

Unsupervised Colonoscopic Depth Estimation by Domain Translations with a Lambertian-Reflection Keeping Auxiliary Task

Hayato Itoh · Masahiro Oda ·
Yuichi Mori · Masashi Misawa ·
Shin-Ei Kudo · Kenichiro Imai ·
Sayo Ito · Kinichi Hotta ·
Hirotsugu Takabatake · Masaki Mori ·
Hiroshi Natori · Kensaku Mori

Received: 9 November 2020 / Accepted: 2 May 2021

H. Itoh

Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
Tel.: +81-52-789-5688
Fax: +81-52-789-3815
E-mail: hitoh@morim.m.is.nagoya-u.ac.jp

M. Oda, and K. Mori

Graduate School of Informatics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan

Y. Mori

Clinical Effectiveness Research Group, University of Oslo, Gaustad Sykehus, Bygg 20, Sognsvannsveien 21, Oslo, 0372, Norway

Y. Mori, M. Misawa, and S.-E. Kudo

Digestive Disease Center, Showa University Northern Yokohama Hospital, Chigasaki-chuo 35-1, Tsuzuki-ku, Yokohama, 224-8503, Japan

K. Imai, S. Ito, and K. Hotta

Division of Endoscopy, Shizuoka Cancer Center, 1007 Shimonagakubo, Nagaizumi-cho, Sunto-gun, Shizuoka, 411-8777, Japan

Hirotsugu Takabatake

Department of Respiratory Medicine, Sapporo-Minami-Sanjo Hospital, Nishi-6-chome, Minami-3-jo, Chuo-ku, Sapporo, Hokkaido 060-0063, Japan

Masaki Mori

Department of Respiratory Medicine, Sapporo-Kosei General Hospital, Higashi-8-chome, Kita-3-jo, Chuo-ku, Sapporo, Hokkaido 060-0033, Japan

Hiroshi Natori

Department of Respiratory Medicine, Keiwakai Nishioka Hospital, 1-52, 4-jo 4-chome, Nishioka, Toyohira-ku, Sapporo, Hokkaido, 062-0034, Japan

Abstract *Purpose:* A three-dimensional (3D) structure extraction technique viewed from a two-dimensional image is essential for the development of a computer-aided-diagnosis (CAD) system for colonoscopy. However, a straightforward application of existing depth-estimation methods to colonoscopic images is impossible or inappropriate due to several limitations of colonoscopes. In particular, the absence of ground-truth depth for colonoscopic images hinders the application of supervised machine-learning methods. To circumvent these difficulties, we developed an unsupervised and accurate depth-estimation method.

Method: We propose a novel unsupervised depth-estimation method by introducing a Lambertian-reflection model as an auxiliary task to domain translation between real and virtual colonoscopic images. This auxiliary task contributes to accurate depth estimation by maintaining the Lambertian-reflection assumption. In our experiments, we qualitatively evaluate the proposed method by comparing it with state-of-the-art unsupervised methods. Furthermore, we present two quantitative evaluations of the proposed method using a measuring device, as well as a new 3D reconstruction technique and measured polyp sizes.

Results: Our proposed method achieved accurate depth estimation with an average estimation error of less than 1 mm for regions close to the colonoscope in both of two types of quantitative evaluations. Qualitative evaluation showed that the introduced auxiliary task reduces the effects of specular reflections and colon wall textures on depth estimation and our proposed method achieved smooth depth estimation without noise; thus validating the proposed method.

Conclusions: We developed an accurate depth-estimation method with a new type of unsupervised domain translation with the auxiliary task. This method is useful for analysis of colonoscopic images and for the development of a CAD system since it can extract accurate 3D information.

Keywords Colonoscopy · depth estimation · medical image understanding · computer-aided diagnosis · domain translation · Lambertian reflection

1 Introduction

The extraction of three-dimensional (3D) structures is an essential medical-image processing task for understanding anatomical structures and surgical scenes. Especially in colonoscopy, extracting 3D structures in a view from two-dimensional images is important for the development of computer-aided diagnosis (CAD) systems, since we can obtain only a set of two-dimensional images by a colonoscope’s monoscopic camera. To assist colonoscopic examination, several depth-estimation-based methods have been proposed [1–4]. Nadeem and Kaufman integrate depth estimation into their polyp-detection method to prevent overlooking polyps [1]. Itoh et al. proposed automatic polyp-size classification using estimated depth to clarify whether a polyp is over or under 10 mm for qualitative diagnosis [2]. Ma et al. proposed a real-time 3D colon reconstruction method using deep-learning-based depth and

camera-motion estimation to detect missing regions and reduce overlooking lesions in colonoscopy [3]. Chen et al. adopted an estimated depth map to obtain accurate localisation and mapping for the construction of an accurate navigation system [4]. In these methods, depth estimation plays an essential role. Therefore, an accurate depth-estimation method from a single-shot image has great potential for developing an accurate CAD system in colonoscopy.

In spite of the importance of depth estimation in colonoscopy, the settings of this estimation problem are challenging. Even though supervised depth-estimation methods for a single-shot image have been reported in computer vision [5–7], depth information cannot be measured with a colonoscope due to its hardware limitations. Since ground truth depth is unavailable, supervised methods are inapplicable to colonoscopic images. On the other hand, based on classical shape from shading [8] and multiple view geometry [9], unsupervised deep-learning-based methods have been proposed for an autonomous car-driving system [10–13]. These methods construct convolutional neural network (CNN) models without the ground truth of depth. Unsupervised training is achieved by the optimisation of depth-estimation CNN and camera-pose estimation CNN for temporary-sequential images. For the optimisation, a Lambertian-reflection assumption and geometrical constraint with corresponding matching among sequential images are used instead of the ground truth of depth. However, colonoscopic images include non-Lambertian reflections, such as the specular reflections and textures of a colon wall. Furthermore, colonoscopic images have fewer discriminative textures and poor local geometrical features for fine corresponding matching [3, 4]. In this difficult situation, we can achieve only blurred or partly corrupted depth images with the current unsupervised deep-learning method for colonoscopic images [2].

