Depth estimation from single-shot monocular endoscope image using image domain adaptation and edge-aware depth estimation

Masahiro Oda, Hayato Itoh, Kiyohito Tanaka, Hirotsugu Takabatake, Masaki Mori, Hiroshi Natori & Kensaku Mori

Published online: 21 Dec 2021.

Submit your article to this journal ↗

View related articles ↗

View Crossmark data ↗

Taylor & Francis
Taylor & Francis Group

Check for updates

Depth estimation from single-shot monocular endoscope image using image domain adaptation and edge-aware depth estimation

Masahiro Oda[a,b], Hayato Itoh[b], Kiyohito Tanaka[c], Hirotsugu Takabatake[d], Masaki Mori[e], Hiroshi Natori[f] and Kensaku Mori[a,b,g]

[a]Information and Communications, Nagoya University, Nagoya, Japan; [b]Graduate School of Informatics, Nagoya University, Nagoya, Japan; [c]Department of Gastroenterology, Kyoto Second Red Cross Hospital, Kyoto, Japan; [d]Department of Respiratory Medicine, Sapporo-Minami-Sanjo Hospital, Sapporo, Japan; [e]Department of Respiratory Medicine, Sapporo-Kosei General Hospital, Sapporo, Japan; [f]Department of Respiratory Medicine, Keiwakai Nishioka Hospital, Sapporo, Japan; [g]Research Center for Medical Bigdata, National Institute of Informatics, Tokyo, Japan

ABSTRACT

We propose a depth estimation method from a single-shot monocular endoscopic image using Lambertian surface translation by domain adaptation and depth estimation using multi-scale edge loss. We employ a two-step estimation process including Lambertian surface translation from unpaired data and depth estimation. The texture and specular reflection on the surface of an organ reduce the accuracy of depth estimations. We apply Lambertian surface translation to an endoscopic image to remove these texture and reflections. Then, we estimate the depth by using a fully convolutional network (FCN). During the training of the FCN, improvement of the object edge similarity between an estimated image and a ground truth depth image is important for getting better results. We introduced a muti-scale edge loss function to improve the accuracy of depth estimation. We quantitatively evaluated the proposed method using real colonoscopic images. The estimated depth values were proportional to the real depth values. Furthermore, we applied the estimated depth images to automated anatomical location identification of colonoscopic images using a convolutional neural network. The identification accuracy of the network improved from 69.2% to 74.1% by using the estimated depth images.

1. Introduction

Diagnosis and treatment using endoscopes can encounter some common problem that the difficulty of understanding 3D structures. Although some laparoscopes are equipped with stereo scope cameras, most endoscopes, including colonoscopes, have monocular cameras because of their small size. Image-based assistance by computers can help solve the problem. Examples of image-based endoscope assistance include endoscope navigation (Hayashi et al. 2016), tracking (Luo et al. 2015), lesion detection (Brandao et al. 2017; Yuan et al. 2018), and scene understanding during surgery (Twinanda et al. 2017; Aksamentov et al. 2017). However, the lack of 3D information in monocular endoscopic images makes using such assistance difficult. Therefore, reconstruction of 3D structures or depth estimation from endoscopic images is needed in many endoscope assistance applications.

Previous research has proposed 3D structure reconstruction or depth estimation from endoscopic images, including *Shape from Shading* and feature-point matching techniques Mair-(Mair-Hein et al. 2013). Because sensors for depth measurement cannot be used in combination with endoscopes due to the limitation of the size, image-based estimation is commonly performed to estimate depth. Depth estimation from endoscopic images is commonly performed. However, these approaches easily fail to estimate depths from real endoscopic images. This is because endoscopic images can portray a huge variety of organ-surface textures. Furthermore, organs appearing in endoscopic images

show non-rigid deformations. Such deformations decrease the matching accuracies of feature points in images. Thus, a new depth estimation method that does not rely on these previous techniques is needed. Recently, many deep learning-based depth estimation methods from indoor or driving images (Godard et al. 2017; Luo et al. 2018; Prasad and Bhowmick 2019; Liu 2019a; Guo et al. 2019; Ma et al. 2019; Zhang et al. 2019; Chabra et al. 2019; Ren et al. 2019) and endoscopic images (Visentini-Scarzanella et al. 2017; Mahmood 2018a, 2018b, 2018c; Rau et al. 2019; Liu et al. 2018, 2019b; Luo et al. 2019) have been proposed. Among the estimation methods from endoscopic images, Visentini-Scarzanella et al. (Visentini-Scarzanella et al. 2017) trained and tested only using images taken from a phantom. Mahmood et al. (Mahmood 2018a, 2018b, 2018c) performed quantitative evaluation on a phantom and a porcine colon datasets. Rau et al. (Rau et al. 2019) also evaluated their method on a phantom dataset. Quantitative performances of the previous methods on real human dataset remain obscure. Liu et al. (Liu et al. 2018, 2019b) and Luo et al. (Luo et al. 2019) performed estimations from time-series and stereo images, respectively. Depth estimation from a single-shot image is still a challenging task as a baseline of depth estimation methods using time-series or stereo camera images. The single-shot image-based depth estimation is useful to perform automatic classification or recognition of large number of previously stored single-shot monocular endoscopic images in retrospective studies of endoscopic image analysis.

