Unsupervised Odometry and Depth Learning for Endoscopic Capsule Robots

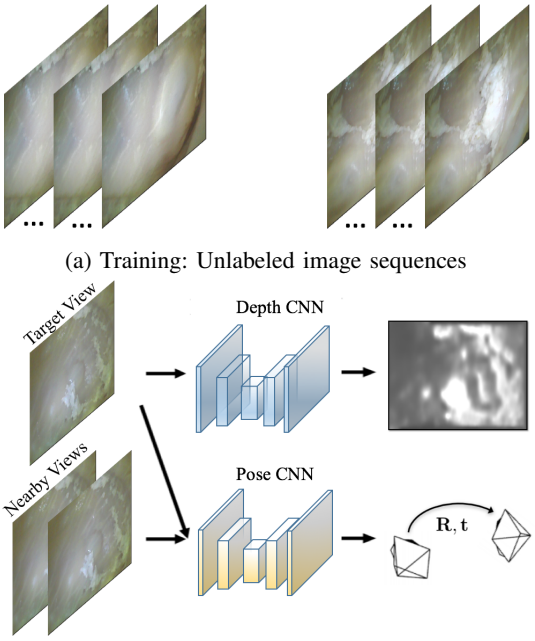
Mehmet Turan[1], Evin Pinar Ornek[2], Nail Ibrahimli [2], Can Giracoglu[2], Yasin Almalioglu[3], Mehmet Fatih Yanik [4], and Metin Sitti[5]

***Abstract*— In the last decade, many medical companies and research groups have tried to convert passive capsule endoscopes as an emerging and minimally invasive diagnostic technology into actively steerable endoscopic capsule robots which will provide more intuitive disease detection, targeted drug delivery and biopsy-like operations in the gastrointestinal(GI) tract. In this study, we introduce a fully unsupervised, real-time odometry and depth learner for monocular endoscopic capsule robots. We establish the supervision by warping view sequences and assigning the re-projection minimization to the loss function, which we adopt in multi-view pose estimation and single-view depth estimation network. Detailed quantitative and qualitative analyses of the proposed framework performed on non-rigidly deformable ex-vivo porcine stomach datasets proves the effectiveness of the method in terms of motion estimation and depth recovery.**

I. Introduction

Advancements in various fields of science and technology in the last decade has opened new pathways for non-invasive examination of patient's body and detailed investigation about diseases. Hospitals are using innovative ways to provide accurate data from inside of the human body. As an emerging example, various diseases such as colorectal cancer and inflamatory bowel disease are diagnosed by the usage of swallowable capsule endoscopes, which are non-invasive, painless, suitable to be used for long duration screening purposes which can access difficult body parts (e.g.,small intestines) better than standard endoscopy. Such benefits make swallowable, non-tethered capsule endoscopes an exciting alternative over standard endoscopy [1], [2].

Current capsule endoscope technology employed in GI tract monitoring and disease detection consists of passive devices which are locomated by random peristaltic motions. The doctor would have an easier access to fine-scale body parts and could make more intuitive and correct diagnosis in case of a precise and reliable control over the position of the capsule. Many research groups attempted to build remotely controllable active endoscopic capsule robot systems with additional functionalities such as local drug delivery, biopsy and other medical functions [2]–[17], which are, on the other hand, heavily dependent on a real-time and precise pose estimation capability.



(a) Training: Unlabeled image sequences

(b) Testing: Pose and depth prediction

Fig. 1: Unsupervised training approach consists of two separate neural networks, one for depth prediction and another one for multi-view pose estimation. It requires unlabeled image sequences from different temporal points to establish a supervision basis. Models produce pose estimation between two views from different perspectives parameterized as 6-DoF motion, and depth prediction as a disparity map for a given view.

In this work, we propose a novel real-time localization and depth estimation approach for endoscopic capsule robots which mimic the remarkable ego-motion estimation and scene reconstruction capabilities of human beings by training an unsupervised deep neural network. The proposed network consists of two simultaneously trained sub-networks, the first one assigned for depth estimation via encoder-decoder strategy, the second assigned to regress the camera pose in 6-DoF. The model observes sequences of monocular images and aims to interpret them to estimate executed camera motion in 6-DoF and the depth map of the observed scene as shown in Fig. 1. Our framework estimates the camera motion and depth information in an end-to-end and unsupervised

[1]Mehmet Turan is with Physical Intelligence Department, Max Planck Institute for Intelligent Systems, Germany turan@is.mpg.de

[2]Evin Pinar Ornek, Nail Ibrahimli, Can Giracoglu is with the Informatics Faculty, Technical University of Muenich, Germany evin.oernek, nail.ibrahimli, can.giracoglu@tum.de

[3]Yasin Almalioglu is with Computer Science Department, University of Oxford, Oxford, UK yasin.almalioglu@cs.ox.ac.uk

[4]M. Fatih Yanik is with the Department of Information Technology and Electrical Engineering, Zurich, Switzerland yanik@ethz.ch