To circumvent the above challenges, we propose a new unsupervised depth-estimation method for a colonoscope. Instead of the conventional multiple-view-geometry approach, we adopt a model-based domain translation approach. We first generated virtual RGB-D colonoscopic images, based on an ideal Lambertian-reflection model with an unique albedo for depth estimation, from human computed tomography (CT) colonography data by virtual-colonoscopy settings [15]. We used these generated virtual images and the real colonoscopic images collected in daily colonoscopy to find the translations between the domains of these two types of images. In general, textures of colon walls and specular reflections lead to depth-estimation errors, since these are violations of the Lambertian-reflection model. However, we can reduce the effects of these violations by using Lambertian-reflection-based images for the training of translations. Figure 1 summarises our proposed method. In the training of the proposed method, its loss function evaluates translated Lambertian-reflection-based images in addition to translated depth images. This additional evaluation is backpropagated to the all weights in a model and contributes to keeping the Lambertian-reflection model as an auxiliary task for the depth estimation. From the results of this backpropagation, deep-learning architecture learns how to ignore the textures and specular reflections in input images. Therefore, we achieved an accurate depth estimation from a

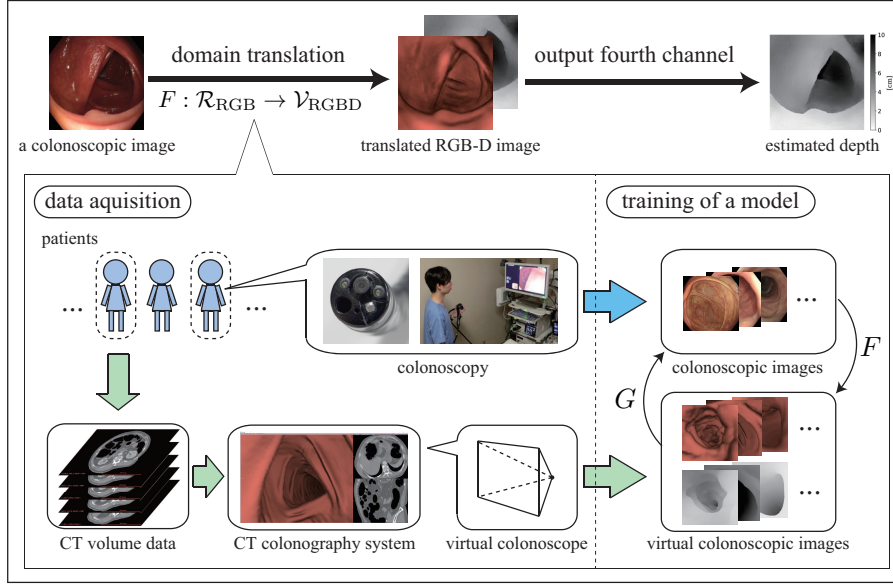


Fig. 1 Proposed depth-estimation method for a colonoscope.

real colonoscopic image by adopting this auxiliary task for unsupervised learning.

2 Methodology

2.1 Depth estimation and domain translation

Computing a 3D shape of a surface as a set of 3D points $(X, Y, Z)^\top \in \mathbb{R}^3$ from a brightness image $\mathbf{I}(\mathbf{u})$, $\mathbf{u} = (u, v)^\top \in \Omega$ of an object's surface captured in an image plane $\Omega \subset \mathbb{Z}^2$ is conventionally known as a *shape-from-shading* problem. For shape from shading, Belhumeur and colleagues proved the following property with the Lambertian-reflection model, which is a model of a surface's diffuse reflection.

Bas-relief Ambiguity *When the lighting direction and the Lambertian reflectance (albedo) of the surface are unknown, the same image can be obtained by a continuous family of the surface [8, 18].*

To circumvent this ambiguity, we assume that the lighting is known and the albedo is uniform. We then can express the brightness on an image by

$$\mathbf{I}(\mathbf{u}) = \rho \mathbf{l}^\top \mathbf{n} \propto \cos \theta, \quad (1)$$

where ρ is a reflectance ratio (albedo) and θ is an angle between lighting direction \mathbf{l} and surface normal vector \mathbf{n} , $\|\mathbf{l}\|_2 = \|\mathbf{n}\|_2 = 1$ at surface point

$(X, Y, Z)^\top$. From the given brightness and lighting directions, we can estimate the normal vectors over an image. By adding a smooth condition to a surface as an assumption, we can estimate the shape of a surface by minimising the sum of the gradient of the surface normal vectors over an image [8]. Note that the solution is not unique due to the properties of this ill-posed problem, which includes concave/convex ambiguities [8].

The depth information in colonoscopy expresses the distance between a colonoscope and the surface of a colon wall. In the depth estimation, we have to capture the distance to a colon wall in addition to the shape of a colon wall. Therefore, the estimation problem of depth $D(\mathbf{u}) \in \mathbb{R}$ from a three-channel discrete image \mathbf{X} in colonoscopy can be interpreted as an extension of shape-estimation problem. Note that the directions of the colonoscope and lighting are the same, where a single light that is attached to the colonoscope's tip. We then have the following proposition with the Lambertian-reflection model for the depth-estimation problem.

Proposition 1 *We set \mathcal{V}_{RGB} and \mathcal{V}_{D} to be domains of virtual RGB colonoscopic images and virtual depth images, respectively. Virtual RGB images are generated under the Lambertian-reflection model of a unique albedo by using CT colonography data. We assume that a virtual depth image expresses a distribution of distances from points on a colon surface to the optical centre of a colonoscope. By using the given pairs of a virtual colonoscopic image and a depth image, CNN can find a translation $\Phi : \mathcal{V}_{\text{RGB}} \rightarrow \mathcal{V}_{\text{D}}$.*

The validity of the above proposition was experimentally supported [16, 17]. In this report, virtual colonoscopic images were generated under the Lambertian reflection of a unique albedo. Therefore, these virtual images are ideal images for depth estimation without violation elements on images such that textures and specular reflections. Toward the application of transformation Φ to real colonoscopic images, we have to remove the violation elements on the colon surface. Therefore, we introduce the following proposition.

Proposition 2 *For given unpaired data of virtual RGB colonoscopic images in domain \mathcal{V}_{RGB} and real RGB colonoscopic images in domain \mathcal{R}_{RGB} , Generative Adversarial Networks (GANs) can find a domain translation $\Psi : \mathcal{R}_{\text{RGB}} \rightarrow \mathcal{V}_{\text{RGB}}$ and its inverse $\Psi^{-1} : \mathcal{V}_{\text{RGB}} \rightarrow \mathcal{R}_{\text{RGB}}$.*

Studies have experimentally supported the validity of this proposition [16, 19]. Unfortunately, however, Ψ leads to cumbersome two-step optimisation and estimation.

Instead of introducing Ψ , training techniques have been proposed for the achievement of a translation $\mathcal{R}_{\text{RGB}} \rightarrow \mathcal{V}_{\text{D}}$. To mitigate the difference between real and virtual RGB images, Rau et al. introduced an adversarial loss of translated real colonoscopic images to the training of Pix2Pix and proposed it as extended Pix2Pix (ExtPix2Pix) [17]. Even though ExtPix2Pix adopted the GAN framework, ExtPix2Pix bases on the paired learning of virtual RGB and depth image; This paired learning is apt to overfit a domain of training data, and does not work well for unseen data. Furthermore, ExtPix2Pix uses only one

real colonoscopic image for each minibatch of 20 images in training. Therefore, its practical advantage is unclear. Mathrew et al. proposed an extended cycle loss and directional discriminator for the training of CycleGAN [14]. Their new loss function reduced the effects of patient-specific texture and specular reflections in depth estimation. However, we found that the extended cycle loss causes over smoothing and artefact generation. Thus, to achieve accurate depth estimation, we need to integrate the reflection model to depth estimation.

