2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Active Stereo Method for 3D Endoscopes using Deep-layer GCN and Graph Representation with Proximity Information

Michihiro Mikamo[1], Ryo Furukawa[1], Shiro Oka[2], Takahiro Kotachi[2], Yuki Okamoto[2], Shinji Tanaka[2], Ryusuke Sagawa[3], Hiroshi Kawasaki[4]

***Abstract*— Techniques for 3D endoscopic systems have been widely studied for various reasons. Among them, active stereo based systems, in which structured-light patterns are projected to surfaces and endoscopic images of the pattern are analyzed to produce 3D depth images, are promising, because of robustness and simple system configurations. For those systems, finding correspondences between a projected pattern and an original pattern is an open problem. Recently, correspondence estimation by graph neural networks (GCN) using graph-based representation of the patterns were proposed for 3D endoscopic systems. One severe problem of the approach is that the graph matching by GCN is largely affected by the stability of the graph construction process using the detected patterns of a captured image. If the detected pattern is fragmented into small pieces, graph matching may fail and 3D shapes cannot be retrieved. In this paper, we propose a solution for those problems by applying deep-layered GCN and extended graph representations of the patterns, where proximity information is added. Experiments show that the proposed method outperformed the previous method in accuracies for correspondence matching for 3D reconstruction.**

I. INTRODUCTION

3D reconstruction for endoscopic systems has been attracting many researchers. Among them, active-stereo-based techniques have been considered promising for practical usage [1]–[4], because they are easily built by just adding a static pattern projector to existing systems and textureless regions can be densely reconstructed.

Recently, Furukawa *et al.* developed a 3D endoscopic system, in which a micro-sized pattern projector with optical fiber is inserted through an instrument channel of a common monocular endoscope to recover 3D depth images. They proposed a method to robustly obtain correspondences between the detected pattern and the original pattern by representing the pattern by graphs and applying a graph convolutional network (GCN) [1]. By using a GCN, they achieved frame-wise 3D reconstruction even without a pre-calibration of the system, such as projector's position and orientation from the camera.

One severe problem of the approach is that the graph matching by GCN is largely affected by the stability of the graph construction process using the detected patterns of a captured image. If the detected pattern is fragmented into small pieces, graph matching may fail and 3D shapes cannot be retrieved. Since endoscopic images of internal organs usually include severe disturbances, such as occluding boundaries, specular noises or subsurface scattering, the projected pattern includes many unclear pattern features as well as unexpected discontinuities, which causes unstable pattern detection.

In this paper, we propose a solution for those problems by extending the GCN and the graph for the input. In the proposed method, the pattern is represented as an extended graph representation that has not only a grid structure, but also a non-grid structure of proximity edges that connects nodes that are near in 2D positions. Also, GCN model for graph matching is extended by constructing a deep-layer GCN.

The Contributions of the paper are as follows:
(1) Node-wise feature vectors are extracted to better represent each node using GCN with deep layers to increase matching accuracy between two graphs.
(2) New graph representation of projected patterns where nodes are connected not only for vertical and horizontal directions to build grid structures, but also for all the adjacent nodes within threshold to increase the density of recovered shape is proposed.
(3) Deep layered GCN [5] is adopted and modified to achieve matching between two graphs even though the topologies are different.

II. RELATED WORKS

For endoscopic diagnosis and treatment, 3D information is desired for many purposes. Thus, many researches are developing 3D reconstruction methods. These methods include photometric information [6]–[10], or texture information based on shape-from-motion (SfM) techniques [11], [12]. Photometric information heavily rely on source light intensity and characteristics and surface albedo; thus, absolute distance accuracy is limited. SfM-based approaches have problems with texture requirements and scale ambiguity.

Projecting structured light for has been used for practical applications for 3D scanning purposes [13]. For endoscope systems, scale factor of the the pattern projector is important. Thus, projection of static pattern is an inevitable choice [1]–[4]. One severe problem for static-pattern-projection approach is that the captured pattern that are distorted or fragmented should be stably matched with the original pattern. despite that the captured patterns tend to be degraded by environmental conditions, such as noise, specularity, blur, etc.

*This work was supported by 20H00611, 18K19824, 18H04119, 16KK0151, and Nedo 3.0 in Japan
[1] Hiroshima City University, JAPAN
[2] Hiroshima University Hospital, JAPAN
[3] Natl. Inst. of Advanced Industrial Science and Technology, JAPAN
[4] Kyushu University, JAPAN

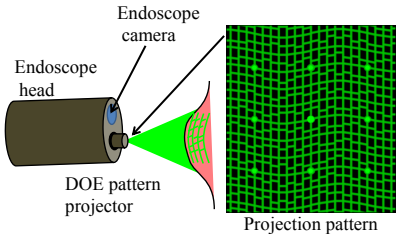

Fig. 1. System configuration.

