become unreliable when objects deform (one example of correspondences estimation by SIFT feature tracking [5], SIFT flow [1] and our method (EndoRegNet) is shown if Fig.1).

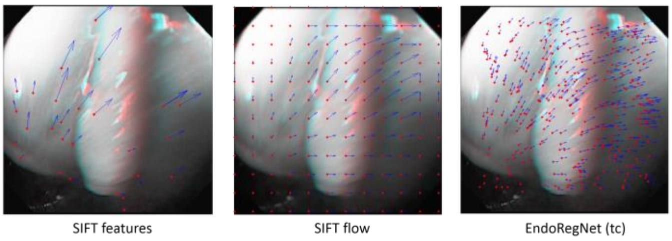

Fig. 1. Example of correspondences estimation by the SIFT feature tracker, SIFT flow, and our proposed method (EndoRegNet) from consecutive colonoscopy frames, frames are overlaid, SIFT flow and EndoRegNet are shown sparsely for better visualization of the motion.

In recent years, methods based on deep Convolutional Neural Networks (CNN) have been shown to be accurate in image correspondence estimation. Ji et al. [6] developed a deep view morphing network that can predict the middle view and image correspondences between two frames. Fischer et al. proposed FlowNet [7] which can predict dense motion flow between two frames. However, these methods need a large amount of labeled data for training and testing, which hamper performance when not available because it is very difficult to generate a ground-truth for correspondences of endoscopy images (even when using a simulator). The lack of ground-truth to allow end-to-end network training, especially in medical imaging, has increased the popularity of unsupervised or semi-supervised CNNs. For instance, Zhou et al. [8] and Garg et al. [9] have estimated depth, and Yin and Shi [10] estimated depth, camera pose and optical flow from images without using labeled data. Meister el al. [11] and Wang et al. [12] however, focused mainly on unsupervised flow estimation by estimating back and forth motion using FlowNet architecture and introducing an loss function to deal with occlusion. Although, they have shown remarkable results in comparison to supervised methods (e.g. FlowNet), for a more challenging dataset such as Sintel [13] which include deformation and occlusion, their method cannot outperform supervised methods, and needs improvements. Besides, using FlowNetS as the base of their network structure means a requirement of a huge dataset for training. In our method, we tackled deformation by learning parameters of a global polynomial transformation between consecutive frames, and inspired by deep view morphing [6] we developed a CNN that can be trained with smaller dataset. In medical imaging, De vos et al. [14] registered cardiac MRI images through implementing a cubic B-spline transformer and spatial transformer network [15]. Although their method can deal with deformable MRI images, it cannot handle occlusion, which is common in colonoscopy images.

In this paper, we propose a novel CNN architecture to predict correspondences of deformable, sparse texture endoscopy images through image registration while being robust to occluded areas. Our method does not require labeled data. We achieved this by developing a network comprising three components: (i) a Dense Image Correspondences (DIC) sub-network that predicts pixel displacement between two frames as (dx,dy) and allows local deformation; (ii) a Polynomial Transformer Parameters (PTP) sub-network, which estimates polynomial parameters between two frames and can produce a global motion flow which is used to regularize the output of the DIC network; (iii) and a Visibility Mask (VM) sub-network, which predicts occluded areas in the second frame. The output of the dense image correspondences and the polynomial subnetwork are the input to a bilinear image transformer which transforms the second image to the first one. The loss function is computed as absolute difference between first image I_1 and a transformation of second image I_2 to I_1 based on both motion and polynomial transformation estimated by the DIC and PTP networks, along with absolute difference between correspondences obtained by the PTP and DIC network. Since our model performs image registration for endoscopy, we call our network EndoRegNet. The EndoRegNet is unsupervised and there is no need for any labeled data for training. We train the network with both simulated and real colonoscopy video frames. Our results show excellent performance in image registration of colonoscopy frames that are non-rigid and have sparse texture. Further, EndoRegNet can be used to register any endoscopy video frames, or indeed other non-rigid scenes. We test EndoRegNet on vivo datasets [16, 17]. The key contributions of the EndoRegNet can be summarized as (i) using a polynomial transformation to regularize local pixel displacement (a polynomial transformation unlike affine transformation can model deformation between two frames, which is a main difference between our method and other unsupervised method such as [11]) ; (ii) dealing with deformation by using absolute pixel-by-pixel transformations regularized by a polynomial transformation; (iii) refining image correspondences for occluded areas by calculating a visibility mask. We could obtain good results by training our network even on a small medical image dataset. The overview of our method is shown in Fig.2.

