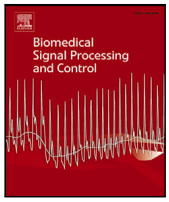
Biomedical Signal Processing and Control 95 (2024) 106410

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Self-supervised endoscopy depth estimation framework with CLIP-guidance segmentation

Zhuoyue Yang [a,b], Junjun Pan [a,b,*], Ju Dai [b,*], Zhen Sun [c], Yi Xiao [c]

[a] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, 37 Xueyuan Road, Haidian District, Beijing, 100191, China
[b] Department of Strategic and Advanced Interdisciplinary Research, Peng Cheng Laboratory, No. 2 Xingke 1st Street, Nanshan District, Shenzhen, Guangdong Province, 518000, China
[c] Division of Colorectal Surgery, Department of General Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, No. 1 Shuai Fu Yuan, Dongcheng District, Beijing, 100730, China

ARTICLE INFO

Keywords:
3D reconstruction
Endoscopy
Depth estimation
Semantic segmentation
CLIP

ABSTRACT

Background and objective: Depth estimation has very broad potential in medical image analysis and is important for applications such as augmented reality surgical navigation and preoperative planning. Compared with segmentation tasks that can obtain ground truth through manual annotation, it is difficult to obtain a large number of real values for depth estimation tasks that are limited by hardware conditions in endoscopic environments.

Methods: To address the challenge, we propose a novel framework that utilizes segmentation tasks to improve encoder performance in a self-supervised depth estimation network. For the first time, we leverage the Contrastive Language-Image Pre-training (CLIP) method to improve the performance of endoscopy segmentation models. Depth estimation networks can also benefit from this training process indirectly. In addition, we design a semantic-guidance loss function to improve the performance.

Results: Our proposed method is systematically evaluated on three datasets. Experiments have verified that the proposed framework can assist the network model in obtaining smaller errors. Compared with other state-of-the-art methods, our framework obtains 0.081 and 0.097 on absolute relative error metrics in quantitative evaluations on SCARED and SERV-CT datasets respectively. In qualitative experiments on real surgery datasets, our proposed method also shows more ideal results.

Conclusion: The experiments in this study illustrate that our proposed method can alleviate the problem of difficulty in improving network performance due to the lack of real values of depth data. The visual performance of our approach illustrates the application potential in the clinic. Our method helps doctors obtain depth perception and visual cues simultaneously, thereby reducing the difficulty of surgery and the pain of patients.

1. Introduction

Minimally invasive endoscopic surgery has been widely used in recent years due to less bleeding and shorter recovery period compared to open surgery. However, due to the narrow field of view and lack of depth perception, endoscopic surgery places stringent demands on surgeons' experience and skills. Nowadays, with the rapid development of virtual reality (VR) or augmented reality (AR) technology, an increasing number of researchers choose AR-based surgical navigation to address these difficulties. For the current AR-based surgical navigation, most work focuses on a single task during the procedure. But laparoscopic surgeries lasting several hours, surgeons need more guidance to interact with the navigation system more efficiently and comfortably. Multi-task navigation systems [1,2] utilizing multi-task learning method [3] to provide multiple auxiliary information are urgently needed but relatively rare at present. In multiple tasks of AR surgical navigation, depth estimation, and semantic segmentation play a very crucial role. Moreover, surgical robotics, surgical planning assistance, and instrument recognition can benefit from the results of depth estimation and semantic segmentation.

Deep learning-based methods in depth estimation and semantic segmentation tasks have grown rapidly in recent years. This is attributed to the acquisition and distribution of large-scale datasets. In endoscopic environments, semantic segmentation tasks can obtain true values by

* Corresponding authors.
E-mail addresses: pan_junjun@buaa.edu.cn (J. Pan), daij@pcl.ac.cn (J. Dai).

https://doi.org/10.1016/j.bspc.2024.106410
Received 24 December 2023; Received in revised form 24 March 2024; Accepted 4 May 2024
Available online 15 May 2024

manually labeling key images. However, the ground truth of depth estimation in the endoscopic environment is difficult to collect. Acquiring the ground truth of depth inside the human body has extremely high requirements on hardware, which are rarely reported in the literature. Most of the existing published datasets with high quality are collected on pig cadavers or in vitro organs. The inability to obtain large-scale depth truth values is the first challenge. The second challenge is that the in vivo environment of each patient has specific lesions. Such idiosyncratic lesions cannot be reflected in the dataset in time. We hardly obtain data captured from the human body, and there is still a gap between animal and human experiments.