We propose a depth estimation method from a single-shot monocular endoscopic image. The difficulty of the depth estimation is that we cannot obtain real endoscopic images and their corresponding depth images. Depth sensors cannot be attached to endoscopes because of the size limitation. To tackle the difficulty, we employ a two-step estimation process including Lambertian surface translation from unpaired data and depth estimation. While depth estimation from a Lambertian surface in a 2D image is possible, light reflections on organ surfaces in real endoscopic images contain not only diffuse but also specular reflections. Also, textures on organ surfaces make depth estimation difficult. We remove such specular reflections and textures on the organ surface by using a real to Lambertian surface translation by a domain adaptation technique. The domain adaptation translation is performed by a fully convolutional network (FCN), which is trained in an unpaired image training framework. Then, the domain translated images are processed by a depth estimation network. We use a DenseNet-based encoder-decoder-style FCN as the depth estimation network. We propose a multi-scale edge loss that helps to give highly accurate depth estimations. Object edge information is important to evaluate the quality of depth estimations. The multi-scale edge loss evaluates the estimation quality of object edges, including clear to blurred edges. We obtain results from the depth estimation network trained using the multi-scale edge loss. We performed quantitative evaluation of depth estimation results on real human dataset that previous methods failed to report.

The contributions of this paper can be summarised as: (1) a Lambertian surface translation process by domain adaptation to improve depth estimation accuracy, (2) a multi-scale edge loss for FCN-based depth estimation, and (3) quantitative evaluation of depth estimation on real human dataset. Use of the loss improved the depth estimation accuracy.

2. Depth estimation method

2.1. Overview

A single-shot monocular real endoscopic image is the input of our method. Lambertian surface translation is applied to the image to remove specular reflections and textures on the organ surface. The translated image is processed by the depth estimation network. The network is trained using the multi-scale edge loss. The output of the network is an estimated depth image. The process flow of the proposed method is shown in Figure 1.

Depth images that correspond to real colonoscopic or bronchoscopic images are difficult to obtain because of the size limitation of such endoscopes. Therefore, we employ an unpaired training framework to estimate depth images from real endoscopic images. Our method establishes a depth estimation network from unpaired training data.

2.2. Lambertian surface translation by domain adaptation

Following the *shape from shading* theory (Horn 1990), we can estimate the shape of a 3D surface from a 2D image that captures a Lambertian surface having the homogeneous reflection property. This is because surface normals are calculated from shading or intensity values on the surface in a perfect diffuse reflection. The intensity value on the surface ϕ is calculated as

$$\phi = \rho \mathbf{N} \cdot \mathbf{S} = \rho \cos \theta, \tag{1}$$

where ρ is the diffuse reflectance rate, $\mathbf{N}$ is the surface normal, $\mathbf{S}$ is the light direction, and θ is the angle between $\mathbf{N}$ and $\mathbf{S}$. In endoscopic images, the light direction is the same as the camera direction because the light and the camera are mounted at almost the same position. From the above equation, we obtain

$$\theta = \cos^{-1} \left(\frac{\phi}{\rho} \right), \tag{2}$$

to estimate the surface normal. The surface normal vector is calculated under the assumption that the surface is smooth and continuous. Based on the above equations, the shape or depth of a 3D surface can be calculated from the surface normals if we have the ρ of the surface.

However, light reflections on organ surfaces contain not only diffuse but also specular reflections. In the dichromatic reflection model (Shafer 1985), light reflection on the surface is represented as the sum of the diffuse and specular components. In a previous method (Umeyama and Godin 2004), parameters in the model were calculated from object images taken by a rotating polariser. This approach is difficult to apply to in-vivo endoscopic images because there is no commercially available endoscopes including colonoscope or bronchoscope that has the function of rotating
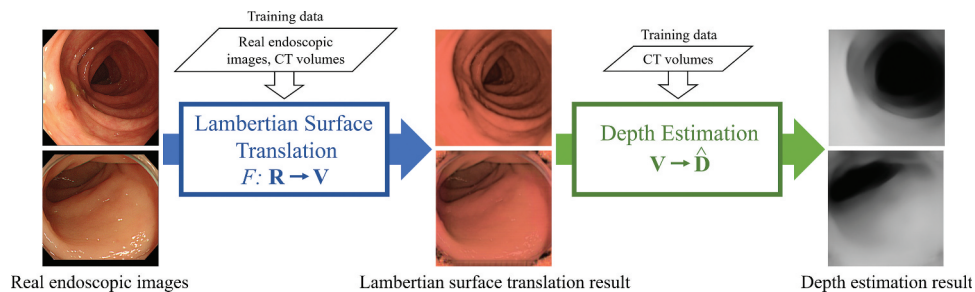

Figure 1. Process flow of proposed depth estimation method. A single-shot monocular real endoscopic image is converted by the Lambertian surface translation process to remove specular light reflection and textures on organ surfaces. The depth estimation method trained using virtual endoscopic images and depth images generated from CT volumes is applied to the image to estimate a depth image.

polarising filters. Because it is quite difficult to measure the light reflection properties on the in vivo organ surfaces, including diffuse and specular reflectance rates that are necessary to calculate the dichromatic reflection model, an alternate approach to reduce the effects of specular reflections on shape estimation is necessary. Also, there are organ-specific textures on the organ surfaces. Such textures make shape estimation difficult.

We remove specular light reflection and textures on organ surfaces from endoscopic images by using a translation based on a domain adaptation technique from real to Lambertian surface domains. Because we build the translator using a data-driven approach, we can skip measurement of the light reflection properties on the in vivo organ surfaces. We use an FCN to perform the domain adaptation. Sets of endoscopic images in real and Lambertian surface domains are denoted as **R** and **V**, respectively. The translator performs mapping $F : \mathbf{R} \to \mathbf{V}$. We train the FCN using CycleGAN (Zhu et al. 2017), which is a training framework that uses unpaired real and virtual endoscopic images. Real endoscopic images $\mathbf{r}_i(i = 1, \ldots, I) \in \mathbf{R}$ are taken from patients during endoscopic diagnoses. Endoscopic images purely containing Lambertian organ surfaces are generated as virtual endoscopic images. Virtual endoscopic images $\mathbf{v}_j(j = 1, \ldots, J) \in \mathbf{V}$ are generated from CT volumes of patients using a volume rendering technique (Mori et al. 2003) that uses diffuse reflectance as a light reflection model on organ surfaces. I and J are the numbers of real and virtual endoscopic images, respectively. The translator F is implemented as a U-Net (Ronneberger et al. 2015) with instance normalisation (Ulyanov et al. 2017) after each convolution layer.