[5]Metin Sitti is with Physical Intelligence Department, Max Planck Institute for Intelligent Systems, Germany sitti@is.mpg.de

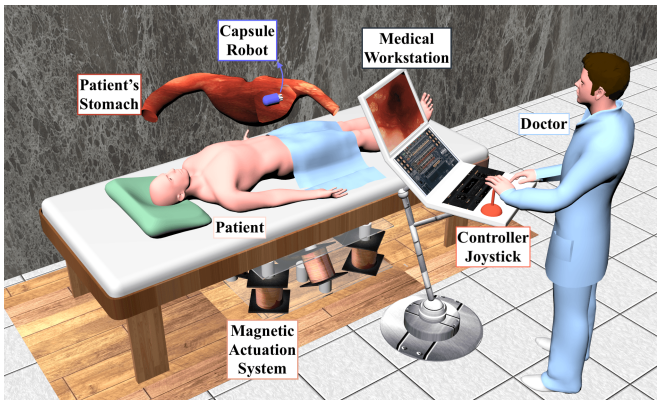


Fig. 2: Demonstration of the active endoscopic capsule robot operation using MASCE (Magnetically actuated soft capsule endoscope) designed for disease detection, drug delivery and biopsy-like operations in the upper GI-tract. MASCE is composed of a RGB camera, a permanent magnet, an empty space for drug chamber and a biopsy tool. Electromagnetic coils based actuation unit below the patient table exerts forces and torques to execute the desired motion. Doctor operates the screening, drug delivery and biopsy processes in real-time using the live video stream onto the medical workstation and the controller joystick to maneuver the endoscopic capsule to the desired position/orientation and to execute desired therapeutic actions such as drug release and biopsy.

fashion directly from input pixels. Training is performed using only unlabeled monocular frames in a similar way to prior works such as [18]–[20].

We formulate the entire pose estimation and map reconstruction pipeline for endoscopic capsule robots as a consistent and systematic learning concept which can improve its performance every day by collecting streamed data belonging to numerous patients undertaken to endoscopic capsule robot and standard endoscopy investigations in hospitals over the world. This way, we want to mimic and transfer a continuous learning functionality from medical doctors into medical robots domain, where experience and adaptation to unexpected novel situations can be much more critical to real-world scenarios.

To summarize, main contributions of our paper are as follows:

- To best of our knowledge, this is the first unsupervised odometry and depth estimation approach for both the endoscopic capsule robots and hand-held standard endoscopes;
- Since the network learns in a fully unsupervised manner, no ground truth pose and/or depth values are required to train the neural network;
- Neither prior knowledge nor parameter tuning is needed to recover the trajectory and depth, contrary to traditional visual odometry(VO) and deep learning(DL) based supervised odometry approaches;
- We simultaneously train a reliability mask which identifies pixels distorted by camera occlusions, non-rigid organ deformations and/or non-Lambertian surface. Such a mask is very crucial for vision based methods applied on endoscopic type of images since occlusions, non-rigid deformations and specularities violating Lambertian surface properties commonly occur in endoscopic types of images.

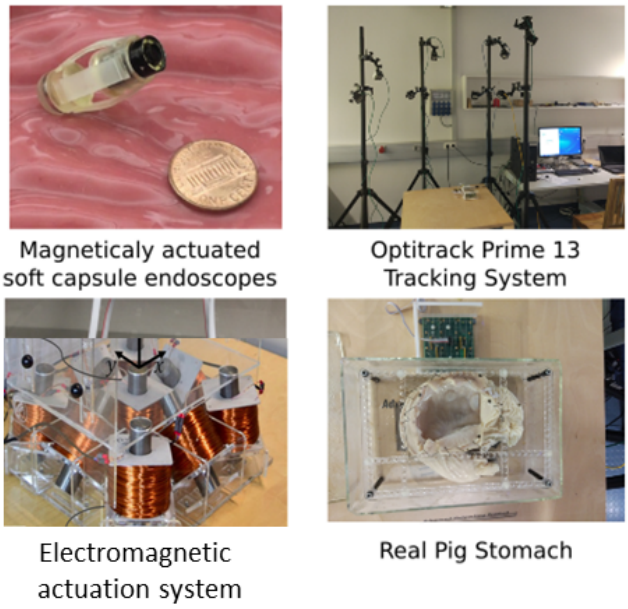


Fig. 3: Illustration of the experimental setup. MASCE is a magnetically actuated robotic capsule endoscope prototype which has a ringmagnet on the body. An electromagnetic coil array consisting of nine coils is used for the actuation of the MASCE. The ringmagnet exerts magnetic force and torque on the capsule in response to the external magnetic field provide by the electromagnetic coil array. Magnetic torque and forces are also used to release drug, as well. OptiTrack system consisting of eight infrared cameras is employed for the ground truth pose estimation. An opened and oiled porcine stomach simulator is used to represent human stomach.