Propositions 1 and 2 lead to the following properties. A real RGB colonoscopic image has information for the domain translation to Lambertian-assumed virtual colonoscopic images, and these virtual colonoscopic images provide ideal information for depth estimation. Therefore, we assume that real RGB colonoscopic images have information for both translations to a virtual RGB colonoscopic image and a depth image. We then have the following proposition.

Proposition 3 *For given unpaired data of real colonoscopic images in domain \mathcal{R}_{RGB} and virtual RGB-D colonoscopic images in domain $\mathcal{V}_{\text{RGBD}}$, CycleGAN can find domain translation $F : \mathcal{R}_{\text{RGB}} \rightarrow \mathcal{V}_{\text{RGBD}}$ and its inverse $G : \mathcal{V}_{\text{RGBD}} \rightarrow \mathcal{R}_{\text{RGB}}$.*

Against Proposition 3, we propose a new method in the next subsection.

2.2 Proposed depth-estimation method

For a spatial size of $H \times W$, and the conditions $i = 1, 2, \dots, n_i, j = 1, 2, \dots, n_j$ and $n_i \neq n_j$, we set discrete images $\mathbf{X}_i \in \mathbb{R}^{H \times W \times 3}$ in \mathcal{R}_{RGB} and $\mathbf{Y}_j \in \mathbb{R}^{H \times W \times 4}$ in $\mathcal{V}_{\text{RGBD}}$. For the domains \mathcal{R}_{RGB} and $\mathcal{V}_{\text{RGBD}}$, we define mapping $F : \mathcal{R}_{\text{RGB}} \rightarrow \mathcal{V}_{\text{RGBD}}$ and $G : \mathcal{V}_{\text{RGBD}} \rightarrow \mathcal{R}_{\text{RGB}}$ as two kinds of domain translations. In addition, we introduced two adversarial discriminators, $D_{\mathcal{V}}$ and $D_{\mathcal{R}}$, where $D_{\mathcal{R}}$ distinguishes between images $\{\mathbf{X}_i\}_{i=1}^{n_i}$ and $\{G(\mathbf{Y}_j)\}_{j=1}^{n_j}$, and $D_{\mathcal{V}}$ distinguishes between images $\{\mathbf{Y}_j\}_{j=1}^{n_j}$ and $\{F(\mathbf{X}_i)\}_{i=1}^{n_i}$. As the extension of Ref. [20], for distributions $p(\mathcal{R})$ and $p(\mathcal{V})$ of real and virtual colonoscopic images, we then define adversarial losses

$$\begin{aligned} \mathcal{L}_{\text{GAN}}(F, D_{\mathcal{V}}, \mathcal{V}_{\text{RGBD}}, \mathcal{R}_{\text{RGB}}) = & \mathbb{E}_j^{\mathbf{Y}_j \sim p(\mathcal{V})} [\log D_{\mathcal{V}}(\mathbf{Y}_j)] \\ & + \mathbb{E}_i^{\mathbf{X}_i \sim p(\mathcal{R})} [\log(1 - D_{\mathcal{V}}(F(\mathbf{X}_i)))] , \end{aligned} \quad (2)$$

$$\begin{aligned} \mathcal{L}_{\text{GAN}}(G, D_{\mathcal{R}}, \mathcal{R}_{\text{RGB}}, \mathcal{V}_{\text{RGBD}}) = & \mathbb{E}_i^{\mathbf{X}_i \sim p(\mathcal{R})} [\log D_{\mathcal{R}}(\mathbf{X}_i)] \\ & + \mathbb{E}_j^{\mathbf{Y}_j \sim p(\mathcal{V})} [\log(1 - D_{\mathcal{R}}(G(\mathbf{Y}_j)))] , \end{aligned} \quad (3)$$

where the first and second terms of the right side of each equation evaluate how accurately the original and translated images are discriminated. These adversarial losses evaluate the matching between the distribution of the translated images to the data distribution of the genuine images in the target domain. If the adversarial losses are small for well-trained discriminators, we can obtain realistic translations.

By extending the cycle-consistency loss [21], we define a loss by

$$\mathcal{L}_{\text{cyc}}(F, G) = \mathbb{E}_{\mathbf{X}_i \sim p(\mathcal{R})} [\|G(F(\mathbf{X}_i)) - \mathbf{X}_i\|_1] + \mathbb{E}_{\mathbf{Y}_j \sim p(\mathcal{V})} [\|F(G(\mathbf{Y}_j)) - \mathbf{Y}_j\|_1]. \quad (4)$$

This cycle-consistency loss evaluates the consistency of two mappings to prevent learned mappings F and G from contradicting each other. By using Eqs. (2)-(4), we define our object function by

$$\begin{aligned} \mathcal{L}(G, F, D_{\mathcal{V}}, D_{\mathcal{R}}) &= \mathcal{L}_{\text{GAN}}(F, D_{\mathcal{V}}, \mathcal{V}_{\text{RGBD}}, \mathcal{R}_{\text{RGB}}) \\ &+ \mathcal{L}_{\text{GAN}}(G, D_{\mathcal{R}}, \mathcal{R}_{\text{RGB}}, \mathcal{V}_{\text{RGBD}}) + \lambda \mathcal{L}_{\text{cyc}}(F, G), \end{aligned} \quad (5)$$

where λ controls the importance of the cycle-consistency loss defined in Eq. (4). In this paper, we use $\lambda = 10$. The object function in Eq. (5) evaluates the translation to RGB-D images, where both the Lambertian and depth images are evaluated. If translation F violates the Lambertian-reflection model in its training, the loss value increases. Therefore, minimisation of the adversarial loss can find translation F while maintaining the model and ignoring textures and specular reflections. The translation from $\mathcal{R}_{\text{RGB}} \rightarrow \mathcal{V}_{\text{RGB}}$ is an auxiliary task for depth estimation $\mathcal{R}_{\text{RGB}} \rightarrow \mathcal{V}_{\text{D}}$ in F , since this task reduces the effects of textures and specular reflections in the translation. This is reason why we adopt $\mathcal{R}_{\text{RGB}} \rightarrow \mathcal{V}_{\text{RGBD}}$ instead of the direction domain translation $\mathcal{R}_{\text{RGB}} \rightarrow \mathcal{V}_{\text{D}}$. Finally, we obtain translations while keeping the Lambertian-reflection assumption as solutions

$$F^*, G^* = \arg \min_{F, G} \max_{D_{\mathcal{V}}, D_{\mathcal{R}}} \mathcal{L}(F, G, D_{\mathcal{V}}, D_{\mathcal{R}}). \quad (6)$$