2.3. Depth estimation network

2.3.1. Network structure
Depth estimation from a monocular single-shot 2D image is an ill-posed problem. However, with the development of deep learning-based depth estimation techniques, reasonable depths can be estimated from such images. Among many network structures for depth estimation, the encoder-decoder-style depth estimation FCN (Alhashim and Wonka 2019) based on the DenseNet-169 (Huang et al. 2017) produces accurate and high-resolution depth estimations. The depth estimation FCN has a pre-trained DenseNet-169 as an encoder. The decoder of the depth estimation FCN consists of upsampling and convolution layers. Feature maps of many resolutions in the encoder are sent to corresponding layers in the decoder by skip connections (concatenation operation). The skip connections help keep spatial resolutions and produce detailed estimations of depth images. DenseNet-169 in the encoder is pre-trained on the ImageNet. Even though the weights on the network are pre-trained to classify images, the transfer learning from classification to depth estimation improves depth estimations.

2.3.2. Multi-scale edge loss
An appropriate loss function for depth estimation is needed to get better training results from the FCN. L1 or L2 norms are commonly used as loss functions that evaluate the

difference between the ground truth and estimated depth images. Object edge information in depth images is also important for training a depth estimation network (Alhashim and Wonka 2019). Alhashim and Wonka (Alhashim and Wonka 2019) introduced first-order differential of depth values in the loss function to consider object edge difference. They calculated the first-order differential of depth values in a small local region, such as a region in 3×3 pixels. However, the use of a small local region results in consideration of very clear edges in the loss function. In some cases, edges of objects in endoscopic depth images are not clear because endoscopic images are blurred when the movement of the endoscope tip was quick. Because the application target of the method proposed by Alhashim and Wonka (Alhashim and Wonka 2019) is made up of natural images, their loss function is not suitable for endoscopic images.

We propose a multi-scale edge loss term in a loss function that takes clear and blurred object edges into account. The loss evaluates the difference of the edges in depth images that have multiple thickness. This term is effective in quality evaluations of endoscopic depth images. We represent a ground truth depth image as **D** and an estimated depth image as $\hat{\mathbf{D}}$. Our loss function L is represented as

$$L(\mathbf{D}, \hat{\mathbf{D}}) = \lambda L_d(\mathbf{D}, \hat{\mathbf{D}}) + L_S(\mathbf{D}, \hat{\mathbf{D}}) + L_e(\mathbf{D}, \hat{\mathbf{D}}), \qquad (3)$$

where $L_d(\mathbf{D}, \hat{\mathbf{D}})$ is the point-wise L1 loss term and $L_S(\mathbf{D}, \hat{\mathbf{D}})$ is the structural similarity (SSIM)-based loss term. Definitions of the terms can be found in Alhashim and Wonka (Alhashim and Wonka 2019). $L_e(\mathbf{D}, \hat{\mathbf{D}})$ is the multi-scale edge loss term, which is described as

$$L_e(\mathbf{D}, \hat{\mathbf{D}}) = \frac{1}{P} \sum_p^P \max\{|\hat{\mathbf{G}}_p^{(3)} - \mathbf{G}_p^{(3)}|, |\hat{\mathbf{G}}_p^{(5)} - \mathbf{G}_p^{(5)}|, |\hat{\mathbf{G}}_p^{(7)} - \mathbf{G}_p^{(7)}|\},$$

$$(4)$$

where p is the index of a pixel in a depth image and P is the total number of pixels in a depth image. $\mathbf{G}^{(3)}$ is the edge image obtained by applying the 3×3 differential filter to **D**. Similarly, $\mathbf{G}^{(5)}$ and $\mathbf{G}^{(7)}$ are obtained by applying the 5×5 and 7×7 differential filters to **D**, respectively. The 3×3, 5×5, and 7×7 differential filters calculate differentials between adjacent pixels, pixels at one pixel intervals, and pixels at two pixel intervals, respectively, in D. $\hat{\mathbf{G}}^{(3)}$, $\hat{\mathbf{G}}^{(5)}$, and $\hat{\mathbf{G}}^{(7)}$ are obtained by applying the three scales differential filters to $\hat{\mathbf{D}}$.

2.3.3. Network training
We train the depth estimation FCN using the multi-scale edge loss. Virtual endoscopic images $\mathbf{v}_j$ and their corresponding depth images $\mathbf{d}_j$ are fed to the FCN. Depth images $\mathbf{d}_j$, which correspond to $\mathbf{v}_j$, are generated from the CT volumes of patients. The depth images have greyscale intensity values that correspond to the distance from a virtual camera position to a position on the surface of an inner wall of a hollow organ.

2.4. Depth estimation

To estimate a depth image from a real endoscopic image, two trained FCNs are used. A real endoscopic image is processed by F to translate into a Lambert surface. The translated image is then processed by the depth estimation FCN to obtain a depth estimation result $\hat{\mathbf{d}}$.