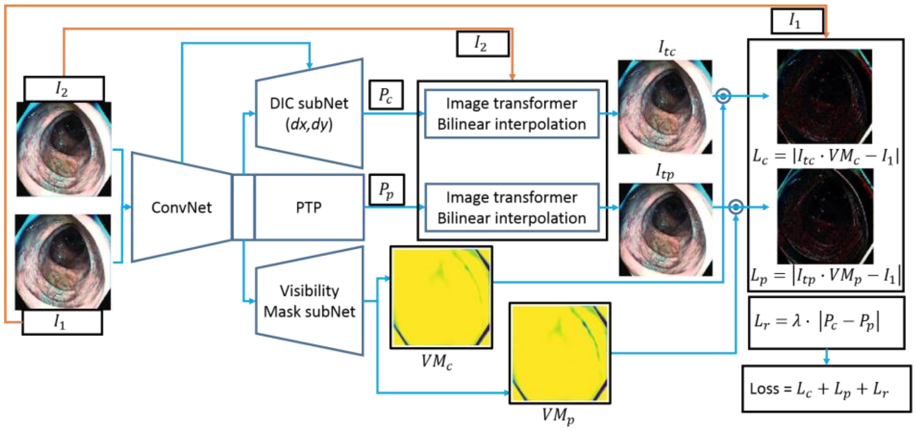

Fig. 2. The endoscopy image registration network (EndoRegNet). DIC and PTP are dense image correspondences and polynomial transformer parameters sub-network, $P_c(x_{c2}, y_{c2})$ and $P_p(x_{p2}, y_{p2})$ are image correspondences estimated by DIC and PTP.

2 Method

Our goal is to register colonoscopy frames and estimate dense image correspondences between consecutive frames through image registration. This can be performed by estimating pixel displacement between two frames, however a network that only estimates pixel displacement can result in outliers and consequently a poor image registration. Here we introduce a new approach to address this through regularizing local pixel displacement by estimating a global transformation. In this paper, we introduce a polynomial function of second order (as it can deal with deformations) to determine the global transformation between two frames. Colonoscopy frames include haustral folds which lead to occlusions, so a visibility mask similar to [6] is also included in the model to improve registration performance by omitting occluded areas. The EndoRegNet is introduced in the following.

2.1 Dense image correspondence (DIC) sub-network

Image correspondences or the dense flow field between two consecutive frames (I_1, I_2) can be estimated as a relative offset of (dx, dy) for each point pair. Each pair of points from I_1 as target image $P(x_1, y_1)$ can be mapped to source image point $P_c(x_{c2}, y_{c2})$ through:

$$x_{c2} = x_1 + dx \ , \ y_{c2} = y_1 + dy \qquad (1)$$

Our DIC sub-network accepts two consecutive images as input, and estimates pixel displacement (dx, dy) for each pixel. By finding the mapping relation between I_1 and I_2 from Equation (1), bilinear sampling which is explained in [15] can be used to generate a transformed image I_{tc} which is a transformation of I_2 onto I_1. The DIC sub-network minimizes the L_1norm; the absolute difference between I_{tc} and I_1, known as photometric loss, which has been used in unsupervised view synthesis algorithms (e.g. [18]): $L_c = \mid I_{tc} - I_1 \mid$.

2.2 Polynomial transformation parameters (PTP)

Similarly to the view synthesis approach, if we only use DIC, we will be highly subject to outliers where individual point pairs have better matches on photometric loss but that are not consistent with their local regions. Here, we introduce a polynomial transformation to regularize the motion of images points between I_1 and I_2. We map a set of grid points $P(x_1, y_1)$ which indicate pixel position in a target image I_1 to a source image I_2 points $P_p(x_{p2}, y_{p2})$ by finding second degree polynomial transformation coefficients (θ_{ij}) between them as $P_p = \theta_{ij} \cdot P$ and can be extended as follows:

$$\begin{bmatrix} x_{p2} \\ y_{p2} \end{bmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} & \theta_{15} & \theta_{16} \\ \theta_{21} & \theta_{22} & \theta_{23} & \theta_{24} & \theta_{25} & \theta_{26} \end{bmatrix} \cdot [\, x_1 \quad y_1 \quad x_1 y_1 \quad {x_1}^2 \quad {y_1}^2 \quad 1]^t \qquad (2)$$

Here, $P_p$determines where to sample pixels from I_2 to obtain transformed image I_{tp} which is a transformation of I_2 onto I_1. The PTP sub-network estimates polynomial coefficients θ_{ij} by minimizing a photometric loss similar to DIC sub-network: $L_p = \left| I_{tp} - I_1 \right|$. Again we incorporate bilinear sampling [15] in a similar manner to DIC to infer I_{tp}.