Evaluations we made on non-rigidly deformable porcine stomach videos prove the success of our depth estimation and localization approach. As the outline of this paper, the previous work in endoscopic capsule odometry is discussed in Section II. Section III introduces the proposed method with its mathematical background in detail and the unsupervised DL architecture. Section IV shows our experimental quantitative and qualitative results achieved for 6-DoF localization and depth recovery. Finally, Section V mentions some bottlenecks and gives future directions for our project. Our code will be made available at https://github.com/mpi/deep-unsupervised-endovo.

II. Background

In the last decade, several localization methods [21]–[25] were proposed to calculate the 3D position and orientation of the endoscopic capsule robot such as fluoroscopy [21],

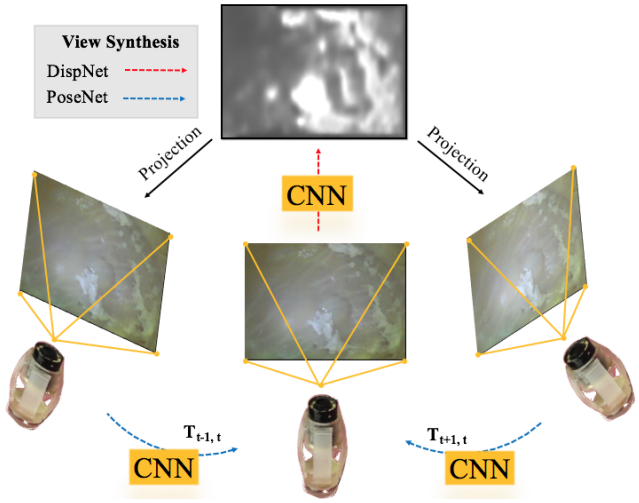


Fig. 4: Training input consists of sequential images from different perspectives, which are noted by $< I_{t-1}, I_t, I_{t+1} >$. After view synthesis creates the supervision baseline, PoseNet is trained to estimate relative motion change between $< I_{t-1}, I_t >$ and $< I_t, I_{t+1} >$, whereas DispNet learns to predict depth for the target image $< I_t >$.

ultrasonic imaging [22]–[25], positron emission tomography (PET) [21], [25], magnetic resonance imaging (MRI) [21], radio transmitter based techniques and magnetic field based techniques. The common drawback of these localization methods is that they require extra sensors and hardware design. Such extra sensors have their own drawbacks and limitations if it comes to their application in small scale medical devices such as space limitations, cost aspects, design incompatibilities, biocompatibility issues and the interference of the sensors with the activation system of the device.

As a solution of these issues, a trend of VO methods have attracted the attention for endoscopic capsule localization. A classic VO pipeline typically consists of many hand-engineered parts such as camera calibration, feature detection, feature matching, outliers rejection (e.g. RANSAC), motion estimation, scale estimation and global optimization (bundle adjustment). Although some state-of-the-art algorithms based on this traditional pipeline have been developed and proposed for endoscopic VO task in the past decades, their main deficiencies such as tracking failures in low textured areas, sensor occlusion issues, lack of handling non-rigid organ deformation still remain. In last couple of years, DL techniques have been dominating many computer vision related tasks with numerous promising result, e.g. object detection, object recognition, classification problems etc. Contrary to these high-level computer vision tasks, VO is mainly working on motion dynamics and relations across sequence of images, which can be defined as a sequential learning problem.

Our proposed method solves several issues faced by typical VO pipelines, e.g the need to establish a frame-to-frame feature correspondence, vignetting artefacts, motion blur, specularity or low signal-to-noise ratio (SNR). We think that DL based endoscopic VO approach is more suitable for such challenge areas since the operation environment(GI tract) has similar organ tissue patterns among different patients which can be learned by a sophisticated machine learning approach easily. Even the dynamics of common artefacts such as non-rigidness, sensor occlusions, vignetting, motion blur and specularity across frame sequences could be learned and used for a better pose estimation, whereas our unsupervised odometry learning method additionally solves the common problem of missing labels on medical datasets from inner body operations [4], [6].

III. METHOD

Different from supervised VO learning [2], [4], [6], where camera poses and/or depth ground truths are required to train the neural network, the core idea underlying our unsupervised pose and depth prediction method is to make use of the view synthesis constraint as the supervision metric, which forces the neural network to synthesize target image from multiple source images acquired from different camera poses. This synthesis is performed using estimated depth image, estimated target camera pose values in 6-DoF and nearby color values from source images. In addition, a reliability mask is trained to detect sensor occlusions, non-rigid deformations of the soft organ tissue and lack of textures inside the explored organ.