We perform a simple correction process of depth values of the depth estimation result. The correction process contains scaling and translation of depth values. The correction process is applied to each depth value in $\hat{\mathbf{d}}$ by

$$\tilde{d}_k = s\hat{d}_k + t, \tag{5}$$

where s and t are the scaling and translation coefficients of the correction process. k is an index of pixels in $\hat{\mathbf{d}}$. $\hat{d}_k$ is a depth value of k-th pixel in $\hat{\mathbf{d}}$. $\tilde{d}_k$ is a corrected depth value of k-th pixel in a corrected depth image $\tilde{\mathbf{d}}$. We obtain the corrected depth image $\tilde{\mathbf{d}}$ as the final depth estimation result.

3. Experiments and results

We evaluated the proposed method quantitatively. We applied the method to real colonoscopic images with a point depth to evaluate the accuracy of the depth estimation. Also, to evaluate the usefulness of the estimated depth images in automated endoscopic scene understanding, we performed automated anatomical location identification of colonoscopic images using a convolutional neural network (CNN). The estimated depth images were used for image classification.

We generated $J = 8,085$ virtual colonoscopic images and corresponding depth images from six cases of colon CT volumes. These images were taken during manual fly-through in the colon in the CT volumes. For the training of the Lambertian surface translator, the virtual colonoscopic images and $I = 13,406$ real colonoscopic images were used. The generators and discriminators were trained in 400 iterations with a 38-minibatch size. The depth estimation network was trained using $J = 8,085$ pairs of virtual colonoscopic and depth images. The training epoch was 7, and the minibatch size was 10. The parameter value was set as $\lambda = 0.1$.

In the correction process of depth values, values of the parameters were set as $s = 0.73$ and $t = -3.0$. These values were selected experimentally.

The size of all images used in our method was 256×256 pixels. The virtual and real colonoscopic images were in colour, and the depth images were greyscale. The brightest and darkest intensity values in the depth images correspond to depth values of 0 and 100 mm, respectively.

3.1. Ablation study and comparative study

We performed an ablation study of the proposed method. We proposed the Lambertian surface translation (LST) method and the depth estimation by the FCN using the multi-scale edge loss (ME loss) (LST + Depth estimation with ME loss (Proposed)). To confirm the effectiveness of using the multi-scale edge loss, we used the mean absolute error loss (MAE loss) as the loss function to train the depth estimation FCN (LST + Depth estimation with MAE loss). We

also compare results obtained by using and without using the LST. To perform the comparison, we made a depth estimation method from real colonoscopic images without using the LST (Depth estimation without LST). We made the method by using an image translation based on a domain adaptation technique trained using CycleGAN. We need to use the unpaired training technique because the real colonoscopic and depth images are unpaired.

Depth estimation results of the above three methods are shown in Figure 2. The results of the proposed method and the LST + Depth estimation with MAE loss represented the shape of the colonic surfaces accurately. In the results of the Depth estimation without LST, depth values were affected by texture and specular light reflections on the surface. Estimated depth values in these areas were not accurate.

We compared the proposed method with a previously proposed depth estimation method from a single-shot monocular image (Iro et al. 2016). The previous method uses a fully convolutional residual network (FCRN) to estimate depth images. We used the method to estimate depth images from the results of the LST (LST + FCRN). The FCRN was trained using 8,085 pairs of virtual colonoscopic and depth images. The training epoch was 40, and the minibatch size was 16. Depth estimation results of the LST + FCRN are shown in Figure 2. Colonic surface shapes were not represented in the results.

3.2. Evaluation of depth estimation accuracy

We used graduated endoscopy forceps for measurement of the sizes or lengths in the endoscopic images. The forceps displays 2 mm-scaled labels along its bar-like body. During colonoscope insertions into patients, we aligned the forceps from the camera position of the colonoscope to the colonic wall to measure depth values. Then, we took real endoscopic images that included the aligned forceps. The measured depth values were used as the ground truth of the depth values. The image was called real endoscopic image with a point depth.

We applied the proposed method to 60 real endoscopic images with a point depth. In the estimated depth images, we picked up an estimated depth value at a position on the colonic wall near the location where the forceps pointed as shown in Figure 3. We avoided picking up an estimated depth value on the forceps because estimated depth value on the forceps is not accurate. We selected an estimated depth value on the colonic surface near the forceps. We compared the ground truth and estimated depth values. The results are shown in Table 1. Even though absolute values of the estimated depth values are different from the ground truth, the averaged estimated depth values were clearly increase along with the ground truth. The correlation coefficient of the ground truth and estimated depth values was 0.45. It means the estimated depth values by the proposed method were correlated with the ground truth depth values.