This min-max problem searches optimal translations by maximising the performance of discriminators and minimising the adversarial losses. The solution of Eq. (6) achieves a depth estimation from a given colonoscopic image \mathbf{X} as the fourth channel of a translated image

$$\hat{\mathbf{Y}} = F^*(\mathbf{X}). \quad (7)$$

3 Dataset construction

We constructed a real and virtual colonoscopic dataset for the development and evaluation of our unsupervised depth-estimation method. We firstly collected colonoscopic movies during typical colonoscopies of 37 patients at two hospitals using the HQ290ZI colonoscope (Olympus, Japan) with Institutional Review Board approval. Next, we divided these movies into training, and test A and test B data without duplication of patients. We then extracted still images from these divided data at 5 frames per second (fps). Each extracted image in the test-A data includes a polyp, whose size is measured. Each extracted image in the test-B data includes measuring forceps. Finally, we generated RGB-D virtual images for training and validation data from the CT colonography

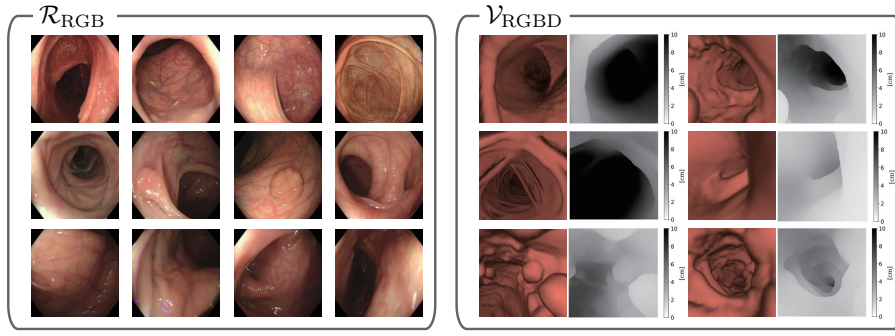


Fig. 2 Examples of images in training data. Left: real RGB three-channel colonoscopic images in domain \mathcal{R}_{RGB} . Right: virtual RGB-D four-channel colonoscopic images in domain $\mathcal{V}_{\text{RGBD}}$. In virtual four-channel images, depth represents physical length captured from CT volumes.

data of seven and two patients, respectively, with CT colonography software: NewVES [15].

In the dataset, there is no correspondence of patients between the extracted real colonoscopic images and the CT data, since the collection of these CT data and the colonoscopic images are independent. Each CT colonography data of patients is a volume data of $512 \times 512 \times 349$ voxels with $0.66 \times 0.66 \times 1.3$ mm thickness. In the NewVES, setting a virtual camera and a light source on a colon, we can observe the inside of a colon such like an usual colonoscopy. In this observation, a virtual colonoscopic view represents only geometrical shape of a colon wall without textures, since the NewVES generate an image by ray-tracing technique against the distribution of colon's CT values with a Lambertian-reflection model of a unique albedo. From this observation, we generated virtual Lambertian-based RGB images. For the generation of the virtual depth image, we set pixel values to a range of $[0, 255]$ for distances in the range of $[0, 10]$ cm. In this setting, this distance represent the physical length from a camera centre to a colon wall. Regions father than 10 cm are ignored and set to 10 cm, since close regions are important for detailed observation of polyps in colonoscopy. From this setting, the optimisation of Eq. (6) tries to find the translation from the real colonoscopic images to the depth images with a fixed scale $[0, 10]$ cm.

Furthermore, we constructed a sequential-image dataset for the multiple-view-geometry approach in the comparative evaluations. Based on results on the previous work [2], we extracted still images at 30 fps and generated triplets of successive images from the movies of 30 patients for training and validation data. Table 1 summarises the number of images in the constructed dataset. Figure 2 shows examples of the training data for the proposed method.

Table 1 Summary of the number of real and virtual colonoscopic images and image triplets used in the experiments. In this table, the number in parentheses expresses the number of patients. Note that there is no duplication of data in patient level among the real and virtual images, and training, validation, test-A, and test-B data.

	training	validation	test A	test B
# of real images	9,189 (30)	—	6000 (11)	4302 (1)
# of virtual images	8,085 (7)	3064 (2)	—	—
# of image triplets	29,693 (27)	3299 (3)	—	—

4 Experimental results

4.1 Validation of training

Firstly, we trained the proposed method using the training data shown in Table 1. In this training, we trained generator F, G and discriminators $D_{\mathcal{R}}, D_{\mathcal{V}}$. We set the sizes of real and virtual images to $256 \times 256 \times 3$ and $256 \times 256 \times 4$, respectively. We then changed the number of channels of the input of $G, D_{\mathcal{V}}$, and the number of the output of F to four from the original Cycle GAN, where the number of channels is three. For $F, G, D_{\mathcal{R}}$ and $D_{\mathcal{V}}$, the setting for sizes, numbers, stride width and padding of kernels is the same to the original CycleGAN. We used minibatch size of 64 and base learning rate $lr = 0.002$ with Adam optimiser for 300 epochs. We used the data argumentation with random flipping for each image in a minibatch. For virtual depth images in training and validation data, we computed the mean absolute error per pixel between the original depth and the cyclically translated depth obtained as the fourth channel of $F(G(\mathbf{Y}))$ for a virtual colonoscopic image \mathbf{Y} at each epoch.

Figure 3(a) illustrates the curves of the mean absolute errors. Figures 3(b) and (c) show the examples of the original virtual images and their cyclically translated depth images at epochs. Figure 4(a) shows an example of the cyclical translation of $G(F(\mathbf{X}))$ for a real colonoscopic image \mathbf{X} . We translated all images in the test-A data with the trained model at 300 epochs. Figure 5 show examples of the translation results.

4.2 Comparative evaluation

Secondly, we qualitatively evaluated the proposed method by comparing it with the primary previous works such that XDCycleGAN [14], CycleGAN [21], ExtPix2Pix [17], Monodepth2 [13] and SfMLearner [11], since there is no existing benchmark dataset for the quantitative evaluation of depth estimation. ExtPix2Pix and XDCycleGAN are the baseline and the latest colonoscopic depth estimation methods. CycleGAN is the original method on which our proposed method and XDCycleGAN based. SfMLearner and Monodepth2, both of which are based on the Lambertian-reflection assumption and geometrical constraint as explained in section 1, are the baseline and the winner