A. View synthesis as supervision metric

To provide a supervision to the neural network, view synthesis is accomplished by training with consecutive images. As input, we take a sequence of 3 consecutive frames, and choose the middle frame as a target frame. Sequences are denoted by $< I_{t-1}, I_t, I_{t+1} >$ where I_t is the target view and rest of images are source views $I_s =< I_{t-1}, I_{t+1} >$, which are used to render the target image (see Fig. 4). The objective function of the view synthesis is:

$$\mathscr{L}_{vs} = \sum_s \sum_p |I_t(p) - \hat{I}_s(p)| \tag{1}$$

where p is pixel coordinate, and $\hat{I}_s$ is the source view I_s warped to the target view making use of the estimated depth image $\hat{D}_t$ and 4×4 camera transformation matrix $\hat{T}_{t \to s}$ [27]. Let p_t represent the homogeneous pixel coordinates in the target view, and K be the camera intrinsics matrix.

p_t is projected coordinate on the source view and p_s is acquired by:

$$p_s \sim K \hat{T}_{t \to s} \hat{D}_t(p_t) K^{-1} p_t \tag{2}$$

Note that the value of p_s is not discrete. To find the expected intensity value at that position, bilinear interpolation among four discrete neighbors of p_s is used [28]:

$$\hat{I}_s(p_t) = I_s(p_s) = \sum_{i \in \{top, bottom\}, j \in \{left, right\}} w^{ij} I_s(p_s^{ij}) \tag{3}$$

Let w^{ij} be the proximity value between projected and neighboring pixels summing up to one and $\hat{I}_s$ be the estimated mean intensity for projected pixel p_s.

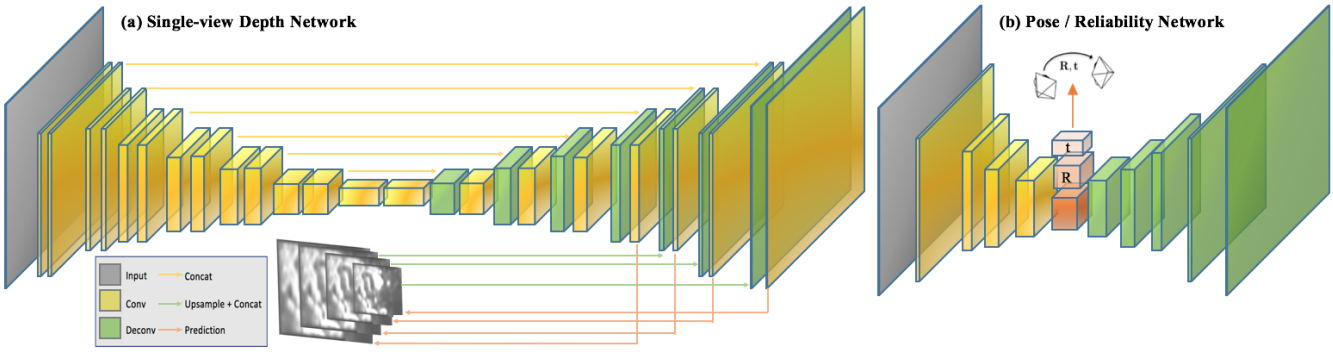


Fig. 5: The proposed neural network architecture for pose/reliability/depth map estimation. The width and height of illustrated blocks reflect the spatial dimensions of layers and output channels which are based on an encoder-decoder design. (a) Single-view depth prediction model is adopted by DispNet [26]. ReLu activations follow the middle convolution layers. Kernel size for first four layers are 7, 7, 5, 5 respectively, and rest of the layers have kernel size 3. (b) Pose/reliability estimation network is motivated by SFM-Learner [19] model and it has decoder-encoder design, as well. The encoder part has five feature extraction layers which are shared for both pose and reliability mask estimation. The pose results are gathered after the encoder network, which has $6*(N-1)$ output channels for 6-DoF motion parameters. The encoder part is followed by a decoder, which has 5 deconvolutional layers, consisting ReLU activations in between.

View synthesis approach assumes that camera sensor is not occluded, non-rigid deformations are avoided and explored organ surface obeys Lambertian surface rules enabling photometric error minimization between target and source views. These assumptions are frequently violated in endoscopic type of videos:

1) Sensor occlusions occur often due to peristaltic organ motions.
2) Inner organs have in general a non-rigid structure meaning deformations cannot be completely avoided.
3) Organ fluids cause specularities which violate the Lambertian surface rules.

To overcome these, we trained a soft reliability mask which labels each target-source pixel pair as reliable to be used for view-synthesis or believed to violate assumptions because of being affected by occlusions, non-rigid deformations and/or specularities. Incorporating the soft-reliability mask $\hat{E}_s$, the view synthesis equation is updated as:

$$\mathscr{L}_{vs} = \sum_{<I_{t-1},I_{t+1}>\in\mathscr{S}} \sum_{p} \hat{E}_s(p)|I_t(p) - \hat{I}_s(p)| \ . \qquad (4)$$

Minimizing this energy function without regularizer will force mask to be zero across the whole image domain. To overcome this problem and obtain a reasonable mask, a regularization term is to use which describes the prior knowledge about reliability mask. Hence, let $\mathscr{L}_{reg}(\hat{E}_s^l)$ be the regularization term that minimizes the cross-entropy loss and prevents trivial solutions. Finally, since gradients are derived from differences between four neighbors and corresponding pixel intensities of source and target frames, a smoothness loss $\mathscr{L}_{smooth}^l$ is needed. The multiscale pyramid and smoothness loss for gradients are extracted from larger spatial regions. This leads to the following energy function:

$$\mathscr{L}_{final} = \sum_l \mathscr{L}_{vs}^l + \lambda_s \mathscr{L}_{smooth}^l + \lambda_e \sum_s \mathscr{L}_{reg}(\hat{E}_s^l) \qquad (5)$$

Here, s indexes source images, l indexes images from different scales, λ_s is the regularization weight for depth smoothness, and λ_e is the weight for reliability mask.