2.3 Visibility mask (VM) sub-network

Colonoscopy frames include haustral folds which cause occlusions. This occlusion prevents a full view of next frame and therefore increases the number of outliers between two consecutive frames. The effect of occlusion has been reduced by determining the visible area between two frames through a visibility mask (VM) [6, 19]. The last layer of VM sub-network has a sigmoid function that assigns one for existing correspondences and zero when correspondences are not found by the DIC sub-network or PTP. We modify the L_c and L_p to learn VM_c and VM_p which are the visibility masks for the DIC and PTP respectively:

$$L_c = |I_{tc} \cdot VM_c - I_1| , \quad L_p = \left|I_{tp} \cdot VM_p - I_1\right| \tag{3}$$

2.4 Regularized DIC and final objective function

To regularize local pixel displacement estimated by the DIC, we reduce the absolute difference between global positions estimated by the PTP sub-network P_p and local position estimated by the DIC sub-network P_c as $L_r = \lambda \cdot \left|P_c - P_p\right|$. Here λ is a weight, and empirically $\lambda = 0.9$ shows good results.

In general, the objective function for whole network can be calculated as sum of L_c and L_p which are estimated from equation 3 and L_r as a regularization term:

$$Loss = L_c + L_p + L_r \tag{4}$$

2.5 Architecture and training details

The first part of EndoRegNet consists of 6 convolutional layers which are shared among other sub-networks. EndoRegNet takes two consecutive RGB frames as input of size 224×224 pixels. PTP consists of three convolution layers followed by a fully connected layer to estimate θ_{ij}. The DIC sub-network is formed by three convolutional layers, and five de-convolutional layers. The VM sub-network has six de-convolutional layers and its last layer is a convolutional layer with a sigmoid activation function. The EndoRegNet architecture is shown in Fig.2.

The whole network was implemented and trained using the GPU version of Tensorflow [20]. We used ADAM solver [21] with the initial learning rate of 0.0001, β_1 and β_2 were 0.9 and 0.999 respectively. We used multi-GPU (Nvidia). Our network began to converge after 150,000 iterations.

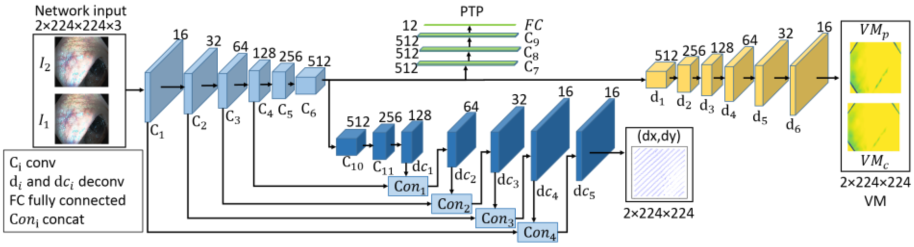

Fig. 3. The EndoRegNet architecture

3 Dataset

Simulated and real colonoscopy frames. Our dataset includes 29,000 pairs of frames which were extracted from simulated and real colonoscopy videos. The simulated frames were generated by a simulator described in [22]. The simulations were of ten different colons, and formed 72% of the data. The real frames extracted from six colonoscopy videos (six different patients). A 190HD Olympus endoscope was used to perform real colonoscopy procedures, which could capture 50 frame/sec (frame size was 1352×1080 pixels). We only used the informative frames for training and validation and removed uninformative frames (e.g. out of focus frames or blurry or those close to the colon wall) [23] from our computation.

Real colonoscopy frames. We used a colonoscopy dataset from Hamlyn Center Laparoscopic (HCL) [24] to validate the generalization performance of our trained network. The video frames were captured either by Olympus NBI endoscope, or a Pentax i-scan endoscope [17]. From HCL colonoscopy videos, the video number 10 (vn10) has been chosen for our test as it contained 1250 pairs of consecutive frames. 25% of these frames were uninformative and ignored in our experiments.