Recently, deep-learning-based approaches are used to find correspondences between stereo-pair images [14], [15]. They have potential for outperforming existing correspondence estimation methods. For active stereo, the matching should be done between a captured image and the original pattern, where cross-domain matching is necessary. For this purpose, Furukawa *et al.* [1] proposed to represent the captured patterns with a graph, and predict the correspondence by GCN-based classification.

III. Method

A. System configuration

In this study, a projector-camera system was constructed by attaching a fiber-like micro-pattern projector to a standard endoscope. The EG-590WR endoscope from Fujifilm and the pattern projector equipped with a diffractive optical element (DOE) were used to achieve structured light illumination. The pattern projector can be inserted into the instrument channel of the endoscope, and the pattern is projected onto the surface in front of the endoscope head (Fig.1(left)). The pattern has a grid structure with a size of 21×21. Each horizontal edge of the grid structure is shifted to vertical direction to make gaps between adjacent edges (Fig.1(right)). To make unique pattern for local region using the gap in order to achieve stable correspondences between graphs, the sign and the size of each gap is carefully encoded in our method. In addition, there are nine grid points which is bigger than other elemnts in the pattern, however, they are not explicitly used in our method, since our new GCN is powerful enough to find correspondences without training them, whereas, the nine markers are explicitly trained and used for calibration in the previous method [1].

B. Overview of the Reconstruction

This section provides an overview of the 3D reconstruction process. The pattern of Fig.1(right) is projected onto the target surface, and the image is captured by the endoscopic camera. The main process of the active stereo method is to extract the pattern from the image and to map each point of the pattern to the original (projection) pattern. The grid structure of the pattern shown in Fig.1(right) is represented as a graph. Also, codes (gaps) of each grid points are represented as node attributes. After a graph is also extracted from the captured image, each node of the graph should be mapped to the corresponding grid points of the original pattern. This process is a graph matching problem.

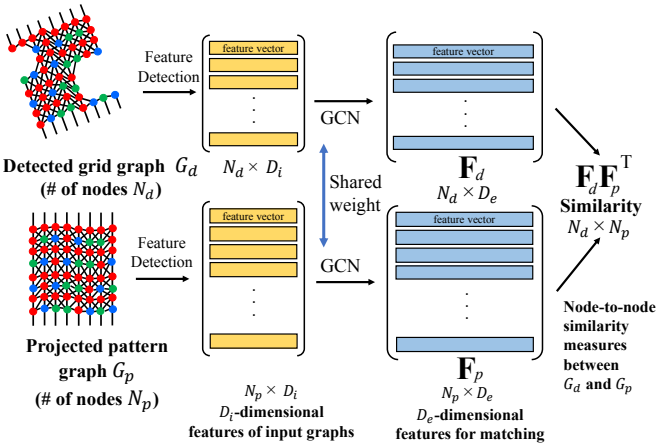

Fig. 2. Node-wise similarity calculation.

For extracting the graph from the captured image, we use deep neural network method to robustly extract the edges to build the grid structure as well as code information for each grid point. Than, the obtained graph from the image is matched with the original graph using GCN, which is a neural network that performs convolution-like operations on node-wise feature vectors. In the method of Furukawa *et al.* [1], an input of GCN was just single graph constructed by captured image and directly output correspondence map of IDs, which was assigned on (original) graph. The problem of this approach is that the network tends to be huge and over-fitted to training data, which is a common problem of classification task on deep neural net. In this paper, to solve the problem, we use a GCN, but feed two graphs as the input and output not single IDs, but simliarity values for each node; details are described in the followings.

C. GCN-based Correspondence Estimation

Let G_d be the graph obtained from the image, and G_p the graph of the original pattern. Furukawa *et al.* [1] solved the problem of assigning a corresponding node of G_p to each node of G_d as a classification problem using a GCN. In their method, GCN is applied to G_d, and for each node of G_d, the corresponding node ID of G_p is directly estimated.

In this paper, we take another approach. We apply a GCN model to both G_d and G_p, instead of only G_d. As a result, the node-wise feature vectors of the both graphs are converted by the GCN. In this process, feature vectors of each node is convolved with neighbor nodes. This operation is repeated by layers-by-layers. In this approach, the role of the GCN is to aggregate the feature vectors of neighboring nodes to compute a feature vector that is suitable for node-wise matching.