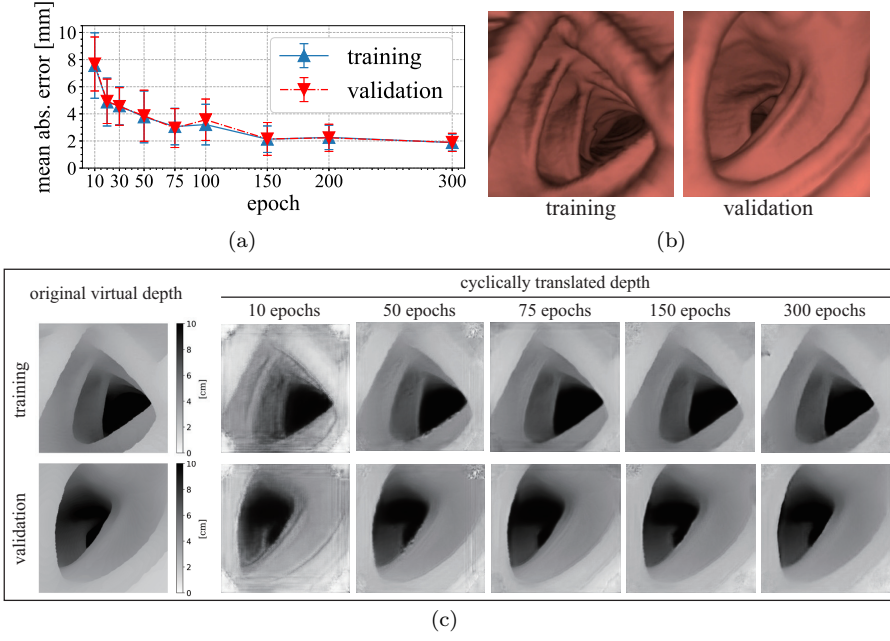


Fig. 3 Qualitative validation of training of the proposed model. (a) Mean absolute error between original and cyclically translated depth at each epoch for the training and validation data. (b) Examples of virtual RGB images. (c) Examples of cyclically translated depth images, which are the fourth channel of the virtual images shown in (b).

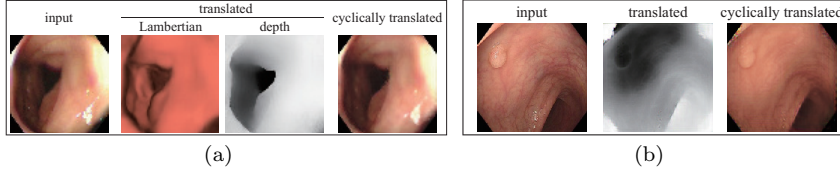


Fig. 4 Examples of correct and incorrect training results. (a) Example of correctly optimised results for the proposed method. (b) Example of incorrectly optimised results for CycleGAN.

of the state-of-the-art methods for mono-view depth estimation, respectively [13]. Note that SfMLearner and Monodepth2 can estimate only relative depth without physical length.

For the training of XDCycleGAN and CycleGAN, which find a domain translation $\mathcal{R}_{\text{RGB}} \rightarrow \mathcal{V}_{\text{D}}$ and its inverse, we used the same training data except for the virtual RGB images of the proposed method. For CycleGAN, we used the same hyperparameters of the proposed method. For XDCycleGAN with the same loss weights to Ref. [14], we started the training with the weights of 200-epoch trained CycleGAN with base learning rate $lr = 1.0 \times 10^{-7}$ for

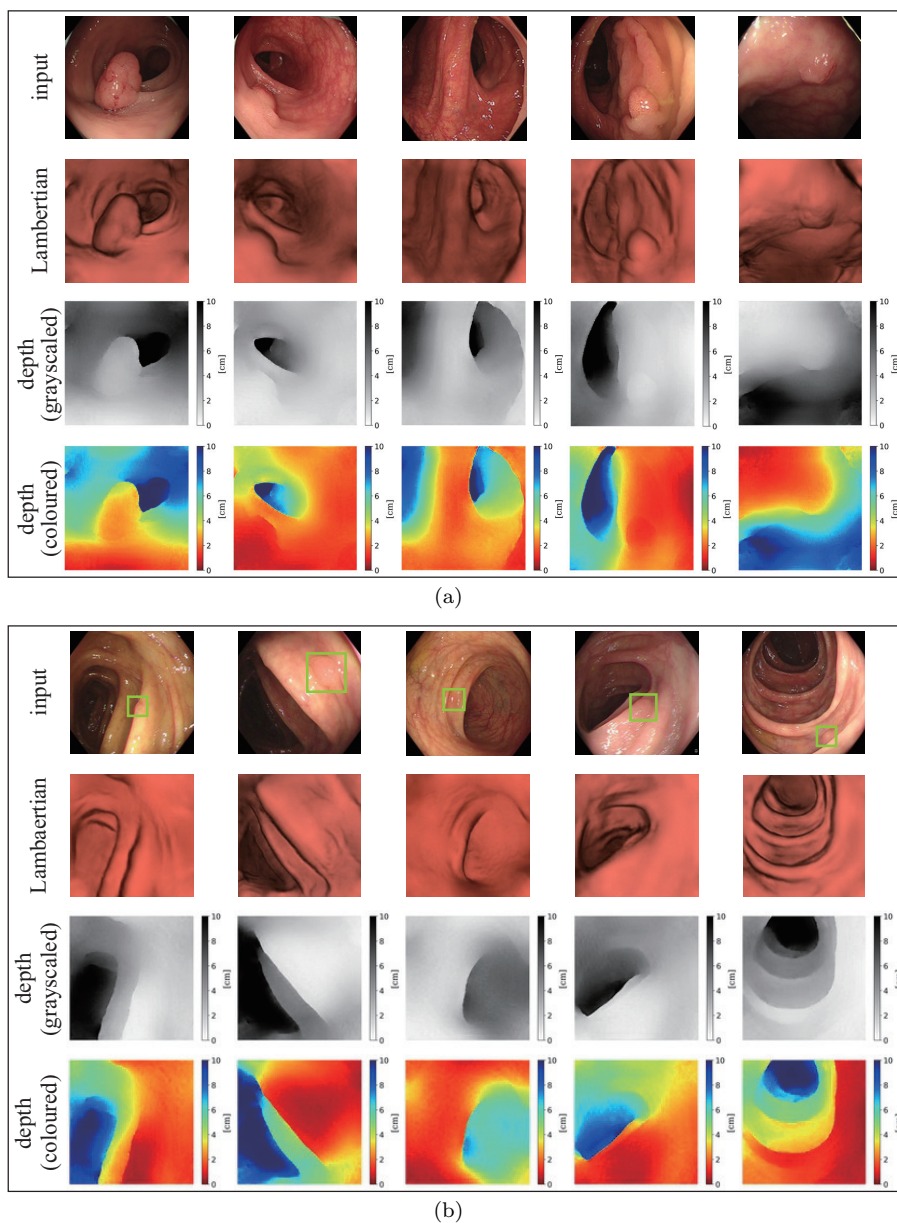
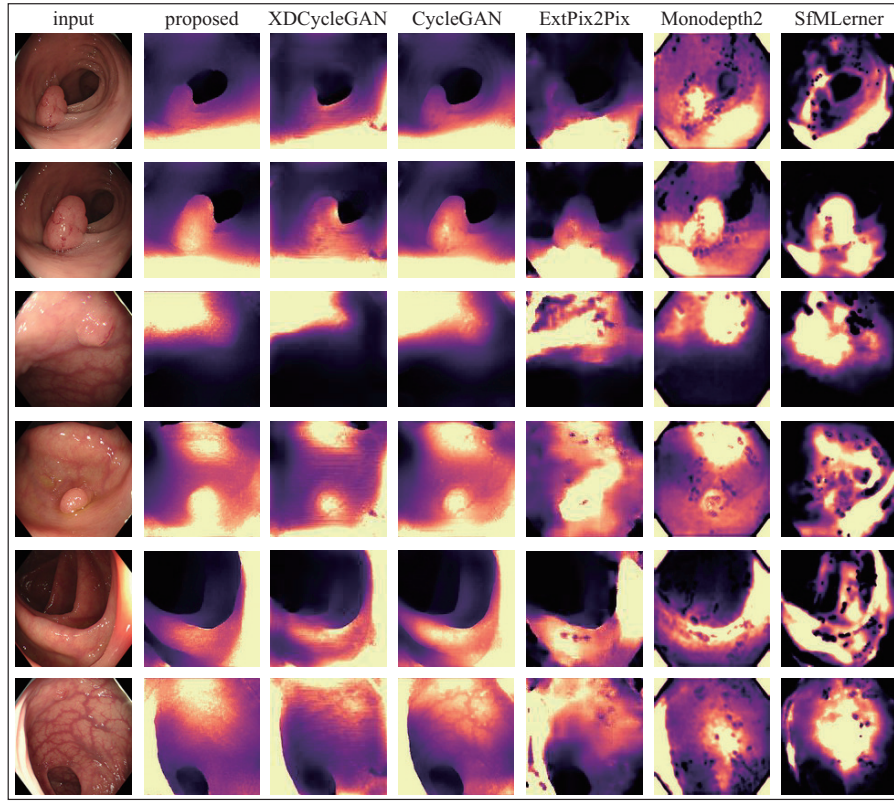
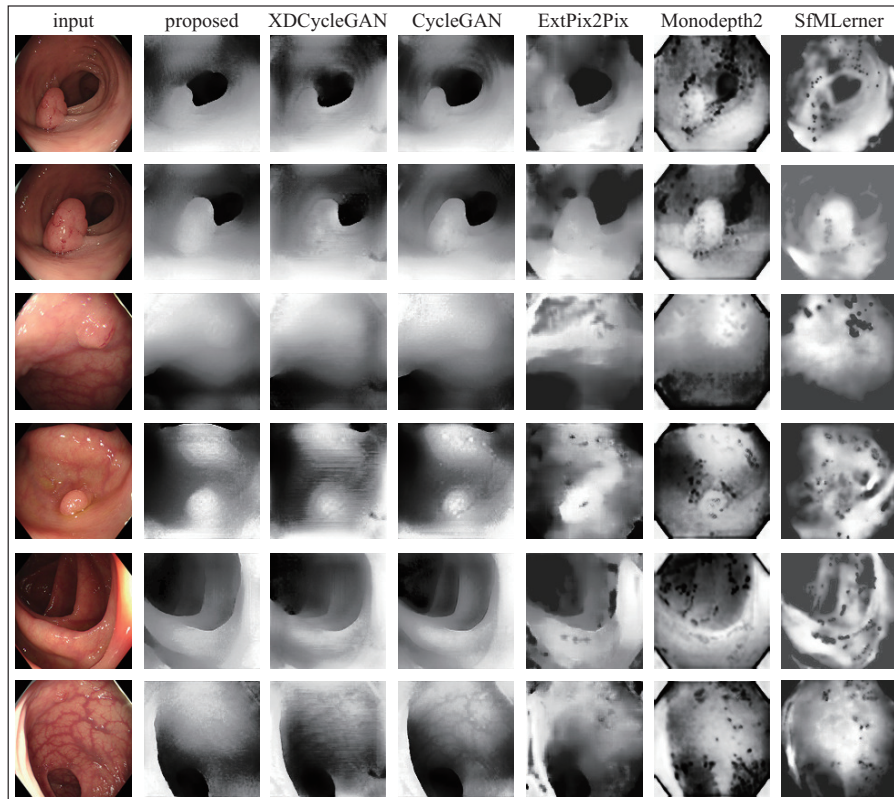