Laparoscopy video frames. In addition to the above, we trained the EndoRegNet with 80% of two set of laparoscopic in vivo video frames [16]. The first set contained 1220 pairs of stereo video frame, and the second set contained 5626 consecutive frames with deformation due to tools interaction.

4 Experiments and results

EndoRegNet was trained with 80% of our colonoscopy data, which was a mix of real and simulated colonoscopy frames (46476 frames). The trained EndoRegNet was then validated on real colonoscopy test data by computing mean absolute difference (MAD) and structural similarity index map (SSIM) (please see [25]) between I_1 and resgitered image. Note that we used default parameters for SSIM as stated in original paper [25]. Examples of SSIMs are presented in Fig. 4 (a) and results as the mean of SSIM and MAD are reported in Fig.6. We evaluated the performance of our trained network on real colonoscopy video frames vn10 which were obtained from [24] (b.1,b.2) in Fig.4. The results are presented in Fig.6.

We trained each set of laparoscopy video frames with the pre-trained EndoRegNet. 80% of data was used for training. Examples of stereo pairs and tool interaction are shown in Fig.4. (c,d) and Fig.5.

In addition, we compared the results of our network with traditional image registration using polynomial transformation and SIFT flow. The correspondences were estimated by using SIFT features explained in [5]. Results are reported in Fig.6. Note that the test set has not been used in training phase and for the sake of comparison we did not apply visibility masks on registered images obtained by EndoRegNet.

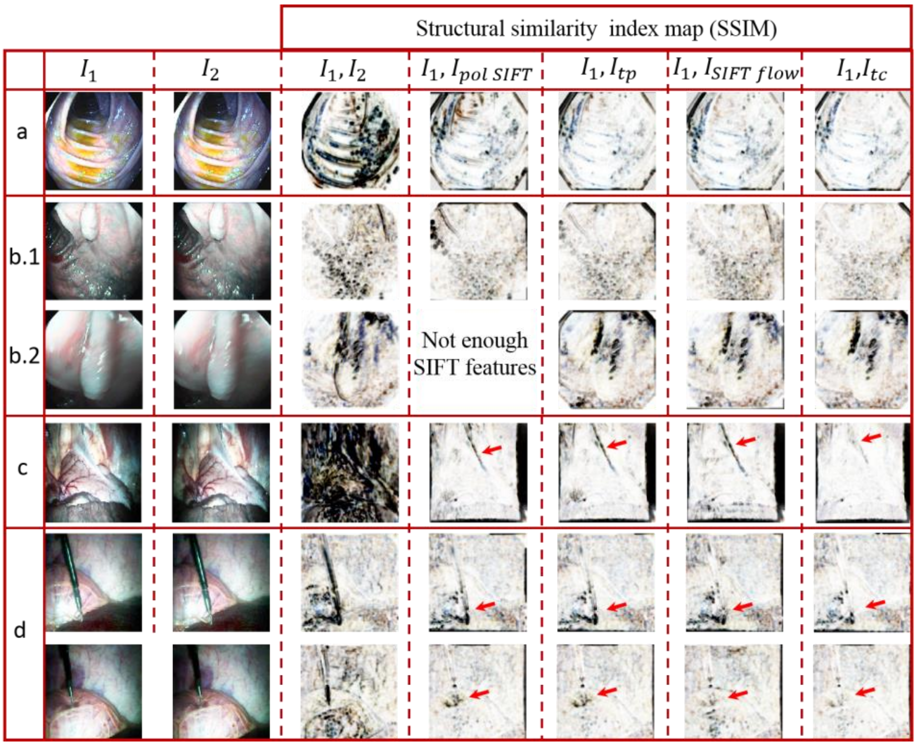

Fig. 4. Examples of images and SSIM between I_1, I_2 and I_1 and registered images by traditional feature-based method polynomial ($\boldsymbol{I_{pol}}$) transformation when SIFT is used as feature detector, EndoRegNet PTP ($\boldsymbol{I_{tp}}$), SIFT flow ($\boldsymbol{I_{SIFT\ flow}}$), and DIC ($\boldsymbol{I_{tc}}$). Real colonoscopy from our dataset (a), colonoscopy frames from Hamlyn (vn10) [17] (b.1,b.2), laparoscopy frame [16] (c), laparoscopy frame when tool interacts with organs and results in deformations (d). The red arrows show areas with deformation. Note that higher similarity leads to brighter area.