The advantage of such an architecture is that it avoids solving a classification problem with a large number of classes. In Furukawa *et al.* [1], the GCN needs to solve the N_p-class classification if G_p has N_p nodes, thus, the size of the network and the required training data increases as N_p increases. In the architecture of this paper, the dimensionality of the GCN output can be much smaller than N_p.

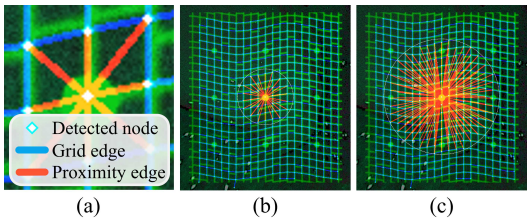

Fig. 3. The radius for adjacency nodes. In (a) the grid edges are shown in blue, and the proximity edges in orange. (b) and (c) show the change of the proximity edges with the thresholds of (b) 100 [pix] and (c) 200 [pix]. The average number of the proximity edges of a node were (b) 35.2 and (c) 127.2, respectively.

The flow of the proposed method is shown in Fig. 2. A GCN is applied to both G_d and G_p. If the number of nodes in G_d is N_d and the number of nodes in G_p is N_p, the GCN produces N_d feature vectors from G_d and N_p feature vectors from G_p. By representing these features as matrices F_d and F_p, each element of the multiplied matrix $F_d F_p^\top$ is the cosine similarity between the nodes of G_d and G_p, and the correspondence is obtained by row-by-row argmax operation.

The GCN is trained by supervised manner. For the cosine similarity, we optimize the GCN by using log softmax cross-entropy as the error function. A large amount of training data with ground-truth correspondences are generated by using CG, which is described in Sec. IV. Note that similarity value can efficiently avoid over-fitting effect on classification, and this is another important advantage of the proposed method.

D. Graph Construction with Proximity Edges

To efficiently conduct matching between two graphs, same topology is usually assumed. For examaple, in the work of Furukawa*et al.* [1], grid structure is expected as the input, *i.e.*, the edges are either vertical or horizontal lines of grid structure. The disadvantage of the method is that graph construction process is usually unstable, because wrong edges are frequently detected by noise or some edges are missing by similar reason. Since the features extraction in GCN is done by convolution via edges, unstable graphs may decreases the accuracy of correspondence estimation.

To deal with this problem, in this paper, edges based on the proximity of the positional relationship of the nodes are also generated in addition to the detected horizontal and vertical lines. In our implementation, two nodes are connected, if their distance is less than a certain length. This allows us to improve the accuracy of GCN estimation, because the additional connections between nodes increases the feature aggregation, even if the pattern in the image is fragmented and the number of grid connections of some nodes are too small. Fig.3 shows the proximity edges generated for distance thresholds of 50, 100, and 200 pixels. Note that simple implementation sometimes does not converge well or take a huge computational time, which is solved by our new GCN using residuals described in the next section.

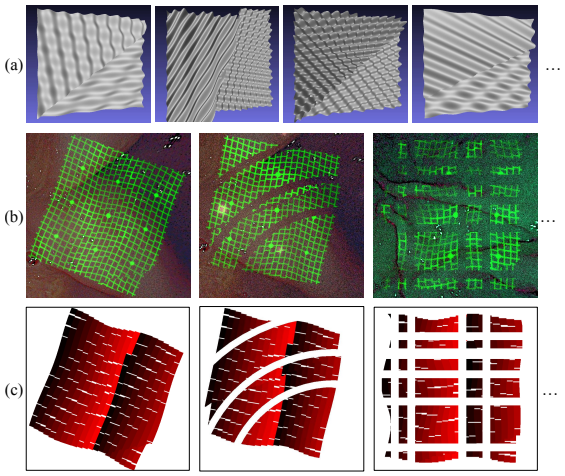

Fig. 4. Some examples of the training data. (a) Shape models for the pattern projection, (b) synthesized training data, and (c) the ID images whose pixel values represent the IDs of the nodes in (b).

E. Deep GCN with Initial Residuals

It has been reported that increasing the number of layers in a GCN leads to over-smoothing, gradient loss, and poor discrimination accuracies [16], [17]. Chen *et al.* proposed GCN-II as a GCN that is less prone to over-smoothing [5]. For each layer of GCN-II, the initial data of the graph is input as a skip connection, and identity mapping is added to the result of operations in each layer.