The purpose of the depth estimation task in endoscopic scenes is to estimate the distance between the soft tissue surface and the camera lens. The goal of the segmentation task is to annotate the tissue and location of the lesion pixel by pixel based on the captured image. Fortunately, depth estimation and semantic segmentation tasks belong to the dense prediction task and both take images as input. There is a strong connection between these two tasks, for example, most networks follow the encoder–decoder structure. On the one hand, the estimated depth image can be applied to anatomical location recognition [4]. On the other hand, semantic segmentation can improve the accuracy of depth estimation explicitly or implicitly. Currently, most work focuses on the design of decoders [5,6]. Some works have confirmed that the performance of the encoder responsible for feature extraction can bring greater improvements to downstream tasks [7,8].

To address the above challenges, we propose a self-supervised multi-task learning framework to improve the performance of the depth estimation network indirectly through a segmentation task, which makes it easy to obtain the true value. The depth estimation and segmentation task share the weight of the image encoder. In the absence of large-scale datasets with ground-truth values, the performance of depth estimation methods is also improved using knowledge distilled from segmentation tasks. For the first time, we apply the Contrastive Language-Image Pre-training (CLIP) method to the semantic segmentation task in endoscopy. The segmentation network takes the corresponding examination image and word prompts as input. Our framework leverages a large number of patient-specific data in the clinical. In addition, we generate 3D annotations based on the results of deep estimation and semantic segmentation. The 3D annotations assist the surgeons with the location of vital organs. In addition, these annotations can provide the surgeon with depth perception and prevent instrument mistouching. The reconstruction of surgical scenes with segmentation can be employed for postoperative review, intra-operative planning, and surgeon training.

Our contributions are as follows:

- We propose a framework that combines the endoscopic depth estimation task with semantic segmentation by improving the performance of the image encoder. The generalization ability of the encoder benefits from the training process of segmentation.
- We introduce the contrastive language-image pre-training (CLIP) training strategy into the field of endoscopic image segmentation for the first time, fully utilizing the correspondence between surgeons' text diagnosis and case images, thereby benefiting from broader supervision.
- We design a novel loss function that performs domain smoothing for different physiological structures based on the semantic segmentation mask to improve accuracy. This comes from our observation that areas belonging to the same organ generally have consistent depths.

2. Related work

In this section, we review the relevant work on depth estimation, semantic segmentation, multi-task learning, and prompt learning that are closely related to our work.

2.1. Depth estimation

With the release of datasets on natural scenes providing a large number of ground truths, depth estimation methods based on deep learning have been rapidly developed. However, due to the limitations of hardware equipment and surgical specifications, the number of endoscopic datasets with ground truth is very small. Facing the challenge of lack of real values, existing mainstream methods start from two angles. Some authors use generative networks to generate pseudo-real values and perform a transformation [9,10]. However, the gap between virtual data and real data in this kind of network is difficult to overcome. Some researchers focus on using unsupervised methods to solve this problem. Godard et al. [11] use the left and right pictures as constraints to train the convolutional network to obtain consistent 3D information. Zhou et al. [12] utilize the similarity between adjacent images in the image sequence as supervision to train the network. This core idea has been adopted by most subsequent methods because of its effectiveness [7,13]. Turan et al. [14] firstly apply this method to the endoscopic depth estimation task. The network follows encoder–decoder architecture and utilizes the ResNet as the encoder. [7]. Recent work has improved on illumination changes and low accuracy in the endoscopic environment, such as increasing reliance on structure from motion (SfM) [15], reducing the impact of reflection through affine transformation [16], using optical flow to perform photometric correction [17], and add long short term memory (LSTM) module to the network structure to improve the pose network [18]. Shao et al. [17] train the optical flow network and an appearance flow network to calibrate the rotation, translation, and illumination changes. Existing methods mainly improve model performance by changing the network structure or adding modules and are still limited by the quality of the dataset. Our framework addresses the problem of lack of ground truth data from another perspective. We utilize segmentation networks, which have easier access to data, to drive the performance of depth estimation tasks based on the similarity between these networks.

2.2. Semantic segmentation

In the endoscopic environment, the results of depth estimation can promote some semantic tasks, such as polyp detection, segmentation, and tracking. Itoh et al. [19] propose a method to improve the accuracy of polyp classification by using depth estimation information and conducting quantitative and qualitative evaluation through different types of polyps. Jonmohamadi et al. [20] present the first knee arthroscope 3D semantic mapping system. The segmentation network and the depth estimation network are separated, and the segmentation results are directly mapped to the depth estimation. Celik et al. [21] use an unsupervised adaptive technology, which can further improve the performance of gastrointestinal polyps. Transunet [22] is proposed to enhance details by combining the advantages of both transformer [23] and U-Net [24]. To summarize, the semantic segmentation network has been further developed by the massive release of datasets and improvements in the network. Different from existing methods, we try to use medical text annotation information and visual information together as supervisory information, which takes advantage of multi-modal features in our proposed framework.