3.3. Application of Depth Estimation Results to Anatomical Location Identification from Colonoscopic Images

Depth information is useful in automated location-identification. To evaluate the usefulness of our method for this identification, we used the estimated depth images in CNN-
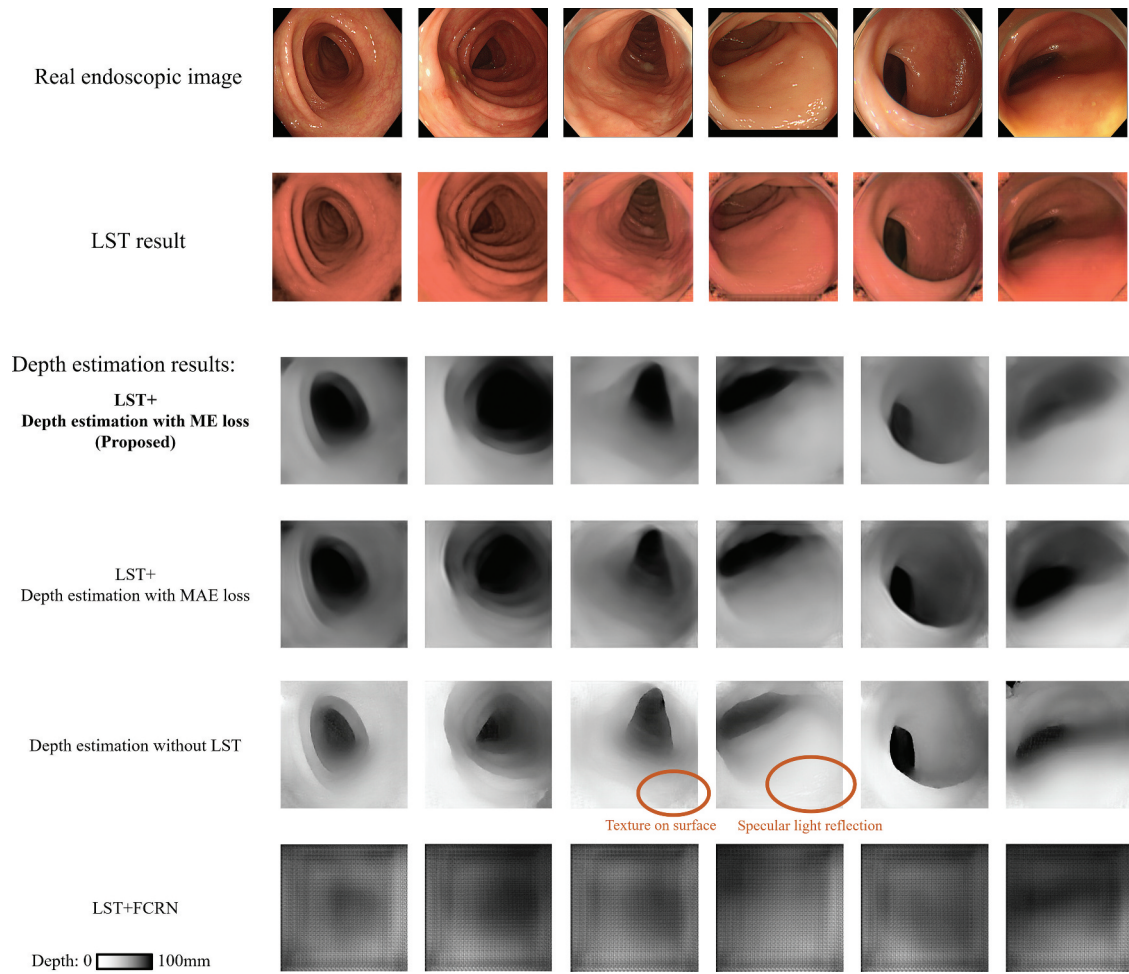
Figure 2. Results of LST and depth estimations. The real colonoscopic images shown in the top row were used as the input of the translator. Results of LST are also shown. Depth estimation results using proposed method are shown as *LST + Depth estimation with ME loss*, which uses LST and depth estimation FCN trained using multi-scale edge loss (ME loss). Results obtained using mean absolute error loss (MAE loss) are indicated as *LST + Depth estimation with MAE loss*. Results obtained without using LST are indicated by *Depth estimation without LST*. Results obtained by FCRN (Iro et al. 2016) are indicated as *LST + FCRN*.
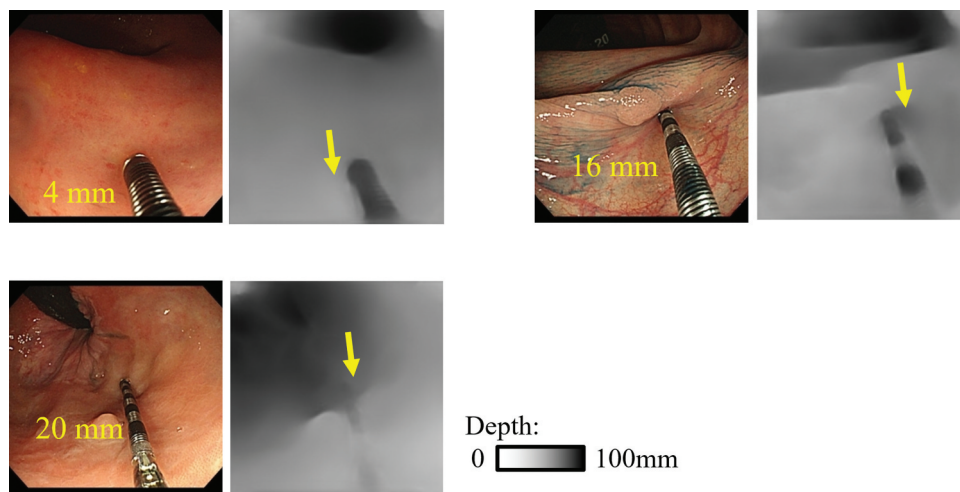

Figure 3. Real endoscopic images with a point depth and depth estimations. Lengths of forceps are shown in real endoscopic images. Depth values estimated by proposed method were measured at the positions on the colonic wall near the location where the forceps pointed (indicated by arrows). Measuring estimated depth values on the forceps was avoided because estimated depth values on the forceps are not accurate.

Table 1. Relationships between ground truth and estimated depth values.

Ground truth (mm)	Average estimated (mm)	Number of images
4 and 6	**5.38**	5
8 and 10	**11.05**	5
12 and 14	**15.73**	11
16 and 18	**16.66**	16
20	**20.21**	23

based anatomical location identification of real colonoscopic images. We made a location identification CNN, as shown in Figure 4. The CNN classified an input image into three classes: the ileocaecal area and ascending colon, the descending colon, and the rectum. We trained the CNN using two sets of images: a set of only real colonoscopic images and a set of combined images of real colonoscopic images and depth images. The combined images were made by combining the real colonoscopic images and depth images in the colour channel. We used 2131 real colonoscopic images that were not used in training either the domain adaptation translator or the depth estimation network. Depth images were generated from them using the four methods used in 3.1. 80% and 20% of the images were used for training and evaluation of the CNN, respectively. Separation of images into the training and evaluation sets were performed randomly. The CNN was trained in 50 epochs with a 50-minibatch size. We compared the classification accuracies of the CNN when trained using the two sets of images.