Fig. 5 Examples of domain translation by the proposed method. (a) protruded polyps (6-14 mm). (b) small flat polyps ($\leq 3mm$). In (b), a green box expresses the location of a flat polyp. In (a) and (b) each row shows the following results. The second and third rows show the first to third channel and the fourth channel, respectively, of the translated images. The fourth row shows the fourth channel with seven colours to emphasise the changes in estimated depth.



(a)



(b)

Fig. 6 Comparative evaluations of the proposed and state-of-the-art methods. (a) disparity maps. (b) histogram-equalised images. In (a), brighter colour expresses larger disparity, that is, the closer regions in a depth image and vice versa. In (b), distribution of histogram-equalised depth shows the estimated shape of a colon wall.

Table 2 Accuracy of polyp size measurement in a 3D reconstructed colon wall.

	Case 1	Case 2	Case 3	Case 4	Case 5
G.T. [mm]	3.0	6.0	8.0	12.0	14.0
measured size [mm]	2.1	6.2	8.1	11.0	15.6
absolute error [mm]	0.9	0.2	0.1	1.0	1.6

70 epochs, since the training of XDCycleGAN is unstable. In the training of CycleGAN, we observed incorrect optimisation, as shown in Fig. 4(b). In the incorrect-optimisation results, CycleGAN outputs the inverse depth for near and far points. These results imply the existence of concave/convex ambiguity even in the deep-learning approach. To obtain the correct optimisation, we restarted the optimisation with different random initial values.

For the training of ExtPix2Pix, which also finds a domain translation $\mathcal{R}_{\text{RGB}} \rightarrow \mathcal{V}_{\text{D}}$, we used the real and virtual RGB images in our training data with the same architecture and hyperparameter setting of the original ExtPix2Pix [17]. Note that this method bases on a supervised manner, where each training data is a pair of a virtual RGB image and a depth image. Therefore, compared with CycleGAN, the training of ExtPix2Pix was stable.

For the training of SfMLearner and Monodepth2, we used the sequential-image dataset. We adopted the same hyperparameter settings as the original SfMLearner and Monodepth2, respectively, since these settings achieved the best results for each. We then predicted the depth for the real colonoscopic images in the test-A data by using the three methods. For the comparison shown in Fig 6, we adopted a disparity map and a histogram-equalised image. The disparity map showed an inverse of depth for qualitative evaluation in the same manner as Ref. [13]. The histogram-equalised image highlights effects of the textures and lighting in the same manner as Ref. [14].

4.3 Quantitative evaluation with a measuring device

Thirdly, we presented quantitative evaluation with measuring forceps. To obtain the ground truth depth, we assumed a pinhole-camera-model setting as shown in Fig. 7(a). We can insert a device in the tip of a colonoscope and the device will appear in a test-B data image. We inserted measuring forceps (M2-3U, Olympus, Japan), which has graduations at every 2 mm.

We selected five still images from the test-B data, as shown in Fig. 7(c), where the measuring forceps touch a colon wall. For each still image, we then computed a mean error of depth estimation per pixel in the small region shown with the green bounding box in Fig. 7(c), which represents the region of a colon wall the measuring device touches. Figure 7(b) summarises the mean errors of the bounding-square regions in the depth estimations.