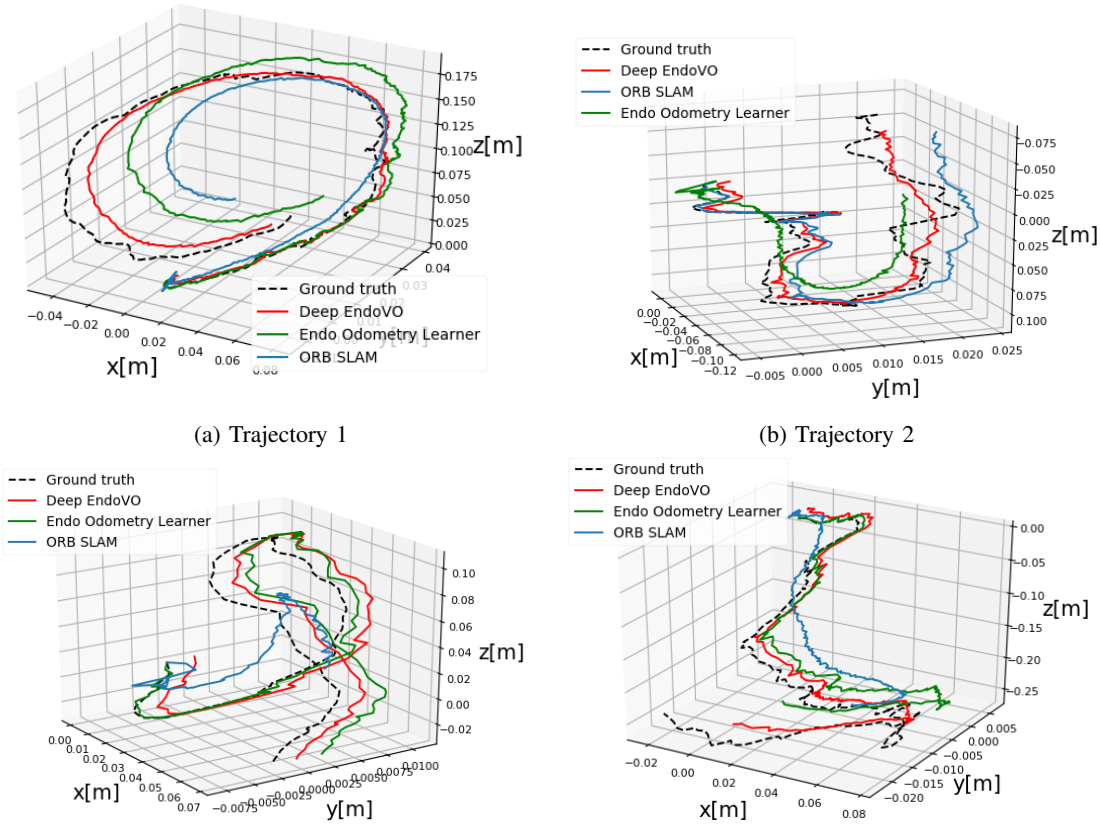
B. Network architecture

As mentioned earlier, our problem is estimating odometry in a textureless scene by using only sequenced RGB frames as input. Since classical methods fail to cope with this problem, we use DL methods where we get our motivation from recent works [29] and [19] which propose improvements by autoencoder based architectures. Our overall DL model as shown in Fig. 5 consists of two end-to-end frameworks.

The first architecture is employed to predict single-view depths by creating disparity map outputs. The encoder-decoder convolutional layers are followed by a prediction layer, whose outputs are constrained by $1/(\alpha*sigmoid(x)+\beta)$ with $\alpha = 10$ and $\beta = 0.1$ to ensure that predictions occur in a desirable interval.

The second network tries to estimate relative pose, parameterized by SE(3) motions between views, and the reliability mask. The encoder part for pose estimation and reliability mask are same, where they share weights in the first five feature extractor convolutional layers and divide into two tracks afterwards. Pose is estimated by encoder's $6*(N-1)$ channels, as translation and rotation parameters. The decoder part consists of five deconvolutional layers and generates multiscale mask predictions. There are four output channels for each prediction layer, and each two of them predict the reliability for input source-target pairs by softmax normalization.

Both networks are trained and optimized jointly. On the other hand, both networks can be tested and evaluated independently. Testing and training pipelines are illustrated in Fig. 1.