We performed the random training/evaluation sets separations and CNN trainings five times. The classification accuracies are shown in 2. From this table, the depth images contributed to improving the classification accuracies of CNN. Among the four depth estimation methods, using the depth images generated by the proposed method resulted in obtaining the highest classification accuracy. This shows that the depth images generated by the proposed method include much useful information for understanding real colonoscopic images.

4. Discussion

Machine learning-based depth estimation from colonoscopic images is difficult because commercially available colonoscopes cannot obtain depth images. By translating real colonoscopic images to virtual images by the LST, the depth estimation network trained in the virtual image domain is applicable for colonoscope depth estimation.

In Figure 2, the depth images generated by the proposed method reflect the distance from the camera to the colonic wall at any point on the image. The importance of the LST can be observed by comparing the results of the proposed method and the results obtained without using LST. In the results obtained without using LST, depth values were affected by texture and specular light reflections on the colonic surfaces. The proposed method obtained better depth estimation results regardless of them. The estimated depth values and the ground truth values had a positive correlation (the correlation coefficient was 0.45) in the experiments using the real endoscopic images with a point depth. This means the estimated depth images represent the shape of colonic walls. However, absolute values of the estimated depth were different from the ground truth in Table 1. The differences were caused by difference of the real and virtual camera parameters. The camera parameters of the real colonoscope and virtual camera should be calibrated to reduce differences of how these cameras map a target object to images. In our method, camera calibration was not performed. We need to calibrate real colonoscope and virtual camera to improve the results.

We conducted an experiment to evaluate the usefulness of the estimated depth images for automated location identification. The averaged accuracy of the location identification from images was improved from 69.2% to 74.1% by using the estimated depth images obtained by using the proposed method. We obtained the highest accuracy among the depth estimation methods shown in Table 2. The results indicate the LST and ME loss proposed in this paper contribute in obtaining high quality depth estimation results. The shape information of objects in scenes is quite important for understanding scenes. However, the sizes of objects are difficult to understand from 2D images because distance information from the camera to the object has lost. The estimated depth images contributed to recover the distance information and improved the automated location identification accuracy. We showed a typical result as one example of using the proposed depth estimation method. In practice, the proposed method can be applied to many automated endoscopic tasks involving scene understanding.

The depth images can be used to improve automated navigation, tracking, scene understanding, lesion detection, and quantitative analysis of lesions. By using the depth images, these methods can utilise not only colour information but also 3D shape information on the surface of organs or surgical
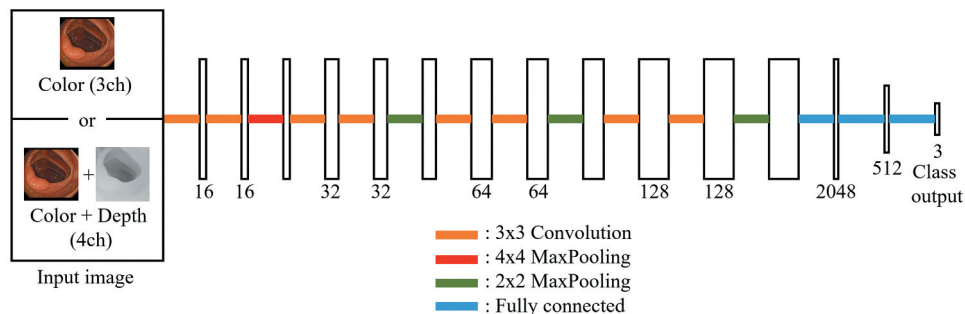

Figure 4. Structure of CNN for anatomical location identification of colonoscopic images. The boxes are layers in the CNN. The numbers shown below the boxes are the numbers of kernels or neuron units. The three neurons in the final layer correspond to the three classes.

Table 2. Classification accuracies of location identification CNN. R and D mean real colonoscopic and depth images. Averages and standard deviations in five times experiments were shown.

Input of CNN	Classification accuracy (Ave. $\pm$ S.D.)
R	69.2 $\pm$ 1.5 %
R, D generated by	**74.1 $\pm$ 2.5 %**
LST + Depth estimation with ME loss	
(Proposed)	
R, D generated by	70.8 $\pm$ 2.1 %
LST + Depth estimation with MAE loss	
R, D generated by	69.9 $\pm$ 2.4 %
Depth estimation without LST	
R, D generated by	63.7 $\pm$ 8.8 %
LST + FCRN (Iro et al. 2016)	

tools. The proposed method shows quite promising results for enhancing endoscopic diagnosis/treatment assistance. Our method can estimate depths from monocular single-shot images. This is an important result to extend our application to stereo cameras and time-series images.

An important clinical application of the depth estimation is size measurement of a colonic polyp during a colon inspection. Colonic polyps larger than 5 mm should be confirmed by a physician whether they are benign or malignant. Our method provided accurate depth estimation results when a measurement target was close to the colonoscope camera. Our method can be used to measure a colonic polyp size when a physician finds a suspicious region during an inspection. To improve clinical value of our method, depth estimation accuracy of targets that are distant from the colonoscope camera should be improved.