We apply GCN-II for better correspondence estimation. In GCN-II, a hyper parameter α_l that define the weight of the skip connection, and β for the identity mapping are used. After experimental trials, we decided to use $\alpha_l = 0.5$ and $\beta = 0$. In the experiments shown later, we tried up to 20 layers as the GCN model.

IV. Experiments

Since it is difficult to obtain a large number of endoscopic images and to annotate the ground-truth correspondences manually, a training data set was synthesized using CG-generated images as follows. First, we prepared a shape model of a plane modulated by sin waved bumps (Fig.4(a)). The sinusoidal modulation mimics shapes inside organs with folds. The pattern of Fig.1(right) was projected onto the shape model using projection mapping. Then, by overlaying real endoscopic images onto the synthesized pattern as textures, images shown in Fig.4(b) were obtained. Simultaneously, images colored with correct node ID were generated as shown in Fig.4(c), where code IDs were encoded by R and G channels of the image with $ID = R + 256 * G$.

Further image processing was used to create disconnections in circular and grid-like region separations as shown in the second and the third columns of Fig.4(b),(c). These are for imitating disconnections or pattern fragmentations from occluding boundaries often caused by folds of surfaces inside organs. From these images, we created pairs of graphs and ground-truth correspondences for the training data.

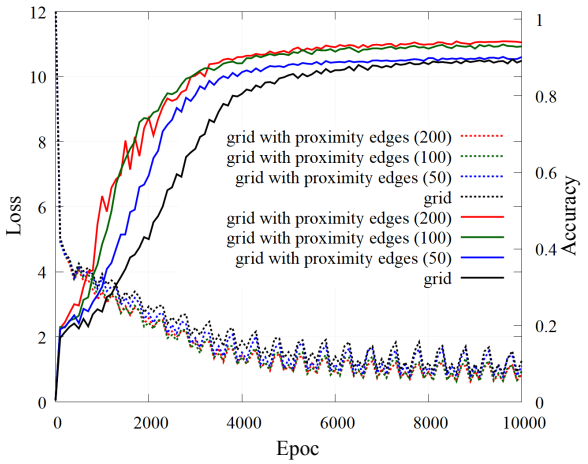

Fig. 5. Learning processes with different configurations of the proximity edges (different distance thresholds). The dashed lines mean loss, and the solid lines means accuracies. The model is GCN-II with 20 layers.

TABLE I

ACCURACIES FOR CONFIGURATIONS WITH GCN/GCN-II, WITH/WITHOUT PROXIMITY EDGES, AND SHALLOW/DEEP LAYERS.

	Layer: 5		Layer: 20	
	with proximity	without proximity	with proximity	without proximity
GCN	0.929	0.851	0.721	0.790
GCN-II	0.923	0.840	**0.942**	0.892

To show effectiveness of the proximity edges, we constructed graphs with and without proximity edges. We generated graphs with different distance thresholds for the proximity edges, with 50, 100, and 200 pixels. As the threshold increases, more proximity edges are generated. Fig.5 shows the accuracies (solid) and the training losses (dashed) in the training processes, for the same inputs with different proximity edge configurations. The GCN model is GCN-II with 20 layers. The dashed lines show the changes in losses and the solid lines show the accuracies. Accuracies are evaluated from the CG-generated data with ground-truth that are separated with the training data. Fig.5 shows that, as the proximity edges increase, accuracy of the correspondence predictions becomes better.

We also examined accuracies for different configurations of GCN models and graph representations, where the GCNs are constructed with a normal GCN layers or GCN-II layers, with or without proximity edges, and with GCN models with 5 or 20 layers. Table I shows the accuracies for the configurations. The table shows that use of proximity edges improved the result, and combination of GCN-II and deep layer improved the result. 20 layers of normal GCN operations was badly performed than 5 layers of normal GCN operations.