(a) Trajectory 1

(b) Trajectory 2

(c) Trajectory 3

(d) Trajectory 4

Fig. 6: Sample trajectories comparing the unsupervised learning method with ORB SLAM, EndoVO and OptiTrack ground truth in millimetric scale. Deep EndoVO shows the best odometry estimations, whereas ORB SLAM fails to track some fine-scale motions. Tracking performance of unsupervised odometry lies inbetween of ORB SLAM and Deep EndoVO; many fine-scale motions are successfully caught in detail, however there is still a certain amount of drift.

IV. EVALUATION AND RESULTS

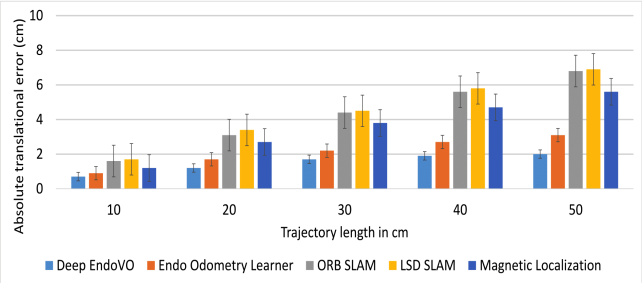
A. Dataset and Transfer learning

We used transfer learning to have an initialization for neural network weights since we lack huge amounts of labeled data. For pretraining, DL model proposed by Zhou et al. [19] is employed. The model is implemented with publicly available Tensorflow framework and pretrained with the KITTI dataset. Batch normalization is used for all of the layers except the outputs. Adam optimization is chosen to increase the convergence rate, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate of 0.1 and mini-batch size of 8. We used the model which was trained with $50K$ images and converged after 150K iterations. The model requires sequential images with size 128 x 416. On top of the model pretrained by a KITTI dataset, we fine-tuned the architecture with our domain data from endoscopic capsule robot by employing a GeForce GTX 1070 model GPU. Our dataset was collected in an experimental setup for an ex-vivo parcine stomach shown in Fig. 3 and it contains $12K$ frames with ground truth odometry obtained by OptiTrack visual tracking system. In this experiment, we fix the length of input image sequences into three frames. We used $10K$ frames for training, $1K$ for cross validation and $1K$ for evaluation and testing.

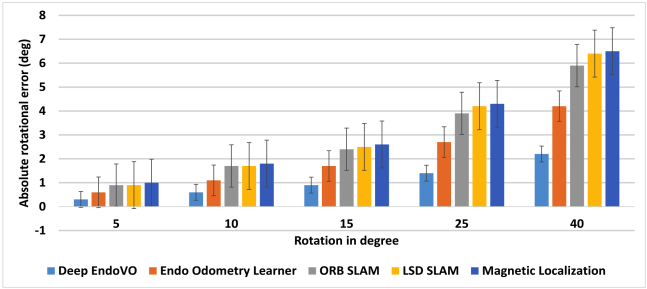
B. Pose estimation and Odometry benchmark

Our pose estimation network is tested with $1K$ frames. The network outputs the pose predictions as 6-DoF motion (Euclidean coordinates for translation and rotation) between sequences. Ground truth data was established with the OptiTrack mechanism. Some examples from odometry outputs can be seen in Fig. 6. Here, we illustrate only short sequences qualitatively. It can be seen that the main trajectory results successfully differentiate the major displacements with a minor amount of drift.

We compare our ego-motion estimation method with monocular ORB-SLAM [30], Deep EndoVO [2], LSD SLAM [31] using Absolute Trajectory Error (ATE) [30] for the alignment with the ground truth. As shown in Fig. 6 and error bars in Fig. 7a, 7b, our method outperforms ORB SLAM and LSD SLAM which are state-of-the art widely used SLAM methods. Because of the geometric and photometric properties of scenes, these methods fail to find and match proper keypoints. Magnetic localization also outperforms ORB-SLAM and LSD-SLAM, because magnetic localization does not depend on textural geometry of the



(a) Translational error results



(b) Rotational error results

Fig. 7: Translational (a) and rotational (b) error results for ORB SLAM, LSD SLAM, Deep EndoVO, magnetic localization and our proposed supervised method. It is clear that in both rotational and translational motions, our unsupervised odometry outperforms ORB SLAM, LSD SLAM and magnetic localization, whereas Deep EndoVO shows best performance. For example, for trajectory length of 10 cm, Deep EndoVO and our method results in a translational error less than 1 cm, and others are slightly above 1 cm. In terms of rotational motion, a 5 degree change has an effect of less than 1 degree in Deep EndoVO and our method, however rest of the methods are closer to 1 degree. Translational results indicate that the proposed method shows robustness for increasing trajectory lengths and remains close to the ground truth trajectory. The trajectory length increase from 10 cm to 50 cm results a change of more than 4 cm in both ORB SLAM and LSD SLAM methods, whereas our error increases around 1 cm.