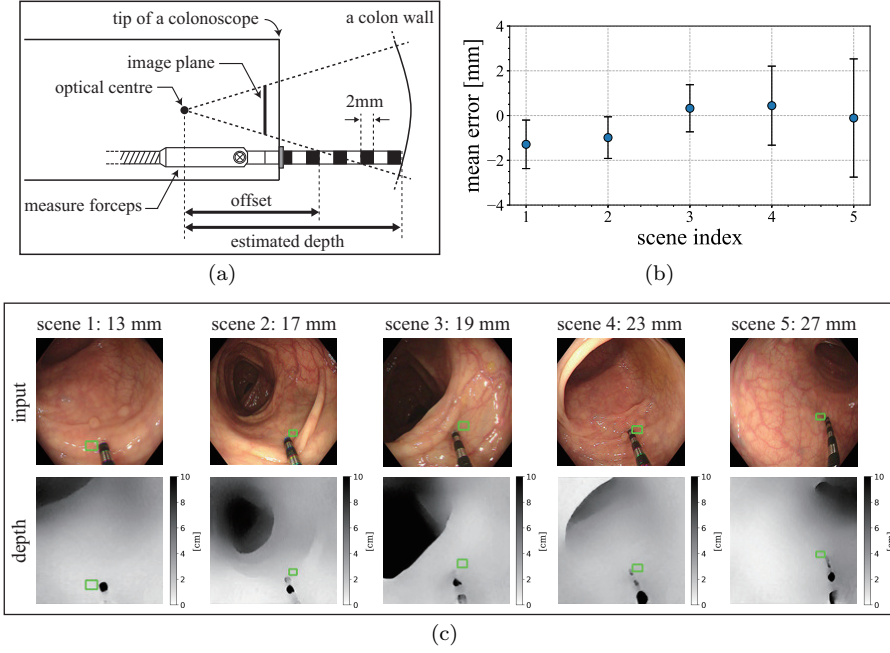


Fig. 7 Qualitative evaluation with measuring forceps. (a) Evaluation setting. Ground truth of depth is defined as the sum of the length of appeared measuring forceps in an image and offset. Maximum length of the measuring forceps is 20 mm. We set the offset at 9 mm based on the pre-calibration for a colonoscope. (b) Mean errors of depth estimation of the target regions. Error bars express standard deviations of errors. (c) Input and estimated depth images.

4.4 Quantitative evaluation with measured polyp sizes

Lastly, we present the quantitative evaluation with the 3D-reconstruction. As preprocessing, we performed calibration [22] of a colonoscope—the same one in the data collection—with a checkerboard and Matlab computer vision toolbox. For the images of the test-A data, we obtained a 3D reconstructed colon wall by reprojection with the inverse of an intrinsic matrix and a scale factor [9], where we set the estimated depth to be the scale factor, and measured the long diameter of each polyp in the reconstructed colon walls.

For the polyps in the test-A data images, expert endoscopists measured their sizes with a device as the ground truth of the polyp sizes. Figure 8 shows the 3D reconstructed colon walls from three viewpoints. Table 2 summarises the ground truth, measured polyp size, and reconstructed error.

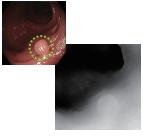
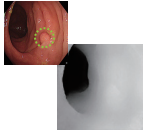
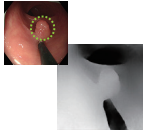
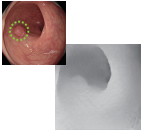
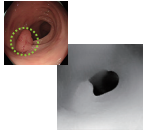
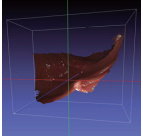
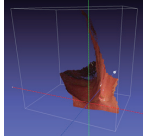
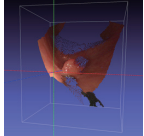
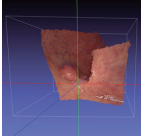
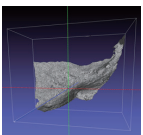
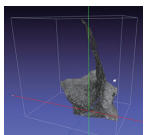
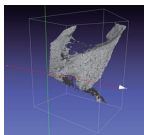
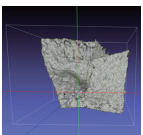
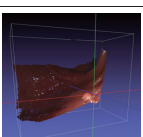
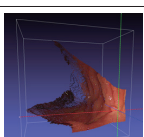
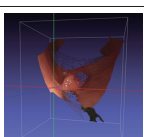
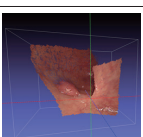
	case 1: 3 mm	case 2: 6 mm	case 3: 8 mm	case 4: 12 mm	case 5: 14 mm
input and estimated depth					
3D reconstruction	right view				
	front view				
	left view				

Fig. 8 Examples of 3D reconstructions with estimated depth. In each view, the two rows show the 3D point reconstruction and surface reconstruction, respectively, by MeshLab [24].

5 Discussion

Figure 3(a) shows the convergence of training of the proposed method in 150-300 epochs. As shown in Fig. 3(c), the depth channel of an input is precisely reconstructed by cyclical domain translation $\mathcal{V}_{\text{RGBD}} \rightarrow \mathcal{R}_{\text{RGB}} \rightarrow \mathcal{V}_{\text{RGBD}}$ after 150 epochs. Figure 5 shows the assumption-satisfied domain translation to RGB-D virtual colonoscopic images. In the second rows of Figs. 5(a) and

5(b), the proposed method outputs smooth and sterical Lambertian-based images without the textures of colon walls. The third and fourth rows in Figs. 5(a) and 5(b) show the smooth and boundary-clear depth distributions by the proposed method. In the fourth row of Fig. 5(b), we can confirm small changes in depth around small folds and flat polyps. These Lambertian and depth images express different contents. A depth image simply shows the distance to the colon wall. A Lambertian image, on the other hand, shows the distribution of reflections and shades. We think that shades on Lambertian images make these images look more sterical than depth images and make depth images look smoother than Lambertian images. The results in Fig. 5 did not reveal any tradeoff in the translation to Lambertian and depth images.

In the seventh column of Figs. 6(a) and (b), SfMLearner failed to capture 3D colon structures. In the sixth column of Figs. 6(a) and (b), Monodepth2 captured the 3D structure of a colon depicting the shapes of polyps and smooth colon walls. In the results of both SfMLearner and Monodepth2, we see many imprecise depth estimations as black spots. These might be caused by non-Lambertian reflection in inputs, because this is a violation of the Lambertian-reflection assumption in estimation. In the fifth column of Figs. 6(a) and (b), ExtPix2Pix captured only a rough 3D structure of a colon and lost its detail shape. Furthermore, we confirmed the black spots on estimation results. These results show that ExtPix2Pix failed to mitigate the difference between real and virtual domains in their depth estimation.