For further examining how the proximity edges and deep/shallow GCNs improve the accuracies, we generated images with three different conditions of data disturbance levels: 'condition A' without data disturbances, 'B' with a moderate noises and pattern fragmentation, and 'C' with a strong noises and string pattern fragmentation as shown in

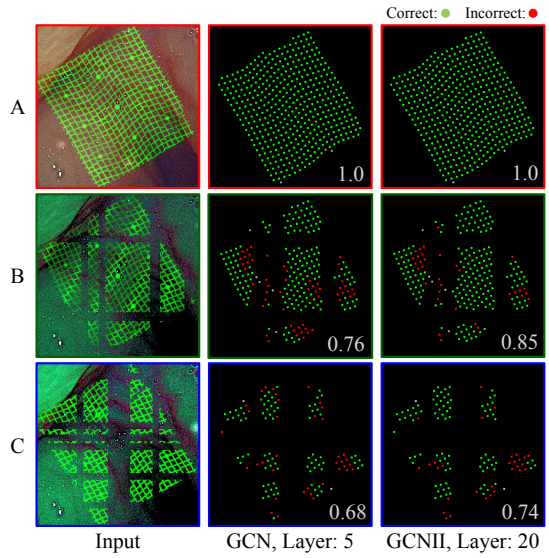

Fig. 6. Visualization of correctness/incorrectness by the inference. We applied our method to the three inputs as examples. A:good condition, B:middle condition with moderate pattern fragmentation and small noises, C: bad condition with larger fragmentation and noises.

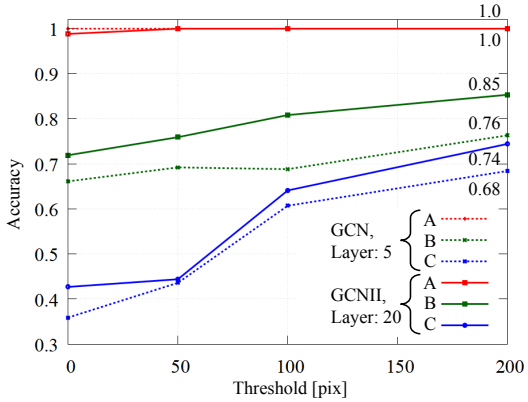

Fig. 7. Accuracies for 5-layer GCN and 20-layer GCN-II for samples A, B and C in Fig.6.

Fig.6 (left column). They are then tested by deep (20-layer GCN-II) and shallow (5-layer GCN) models for different proximity edge conditions (distance thresholds with 0, 50, 100, 150 and 200).

Fig.6 (middle/right columns) and Fig.7 show the results. From Fig.7, as the number proximity edges increases, the accuracies improved. Also, badly conditioned images are more improved by more proximity edges. Note that increasing proximity edges results in more computational cost, thus there are trade-offs between computational costs and accuracies. Fig.7 also shows that deep models constantly outperformed the shallow model. Fig.6 (middle/right columns) show that the accuracies at pattern discontinuities are improved for the deep model.

Examples of 3D reconstruction results of real endoscopic images are shown in Fig.8. The input images are surfaces inside a pig's stomach taken by an endoscope with structured-light projection. Fig.8 (a.1) shows the image with small folds, and Fig.8(b.1) shows the image with large folds. We compared the restoration results using a shallow model (5-

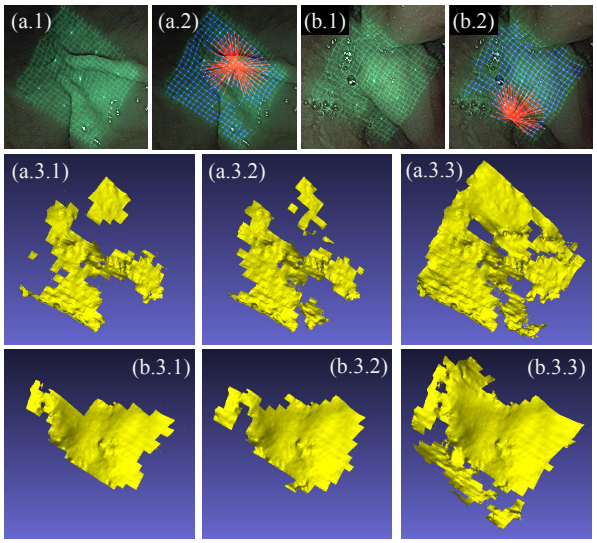

Fig. 8. Reconstruction results for real endoscopic images. The input image include (a.1) shallow fold and (b.1) deep fold. (a.2) is visualized proximity edges of (a.1). Results from (a.1) are (a.3.1) 5-layer GCN without proximity edges, (a.3.2) 20-layer GCN-II without proximity edges, and (a.3.3) 20-layer GCN-II with with proximity edges. The results obtained from (b.1) are shown similarly.

layer GCN) and deep models (20-layer GCN-II with and without proximity edges.