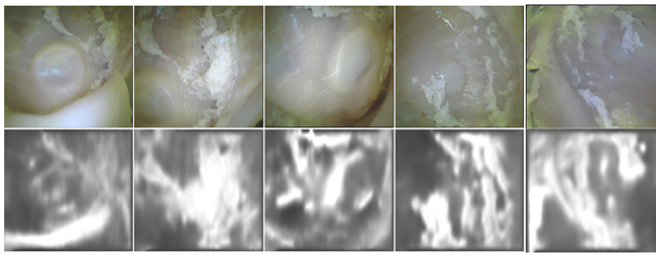


Fig. 8: Sample disparity map estimations from ex-vivo porcine stomach dataset. Even though depth estimations lack fine-scale details in low textured areas, major depth differences were successfully caught.

(a) Without fine-tuning (KITTI) (b) Original image (c) After transfer learning

Fig. 9: Disparity map outputs before and after fine-tuning on top of KITTI. (a) shows the estimation without fine-tuning. Since there is no object in front of the camera in KITTI images, the resulting disparity maps have a dark region in the center. Moreover, the disparity map has a poor quality. After transfer learning and training with porcine stomach dataset in addition to KITTI images, the quality of the disparity map drastically increases and the dark hole in the center of the image dissapears (c).

scene. Even though the proposed method is unsupervised, its translational and rotational accuracies are comparable with Deep EndoVO approach which is a supervised odometry learning method.

C. Depth Estimation

The neural network model creates depth estimation as a disparity map for a given view. Some estimation results can be seen in Fig. 8. It is clear that major depth differences are captured by the network. However, since stomach surface is non-Lambertian and the light source is attached to camera, it becomes more challenging to reproduce a robust algorithm. In the disparity map output of the network, it is observable that there are minor errors at some low textured regions or on high gradient parts such as sharp edges. However, our improvement on overall depth estimation with fine-tuning can be seen in Fig. 9.

V. CONCLUSIONS

In this paper we applied unsupervised DL method for estimating VO and depth for endoscopic capsule robot videos. Even though our method performs comparably well to supervised EndoVO method and outperforms existing state of the arts SLAM algorithms ORB and LSD SLAM, some playroom for the improvements of the method still remains:

- Accuracy of the results can be improved by increasing sequence size of inputs. As well, additional training data generated by augmentation techniques could improve the performance of the method for cases where non-rigid deformations, occlusions and heavy specularities exist.
- Since our capsule robot also uses rolling shutter camera, instead using KITTI dataset captured by global shutter camera, we could also incorporate Cityscapes dataset captured by rolling shutter camera.
- The quality of estimated depth maps can be improved by combining the depth output of our method with shading based depth estimation. In that way, a more realistic and therapeutically relevant 3D reconstruction

of the explored inner organ could be achieved. Even though the depth estimation results qualitatively look promising, a quantitative and more deeper analysis is needed in-vivo testings.

- The dependency of the proposed method on the camera intrinsics matrix makes it rather impractical to be used for random videos streaming from hospitals with unknown calibration matrix.
- It would be interesting to extend our network to perform further tasks such as tissue segmentation and disease detection.
- In future, we aim to improve pose and depth estimation of our method further to achieve even sub-millimeter accuracies which is a prerequisite to be accepted for in-vivo applications.

ACKNOWLEDGMENT

Mehmet Turan would like to thank Max Planck ETH Center for Learning Systems for funding the project.

REFERENCES

[1] M. Sitti, H. Ceylan, W. Hu, J. Giltinan, M. Turan, S. Yim, and E. Diller, "Biomedical applications of untethered mobile milli/microrobots," *Proceedings of the IEEE*, vol. 103, no. 2, pp. 205–224, 2015.

[2] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, and M. Sitti, "Deep endovo: A recurrent convolutional neural network (rcnn) based visual odometry approach for endoscopic capsule robots," *arXiv preprint arXiv:1708.06822*, 2017.

[3] M. K. Goenka, S. Majumder, and U. Goenka, "Capsule endoscopy: Present status and future expectation," *World J Gastroenterol*, vol. 20, no. 29, pp. 10 024–10 037, 2014.

[4] M. Turan, Y. Almalioglu, H. Gilbert, A. E. Sari, U. Soylu, and M. Sitti, "Endo-vmfusenet: Deep visual-magnetic sensor fusion approach for uncalibrated, unsynchronized and asymmetric endoscopic capsule robot localization data," *CoRR*, vol. abs/1709.06041, 2017. [Online]. Available: http://arxiv.org/abs/1709.06041

[5] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, and M. Sitti, "Deep endovo: A recurrent convolutional neural network (rcnn) based visual odometry approach for endoscopic capsule robots," *Neurocomputing*, vol. 275, pp. 1861 – 1870, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092523121731665X

[6] M. Turan, Y. Almalioglu, H. Gilbert, H. Araújo, T. Cemgil, and M. Sitti, "Endosensorfusion: Particle filtering-based multi-sensory data fusion with switching state-space model for endoscopic capsule robots," *CoRR*, vol. abs/1709.03401, 2017. [Online]. Available: http://arxiv.org/abs/1709.03401