5. Conclusions

We proposed a depth estimation method from monocular single-shot endoscopic images. A domain adaptation technique was used to translate a real endoscopic image into a Lambertian surface domain. The translated image was processed by the depth estimation FCN. The FCN has a DenseNet-based encoder-decoder structure. The FCN was trained using the multi-scale edge loss. In the experiment using real endoscopic images with a point depth, the estimated depth values had positive correlation with the ground truth values. Also, in the experiment using the location identification CNN, use of the estimated depth image resulted in improvement of averaged identification accuracy, from 69.2% to 74.1%. Our future work will include investigation of network structures for domain adaptation or depth estimation and applications to scene understanding of other types of images.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Masahiro Oda (Ph.D. Nagoya University '09) is an associate professor at the Nagoya University, Japan. His research is in medical image processing and mainly concerns computer-aided diagnosis and computer-assisted surgery in many application areas. He has (co-)authored over 200 peer-reviewed full papers in international conferences and journals and is the recipient of RSNA Certificate of Merit (2009, 2014, 2019) awards.

Hayato Itoh is currently a designated assistant professor at the Graduate School of Informatics, Nagoya University, Japan. He received his PhD degree from Chiba University in 2017. He served as a Research Fellow at the Graduate School of Informatics, Nagoya University, from 2017 to 2020. His research is pattern recognition, including its applications to computer-aided diagnosis. His research interests range from multilinear algebra to machine learning with a particular focus on tensor analysis for pattern recognition, computer vision and medical image processing.

Kiyohito Tanaka M.D. is chief information officer of the Kyoto Second Red Cross Hospital and director of the Japan Gastroenterological Endoscopy Society. His research is in medical area research.

Hirotsugu Takabatake (M.D., Ph.D, Sapporo Medical University 1990) is a hospital staff at the Sapporo Minami-Sanjo-Hospital, Japan. His research is in respiratory medicine and computer-aided diagnosis in pulmonary area. He is the recipient of the Certificate of Merit awards in recognition of infoRAD Exhibits in 1996, 2004, and for Education Exhibits in 2009, 2019 at the Radiological Society of North America.

Masaki Mori (M.D., Ph.D, Sapporo Medical University 1989) is a hospital staff at the Sapporo-Kosei General Hospital, Japan. His research is in respiratory medicine and computer-aided diagnosis in pulmonary area. He is the recipient of the Certificate of Merit awards in recognition of infoRAD Exhibits in 1996, 2004, and for Education Exhibits in 2009, 2019 at the Radiological Society of North America.

Hiroshi Natori (M.D., Ph.D., FCCP) is a professor emeritus, the Sapporo Medical University School of Medicine, and honorary director of the Keiwakai Nishioka Hospital, Japan. His research is in respiratory medicine and medical imaging. He is the recipient of the Certificate of Merit awards in recognition of infoRAD Exhibits in 1996, 2004, and for Education Exhibits in 2009, 2019 at the Radiological Society of North America.

Kensaku Mori has received the B.S degree in electronics engineering, the M. S degree in Information Engineering, and the Ph.D in Information Engineering from Nagoya University, Japan, under the supervision of Prof. Jun-ichiro Toriwaki, in 1992, 1994 and 1996 respectively. He was a research fellow of the Japanese Society for the Promotion of Science (JSPS) from 1994 to 1997, Research Associate at Department of Computational Science and Engineering, Nagoya University from 1997 to 2000, an assistant professor in 2000. He was Associate Professor at the Research Center for Advance Waste and Emission Management of Nagoya University from 2001 to 2003. He was also Visiting Associate Professor at Department of Neurosurgery, Stanford University from 2001 to 2002. He was Associate Professor of the Graduate School of Information Science, Nagoya University from 2003 to 2009. He has been Professor of Information and Communications Headquarters, Nagoya University since 2009. In 2016, he started his services as the Directory of Information Technology Center, Nagoya University. In 2017, he was appointed as the Vice Chair of Information and Information and Communication, Nagoya University. He has been a Professor of Graduate School of Informatics, Nagoya University since 2017. His current research interests include three-dimensional image processing, computer graphics, virtual reality, and their applications to medical image. Computer aided diagnosis and surgery are his major active research fields. Recently AI-based colonoscope diagnosis system developed in his laboratory has obtained an official certificate as a medical device. He has received many awards from several institutes includes the Young Researcher Award, Ministry of Education, Culture, Sports, the best paper awards from the Japanese Society of Medical Imaging Technology. He served as the general chair of the 16th international conference of Medical Image Computing and Computer Assisted Intervention (MICCAI 2013). He was elected as the President of the International Society of Computer Aided Surgery (ISCAS) in 2018. He was the Chair of SPIE

Medical Imaging CAD conference in 2018 and 2019. He is a member of the Institute of Electronics, Information and Communication Engineering of Japan, Japanese Society of Biomedical Engineering, and Japanese Society of Medical Imaging Technology, MICCAI, IEEE, and SPIE.

References

Aksamentov I, Twinanda AP, Mutter D, Marescaux J, Padoy N. 2017. Deep neural networks predict remaining surgery duration from cholecystectomy videos. MICCAI 2018, LNCS. 10434:586–593.

Alhashim I, Wonka P: High quality monocular depth estimation via transfer learning. arXiv:1812.11941v2 (2019)

Brandao P, Mazomenos E, Ciuti G, Caliò R, Bianchi F, Menciassi A, Dario P, Koulaouzidis A, Arezzo A, Stoyanov D. 2017. Fully convolutional neural networks for polyp segmentation in colonoscopy SPIE Med Imag, Feb 11-16. Orlando (FL). 10134: 101340F .

Chabra R, Staub J, Sweeney C, Newcombe R, Fuchs H: StereoDRNet: dilated residual stereo net. IEEE Conf CVPR. Jun 16-20, Long Beach (CA), pp. 11786–11795 (2019).

Godard C, Aodha OM, Brostow GJ: Unsupervised monocular depth estimation with left-right consistency. IEEE Conf CVPR. Jul 21-26, Honolulu (HI), pp. 270–279 (2017).