In the second and fourth columns of Figs. 6(a) and (b), the proposed method and CycleGAN achieved smooth and boundary-clear depth estimations. However, some outputs of CycleGAN are affected by textures and specular reflections in an input image. For example, the depth image of CycleGAN shown in the first and second bottom rows affected by the texture of blood vessels and specular reflections, respectively, in each input image. Furthermore, as shown in the third columns in Figs. 6(a) and (b), XDCycleGAN lost the shape details of the colon wall and generated some artefacts in the translation, even though it slightly reduced the effects of textures and specular reflections. On the other hand, comparison of the second columns of Figs. 6(a) and (b) finds almost the same contrast between the histogram-equalised images, which are contrast-enhanced images, and the disparity maps. This implies that the proposed method accurately estimated small depth changes on the colon wall. Consequently, the comparative evaluation results clarified the advantage of the domain-translation approach with our auxiliary task.

In Fig. 7(b), the range of mean errors of depth estimation is only 1.0 mm. Figure 8 shows the realistic 3D reconstructions of colon walls. Furthermore, Table 2 shows that the mean absolute error of measured polyp sizes in these 3D reconstructions is 0.76 mm. Some might say that the convergence of the cyclical translation in Fig. 3(a) doesn't make sense, since GAN can embed imperceptible high-frequency signals into an output for the reconstruction of the original sample [23]. However, the results in Figs. 7(b) and (c), and Table 2, where input images do not include any signal generated by the GANs, demonstrate the validity of our method.

These validation of training, and qualitative and quantitative evaluations clarified that the proposed method achieved the most accurate and valid depth estimation where the auxiliary task reduced the effects of textures by keeping the Lambertian-reflection model, among the state-of-the-art method.

6 Conclusions

We proposed a depth-estimation method by introducing a Lambertian-reflection model as the auxiliary task to a domain translation between real and virtual colonoscopic images towards the developing of a CAD system for colonoscopy. Qualitative evaluations demonstrated the advantages of the proposed method by showing smoother and less corrupted depth distribution in its estimation than those of other state-of-the-art methods. Quantitative evaluations clarified that the proposed method achieves accurate depth estimation with an average error of less than 1 mm for regions close to the colonoscope.

Acknowledgements This study was funded by grants from AMED (19hs0110006h0003), JSPS MEXT KAKENHI (26108006, 17H00867, 17K20099), and the JSPS Bilateral Joint Research Project.

Conflicts of interest

Kudo SE and Misawa M received lecture fees from Olympus. Mori Y received consultant fees and lecture fees from Olympus. Mori K is supported by Cybernet Systems and Olympus (research grant) in this work, and by NTT outside of the submitted work. The other authors have no conflicts of interest.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical committee of Nagoya University (No. 357), and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained via an opt-out procedure from all individual participants included in the study.

References

1. Nadeem S, Kaufman A (2016) Computer-aided Detection of Polyps in Optical Colonoscopy Images. Proc. SPIE 9785, Medical Imaging 2016: Computer-Aided Diagnosis: 549-560
2. Itoh H, Roth, HR, Lu L, Oda M, Misawa M, Mori Y, Kudo S-E, Mori K (2018) Towards Automated Colonoscopy Diagnosis: Binary Polyp Size Estimation via Unsupervised Depth Learning. Proc. Medical Image Computing and Computer Assisted Intervention: 611-619

3. Ma R, Wang R, Pizer S, Rosenman J, McGill, SK, Frahm J-H (2019) Real-Time 3D Reconstruction of Colonoscopic Surfaces for Determining Missing Regions. *Proc. Medical Image Computing and Computer Assisted Intervention*: 573-582
4. Chen, RJ, Bobrow TL, Athey T, Mahmood F, Durr NJ (2019) SLAM Endoscopy Enhanced by Adversarial Depth Prediction. *Proc. KDD'19 Workshop on Applied Data Science for Healthcare*
5. Saxena A, Sung HC, Andrew YN (2006) Learning Depth from Single Monocular Images. *Advances in Neural Information Processing Systems* 18: 1161-1168
6. Eigen D, Fergus R (2015) Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. *Proc. IEEE International Conference on Computer Vision*: 2650-2658
7. Ma F, Karaman S (2018) Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image. *Proc. IEEE International Conference on Robotics and Automation*: 4796-4803
8. Prados E, Faugeras O (2006) Shape From Shading, *Handbook of Mathematical Models in Computer Vision*, Springer: 375-388
9. Hartley R, Zisserman A (2003) *Multiple View Geometry in Computer Vision*. Cambridge University Press
10. Garg R, Vijay Kumar BG, Carneiro G, Reid I (2016) Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. *Proc. European Conference on Computer Vision*: 740-756
11. Zhou T, Brown M, Snavely N, Lowe DG (2017) Unsupervised Learning of Depth and Ego-Motion from Video. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*: 6612-6619
12. Wang C, Buenaposada JM, Zhu R, Lucey S (2018) Learning Depth from Monocular Videos Using Direct Methods. *Proc. IEEE Conference on Computer Vision and Pattern*: 2022-2030
13. Godard C, Aodha OM, Firman M, Brostow G (2019) Digging into Self-Supervised Monocular Depth Estimation. *Proc. IEEE International Conference on Computer Vision*: 3827-3837
14. Mathew S, Nadeem S, Kumari S, Kaufman A (2020) Augmenting Colonoscopy Using Extended and Directional CycleGAN for Lossy Image Translation. *Proc. IEEE International Conference on Computer Vision*: 4695-4704
15. 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 Medical Imaging* 5031: 111-122
16. Faisal M, Nicholas JD (2018) Deep Learning and Conditional Random Fields-based Depth Estimation and Topographical Reconstruction from Conventional Endoscopy. *Medical Image Analysis* 48: 230-243
17. 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. *International Journal of Computer Assisted Radiology and Surgery* 14: 1167-1176
18. Belhumeur PN, Kriegman DJ, Yuille AL (1999) The Bas-relief Ambiguity. *International Journal of Computer Vision* 35(1): 33-44
19. Oda M, Tanaka K, Takabatake H, Mori M, Natori H, Mori K (2019) Realistic Endoscopic Image Generation Method Using Virtual-to-real Image-domain Translation. *IET Healthcare Technology Letters* 6(6): 214-219 v
20. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative Adversarial Nets. In: *Advances in neural information processing systems*: 2672-2680
21. Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proc. IEEE International Conference on Computer Vision*: 2242-2251
22. Zhang, Z. (2000) A Flexible New Technique for Camera Calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(11): 1330-1334
23. Chu C, Zhmoginov A, Sandler M (2017) CycleGAN, a Master of Steganography. *Proc. NIPS 2017 Workshop "Machine Deception"*

-
24. Cignoni P, Callieri M, Corsini M, Dellepiane M, Ganovelli F, Ranzuglia G (2008) MeshLab: an Open-Source Mesh Processing Tool. Proc. Eurographics Italian Chapter Conference.