[7] M. Turan, Y. Almalioglu, H. Araujo, E. Konukoglu, and M. Sitti, "A non-rigid map fusion-based direct slam method for endoscopic capsule robots," *International Journal of Intelligent Robotics and Applications*, vol. 1, no. 4, pp. 399–409, Dec 2017. [Online]. Available: https://doi.org/10.1007/s41315-017-0036-4

[8] M. Turan, Y. Y. Pilavci, I. Ganiyusufoglu, H. Araujo, E. Konukoglu, and M. Sitti, "Sparse-then-dense alignment-based 3d map reconstruction method for endoscopic capsule robots," *Machine Vision and Applications*, vol. 29, no. 2, pp. 345–359, Feb 2018. [Online]. Available: https://doi.org/10.1007/s00138-017-0905-8

[9] T. Nakamura and A. Terano, "Capsule endoscopy: past, present, and future," *Journal of gastroenterology*, vol. 43, no. 2, pp. 93–99, 2008.

[10] F. Munoz, G. Alici, and W. Li, "A review of drug delivery systems for capsule endoscopy," *Advanced drug delivery reviews*, vol. 71, pp. 77–85, 2014.

[11] F. Carpi, N. Kastelein, M. Talcott, and C. Pappone, "Magnetically controllable gastrointestinal steering of video capsules," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 231–234, 2011.

[12] H. Keller, A. Juloski, H. Kawano, M. Bechtold, A. Kimura, H. Takizawa, and R. Kuth, "Method for navigation and control of a magnetically guided capsule endoscope in the human stomach," in *Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International Conference on*. IEEE, 2012, pp. 859–865.

[13] A. W. Mahoney, S. E. Wright, and J. J. Abbott, "Managing the attractive magnetic force between an untethered magnetically actuated tool and a rotating permanent magnet," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 5366–5371.

[14] S. Yim, E. Gultepe, D. H. Gracias, and M. Sitti, "Biopsy using a magnetic capsule endoscope carrying, releasing, and retrieving untethered microgrippers," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 2, pp. 513–521, 2014.

[15] A. J. Petruska and J. J. Abbott, "An omnidirectional electromagnet for remote manipulation," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 822–827.

[16] M. Turan, Y. Almalioglu, E. Konukoglu, and M. Sitti, "A deep learning based 6 degree-of-freedom localization method for endoscopic capsule robots," *CoRR*, vol. abs/1705.05435, 2017. [Online]. Available: http://arxiv.org/abs/1705.05435

[17] M. Turan, Y. Y. Pilavci, R. Jamiruddin, H. Araújo, E. Konukoglu, and M. Sitti, "A fully dense and globally consistent 3d map reconstruction approach for GI tract to enhance therapeutic relevance of the endoscopic capsule robot," *CoRR*, vol. abs/1705.06524, 2017. [Online]. Available: http://arxiv.org/abs/1705.06524

[18] R. Szeliski, "Prediction error as a quality metric for motion and stereo," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2. IEEE, 1999, pp. 781–788.

[19] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, vol. 2, no. 6, 2017, p. 7.

[20] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "Deepstereo: Learning to predict new views from the world's imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5515–5524.

[21] T. D. Than, G. Alici, H. Zhou, and W. Li, "A review of localization systems for robotic endoscopic capsules," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 9, pp. 2387–2399, 2012.

[22] M. Fluckiger and B. J. Nelson, "Ultrasound emitter localization in heterogeneous media," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 2867–2870.

[23] J. M. Rubin, H. Xie, K. Kim, W. F. Weitzel, S. Y. Emelianov, S. R. Aglyamov, T. W. Wakefield, A. G. Urquhart, and M. ODonnell, "Sonographic elasticity imaging of acute and chronic deep venous thrombosis in humans," *Journal of Ultrasound in Medicine*, vol. 25, no. 9, pp. 1179–1186, 2006.

[24] K. Kim, L. A. Johnson, C. Jia, J. C. Joyce, S. Rangwalla, P. D. Higgins, and J. M. Rubin, "Noninvasive ultrasound elasticity imaging (uei) of crohn's disease: animal model," *Ultrasound in medicine & biology*, vol. 34, no. 6, pp. 902–912, 2008.

[25] S. Yim and M. Sitti, "3-d localization method for a magnetically actuated soft capsule endoscope and its applications," *IEEE Transactions on Robotics*, vol. 29, no. 5, pp. 1139–1151, 2013.

[26] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," *CoRR*, vol. abs/1512.02134, 2015. [Online]. Available: http://arxiv.org/abs/1512.02134

[27] C. Fehn, "Depth-image-based rendering (dibr), compression, and transmission for a new approach on 3d-tv," in *Stereoscopic Displays and Virtual Reality Systems XI*, vol. 5291. International Society for Optics and Photonics, 2004, pp. 93–105.

[28] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," *CoRR*, vol. abs/1605.03557, 2016.

[29] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, "Sfm-net: Learning of structure and motion from video," *CoRR*, vol. abs/1704.07804, 2017.

[30] R. Mur-Artal, J. Montiel, and J. D. Tardós, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[31] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European Conference on Computer Vision*. Springer, 2014, pp. 834–849.