Guo X, Yang K, Yang W, Wang X, Li H: Group-wise correlation stereo network. IEEE Conf CVPR. Jun 16-20, Long Beach (CA), pp. 3273–3282 (2019).

Hayashi Y, Misawa K, Oda M, Hawkes DJ, Mori K. 2016. Clinical application of a surgical navigation system based on virtual laparoscopy in laparoscopic gastrectomy for gastric cancer. Int J CARS. 11(5):827–836.

Horn BKP. 1990. Height and gradient from shading. Int J Computer Vision. 5:37–75.

Huang G, Liu Z, Maaten L, Weinberger KQ: Densely connected convolutional networks. IEEE Conf CVPR. Jul 21-26, Honolulu (HI), pp. 2261–2269 (2017).

Iro L, Christian R, Vasileios B, Federico T, Nassir N: Deeper depth prediction with fully convolutional residual networks. Fourth Int Conf 3DV Oct 25-28, Stanford (CA), pp. 239–248 (2016).

Liu C, Gu J, Kim K, Narasimhan SG, Kautz J: Neural RGB->D sensing: depth and uncertainty from a video camera. IEEE Conf CVPR. Jun 16-20, Long Beach (CA), pp. 10986–10995 (2019a).

Liu X, Sinha A, Ishii M, Hager GD, Eriter A, Taylor RH, Unberath M: Dense depth estimation in monocular endoscopy with self-supervised learning methods. arXiv:1902.07766 (2019b)

Liu X, Sinha A, Unberath M, Ishii M, Hager GD, Taylor RH, Reiter A: Self-supervised learning for dense depth estimation in monocular endoscopy. Int Workshop on CARE 2018, CLIP 2018, OR 2.0 2018 ISIC 2018, LNCS. Sep 16-20, Granada, Spain **11041**, pp. 128–138 (2018).

Luo H, Hu Q, Jia F. 2019. Details preserved unsupervised depth estimation by fusing traditional stereo knowledge from laparoscopic images. Healthc Technol Lett. 6(6):154–158.

Luo X, Wan Y, He X, Mori K. 2015. Observation-driven adaptive differential evolution and its application to accurate and smooth bronchoscope three-dimensional motion tracking. Med Imag Anal. 24(1):282–296.

Luo Y, Ren J, Lin M, Pang J, Sun W, Li H, Lin L: Single view stereo matching. IEEE Conf CVPR. Jun 18-22, Salt Lake City (UT), pp. 155–163 (2018).

Ma W-C, Wang S, Hu R, Xiong Y, Urtasun R: Deep rigid instance scene flow. IEEE Conf CVPR. Jun 16-20, Long Beach (CA), pp. 3614–3622 (2019).

Mahmood F, Chen R, Durr NJ. 2018a. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. IEEE TMI. 37 (12):2572–2581.

Mahmood F, Chen R, Sudarsky S, Yu D, Durr NJ. 2018b. Deep learning with cinematic rendering: fine-tuning deep neural networks using photorealistic medical images. Phys Med Biol. 63(18):185012.

Mahmood F, Durr NJ. 2018c. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. Med Imag Anal. 48:230–243.

Mair-Hein L, Mountney P, Bartoli A, Elha:wary H, Elson D, Groch A, Kolb A, Rodrigues M, Sorger J, Speidel S, et al. 2013. Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery. Med Imag Anal. 17(8):974–996.

Mori K, Suenaga Y, Toriwaki J. 2003. Fast software-based volume rendering using multimedia instructions on PC platforms and its application to virtual endoscopy. Proc SPIE Med Imag. 5031:111–122.

Prasad V, Bhowmick B: SfMLearner++: learning monocular depth & ego-motion using meaningful geometric constraints. 2019 IEEE WACV Jan 8-10, Waikoloa Village (HI), pp. 2087–2096 (2019)

Rau A, Edwards PJE, Ahmad OF, Riordan P, Janatka M, Lovat LB, Stoyanov D. 2019. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. Int J CARS. 14:1167–1176.

Ren H, El-khamy M, Lee J: Deep robust single image depth estimation neural network using scene understanding. IEEE Conf CVPR Workshops Jun 16-20, Long Beach (CA), pp. 37–45 (2019)

Ronneberger O, Fischer P, Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation. MICCAI 2015. 9351:234–241.

Shafer SA. 1985. Using color to separate reflection components. Color Res Appl. 10(4):210–218.

Twinanda AP, Shehata S, Mutter D, Marescaux J, Mathelin M, Padoy N. 2017. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE TMI. 36(1):86–97.

Ulyanov D, Vedaldi A, Lempitsky V: Instance normalization: the missing ingredient for fast stylization. arXiv:1607.08022 (2017)

Umeyama S, Godin G. 2004. Separation of diffuse and specular components of surface reflection by use of polarization and statistical analysis of images. IEEE PAMI. 26(5):639–647.

Visentini-Scarzanella M, Sugiura T, Kaneko T, Koto S. 2017. Deep monocular 3D reconstruction for assisted navigation in bronchoscopy. Int J CARS. 12(7):1089–1099.

Yuan Y, Qin W, Ibragimov B, Han B, Xing L. 2018. RIIS-DenseNet: rotation-invariant and image similarity constrained densely connected convolutional network for polyp detection. MICCAI 2018, LNCS. 11071:620–628.

Zhang F, Prisacariu V, Yang R, Torr PHS: GA-Net: guided aggregation net for end-to-end stereo matching. IEEE Conf CVPR Jun 16-20, Long Beach (CA), pp. 185–194 (2019)

Zhu J-Y, Park T, Isola P, Efros AA: Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv:1703.10593 (2017)