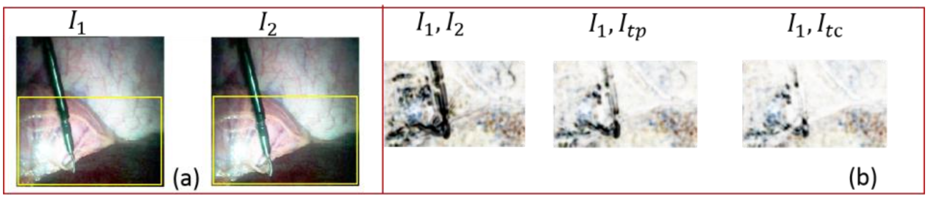

Fig. 5. Sample of deformed endoscopy sequences, two consecutive frames when a tool interacts with organ (deformed region is cropped for better perception, yellow rectangle) (a), SSIM between I_1, I_2 and I_1 and registered image with EndoRegNet (b).

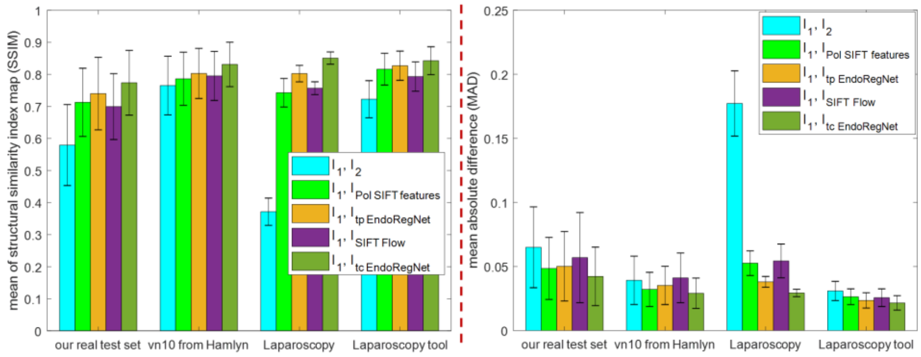

Fig. 6. The mean of SSIM and MAD error of different image registration method including polynomial ($\boldsymbol{I_{pol}}$) transforms when SIFT is used as feature detector, EndoRegNet PTP ($\boldsymbol{I_{tp}}$), SIFT flow, and DIC ($\boldsymbol{I_{tc}}$) over endoscopy frames. Our real colonoscopy test set, vn10 from Hamlyn [17], laparoscopy test set [16] , and deformed laparoscopy test set. Higher the SSIM and lower the MAD is better.

5 Discussion and Conclusion

In this paper, we present an unsupervised method to register deformable endoscopy video frames and estimate their correspondences. This is achieved by introducing a novel CNN model, called EndoRegNet, which has three main parts; (i) a dense image correspondences (DIC) sub-network, which estimates local displacement of pixels; (ii) polynomial transformation parameters (PTP) estimator, which is used to regularizes correspondences estimated by DIC, it can also deal with global deformations; (iii) and a visibility mask VM sub-network, which can refine correspondences in case of an occlusion (this is very common in colonoscopy video frames).

We trained all parts of EndoRegNet at the same time. At the test time, only DIC and VM could be used to predict correspondences between two consecutive frames and refine them. The results of EndoRegNet were compared with feature-based image registration for different set of endoscopy video frames. Our results presented in Fig.6. show high performance of EndoRegNet and its ability to generalize to new datasets. Note that we trained EndoRegNet on a training set and then evaluated its performance on data that has not been observed in the training phase by computing SSIM and MAD.

Further, EndoRegNet showed excellent performance in registering deformed sequences (e.g. Fig.5). As shown in Fig.5 (b) warping functions such as polynomial are inadequate to deal with the deformed images. We used a combination of local pixel displacement DIC and a second degree polynomial transformation PTP to deal with deformation. Particularly in Fig. (4) (b,d) it can be seen that some local strong deformation artefacts are better handled by the combination.
Other unsupervised flow estimation methods introduced by Meister el al. [11] and Wang et al. [12] are using FlowNet architecture but they have over 150 million parameters and thus require a huge training dataset. This is not feasible for our application. Instead, our proposed method provides excellent performance without requiring a large training data. We plan to improve our deformation model by using different objective function and convolution layers to better model long displacement and deformation.