As can be seen from the images, the reconstructed area with the deep model was larger than that of the shallow model. Aslo, using the proximity edges results in better reconstruction. In particular, differences of reconstructed area were concentrated around folded shapes. These results indicate that GCN-II with deep layers using proximity edges performed the best for real images, similarity to the results of CG images shown in Fig.7. The proximity node connections shown in (a.2) and (b.2), connect regions across folded shapes, where disconnection of the grid often occurs, and exchanged feature information between the regions. For (a.1), the proposed method (a.3.3) recovered 201 percent more points than the conventional method(a.3.1), and for (b.1), 183 percent more ((b.3.1) and (b.3.3)).

V. Conclusion

In this paper, we proposed a matching algorithm using a GCN for correspondence estimation in a 3D endoscope system based on active stereo. A grid pattern with code features is projected onto the target, and the grid structure was extracted from the captured image by image processing using U-Nets. The grid structure was graphed, and a GCN was applied to both the graph obtained from the image and the graph of the original pattern, and correspondences are obtained by node-wise matching the outputs. In addition to the edges representing the grid structure, edges representing proximity between grid points are added to improve the stability. We also implemented deep layer GCN models using GCN-II and conducted comparison experiments. As a result, we confirmed that the accuracy becomes better when proximity edges are used, especially when the condition of the input is bad. Also, deep-layered models using GCN-II was shown to outperform shallow models or models based on naive GCN.

References

[1] R. Furukawa, S. Oka, T. Kotachi, Y. Okamoto, S. Tanaka, R. Sagawa, and H. Kawasaki, "Fully auto-calibrated active-stereo-based 3d endoscopic system using correspondence estimation with graph convolutional network," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 4357–4360.

[2] J. Geurten, W. Xia, U. Jayarathne, T. M. Peters, and E. C. Chen, "Endoscopic laser surface scanner for minimally invasive abdominal surgeries," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 143–150.

[3] X. Maurice, C. Albitar, C. Doignon, and M. de Mathelin, "A structured light-based laparoscope with real-time organs' surface reconstruction for minimally invasive surgery," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 5769–5772.

[4] C. Schmalz, F. Forster, A. Schick, and E. Angelopoulou, "An endoscopic 3d scanner based on structured light," *Medical image analysis*, vol. 16, no. 5, pp. 1063–1072, 2012.

[5] M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li, "Simple and deep graph convolutional networks," in *Thirty-seventh International Conference on Machine Learning (ICML)*, 07 2020.

[6] X. Liu, A. Sinha, M. Unberath, M. Ishii, G. D. Hager, R. H. Taylor, and A. Reiter, "Self-supervised learning for dense depth estimation in monocular endoscopy," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer, 2018, pp. 128–138.

[7] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2572–2581, 2018.

[8] F. Mahmood and N. J. Durr, "Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy," *Medical image analysis*, vol. 48, pp. 230–243, 2018.

[9] A. Rau, P. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka, L. B. Lovat, and D. Stoyanov, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *International journal of computer assisted radiology and surgery*, vol. 14, no. 7, pp. 1167–1176, 2019.

[10] M. Visentini-Scarzanella, T. Sugiura, T. Kaneko, and S. Koto, "Deep monocular 3d reconstruction for assisted navigation in bronchoscopy," *International journal of computer assisted radiology and surgery*, vol. 12, no. 7, pp. 1089–1099, 2017.

[11] R. Ma, R. Wang, S. Pizer, J. Rosenman, S. K. McGill, and J.-M. Frahm, "Real-time 3d reconstruction of colonoscopic surfaces for determining missing regions," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 573–582.

[12] X. Liu, M. Stiber, J. Huang, M. Ishii, G. D. Hager, R. H. Taylor, and M. Unberath, "Reconstructing sinus anatomy from endoscopic video–towards a radiation-free approach for quantitative longitudinal assessment," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 3–13.

[13] J. Salvi, J. Pages, and J. Batlle, "Pattern codification strategies in structured light systems," *Pattern recognition*, vol. 37, no. 4, pp. 827–849, 2004.

[14] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4353–4361.

[15] J. Žbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *The journal of machine learning research*, vol. 17, no. 1, pp. 2287–2318, 2016.

[16] K. Xu, C. Li, Y. Tian, T. Sonobe, K. Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," *CoRR*, vol. abs/1806.03536, 2018. [Online]. Available: http://arxiv.org/abs/1806.03536

[17] Y. Rong, W. Huang, T. Xu, and J. Huang, "The truly deep graph convolutional networks for node classification," *CoRR*, vol. abs/1907.10903, 2019. [Online]. Available: http://arxiv.org/abs/1907.10903