Acknowledgement. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

1. Liu, C., Yuen, J., Torralba, A., Sivic, J., Freeman, W.T.: SIFT Flow: Dense Correspondence across Different Scenes. In: Forsyth, D., Torr, P., and Zisserman, A. (eds.) Computer Vision – ECCV 2008. pp. 28–42. Springer Berlin Heidelberg, Berlin, Heidelberg.
2. Jianbo Shi, Tomasi Carlo: Good features to track. Presented at the Computer Vision and Patern Recognition , Seattle, WA (1994).
3. Armin, M.A., Chetty, G., De Visser, H., Dumas, C., Grimpen, F., Salvado, O.: Automated visibility map of the internal colon surface from colonoscopy video. Int. J. Comput. Assist. Radiol. Surg. 11, 1599–1610 (2016).
4. Bell, C.S., Puerto, G.A., Mariottini, G.-L., Valdastri, P.: Six DOF motion estimation for teleoperated flexible endoscopes using optical flow: A comparative study. Presented at the May (2014).
5. Puerto-Souza, G.A., Mariottini, G.L.: Hierarchical Multi-Affine (HMA) algorithm for fast and accurate feature matching in minimally-invasive surgical images. Presented at the October (2012).
6. Ji, D., Kwon, J., McFarland, M., Savarese, S.: Deep View Morphing. CVPR 2017. (2017).
7. Dosovitskiy, A., Fischery, P., Ilg, E., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 2758–2766. IEEE (2015).
8. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised Learning of Depth and Ego-Motion from Video. Presented at the July (2017).
9. Garg, R., B.G., V.K., Carneiro, G., Reid, I.: Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. In: Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 740–756. Cham (2016).
10. Yin, Z., Shi, J.: GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. CVPR. (2018).

11. Meister, S., Hur, J., Roth, S.: UnFlow: Unsupervised Learning of Optical Flow with a Bidirectional Census Loss. AAAI. (2018).
12. Wang, Y., Yang, Y., Yang, Z., Zhao, L., Xu, W.: Occlusion Aware Unsupervised Learning of Optical Flow. CVPR. (2018).
13. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A Naturalistic Open Source Movie for Optical Flow Evaluation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 611–625. Springer (2012).
14. de Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I.: End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network. In: Cardoso, M.J., Arbel, T., Carneiro, G., Syeda-Mahmood, T., Tavares, J.M.R.S., Moradi, M., Bradley, A., Greenspan, H., Papa, J.P., Madabhushi, A., Nascimento, J.C., Cardoso, J.S., Belagiannis, V., and Lu, Z. (eds.) Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 204–212. Springer International Publishing, Cham (2017).
15. Jaderberg, M., Simonyan, K., Zisserman, A., kavukcuoglu, koray: Spatial Transformer Networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., and Garnett, R. (eds.) Advances in Neural Information Processing Systems 28. pp. 2017–2025. Curran Associates, Inc. (2015).
16. Mountney, P., Stoyanov, D., Yang, G.-Z.: Three-Dimensional Tissue Deformation Recovery and Tracking. IEEE Signal Process. Mag. 27, 14–24 (2010).
17. Ye, M., Giannarou, S., Meining, A., Yang, G.-Z.: Online tracking and retargeting with applications to optical biopsy in gastrointestinal endoscopic examinations. Med. Image Anal. 30, 144–157 (2016).
18. Zhou, T., Tulsiani, S., Sun, W., Malik, J., Efros, A.A.: View Synthesis by Appearance Flow. In: Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 286–301. Springer International Publishing, Cham (2016).
19. Zhou, T., Krahenbuhl, P., Aubry, M., Huang, Q., Efros, A.A.: Learning dense correspondence via 3d-guided cycle consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 117–126 (2016).
20. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. ArXiv160304467 Cs. (2016).
21. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. ArXiv14126980 Cs. (2014).
22. De Visser, H., Passenger, J., Conlan, D., Russ, C., Hellier, D., Cheng, M., Acosta, O., Ourselin, S., Salvado, O.: Developing a next generation colonoscopy simulator. Int. J. Image Graph. 10, 203–217 (2010).
23. Armin, M.A., Chetty, G., Jurgen, F., Visser, H.D., Dumas, C., Fazlollahi, A., Grimpen, F., Salvado, O.: Uninformative frame detection in colonoscopy through motion, edge and Color Features. In: International workshop on computer -assisted and robotic. Springer International Publishing, Munich, Germany (2015).
24. Hamlyn Centre Laparoscopic / Endoscopic Video Datasets, http://hamlyn.doc.ic.ac.uk/vision/.
25. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Trans. Image Process. 13, 600–612 (2004).