2.3. Multi-task training

Currently, most work focuses on the design of decoders for multi-task training. Klingner et al. [5] present a method that predicts dynamic objects through a segmentation mask and uses them to guide the depth estimation network to solve the problem of inconsistent lighting in dynamic environments. Jung et al. [6] utilize the transformer block to interact between two network branches. However, it is confirmed that the performance of the encoder responsible for feature extraction can bring greater improvements to downstream tasks. Psychogyios

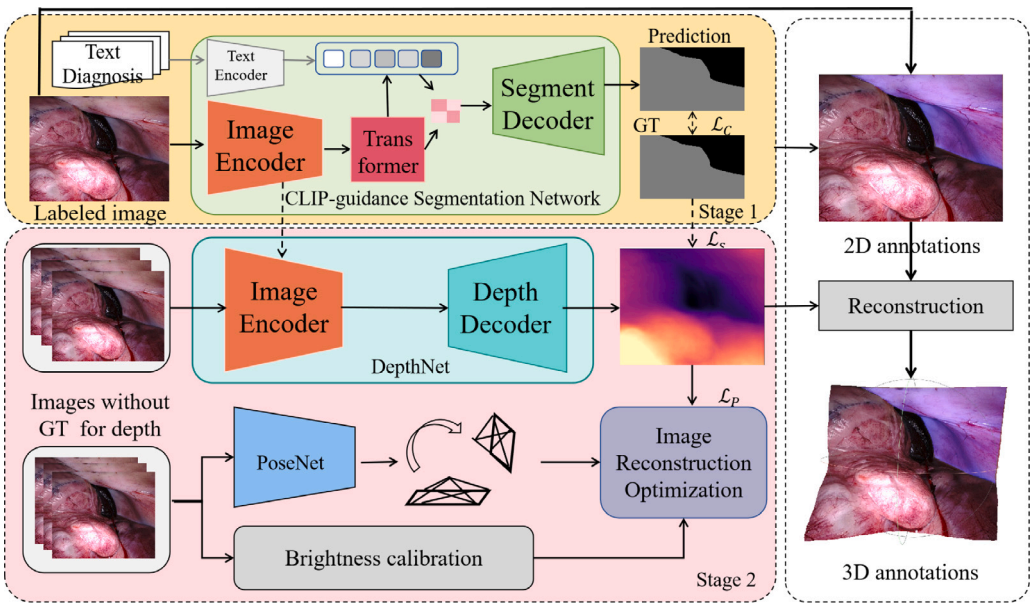


Fig. 1. The pipeline of our proposed framework. We utilize the structural similarity between depth estimation networks and semantic segmentation models to solve the problem of difficulty in obtaining ground truth for depth estimation tasks through semantic segmentation pretraining.

et al. [25] propose a learning framework for joint disparity estimation and device segmentation. Mahjourian et al. [26] and Huang et al. [27] leverage 3D point cloud alignment loss to establish geometry consistency between adjacent frames. For multi-task learning methods, the combination of depth information and segmentation information is mostly used to solve the impact of moving objects such as surgical instruments. In our proposed method, the segmentation task is used to improve the performance of the encoder, which can overcome the challenge of limited data.

2.4. Prompt learning

Contrastive Language-Image Pre-training (CLIP) models have made great strides in learning with fewer samples. Radford et al. [28] firstly convert a pre-training task into a text and picture correspondence problem, thus breaking the object class restriction. CLIP shows better generalizability for downstream tasks and datasets with a wider range of supervised sources [29]. Zhou et al. [30] present the Context Optimization (CoOp) method for word adjustment, which uses learnable vectors to model contextual words for cues and achieves better generalization ability. Rao et al. [29] apply the knowledge from CLIP pre-training to dense prediction tasks such as semantic segmentation, object detection, and instance segmentation through a novel framework. Our framework utilizes the CLIP model to improve the performance of endoscopic segmentation based on the correspondence between diagnostic text and endoscopic images, thus indirectly increasing the accuracy of predicted depth. The proposed method validates the applicability of the CLIP strategy on endoscopic images.

3. Method

In this section, we first describe the two-stage training framework. Then, each module in the pipeline and the corresponding network are introduced. The proposed semantic-aware smooth loss function is illustrated below. Finally, we provide training strategy and implementation details.

3.1. Framework

We propose a two-stage training framework that can improve the performance of a depth estimation task by pre-training on a semantic segmentation network. Most of the works follow the encoder–decoder architecture. It is common to utilize the convolutional network as the image encoder for feature extraction. Some work has shown that encoder improvements can affect the performance of depth estimation networks [7,8]. Utilizing the structural similarity between depth estimation networks and semantic segmentation models, we present a self-supervised framework to deal with the lack of ground truth in endoscopy. Inspired by the state-of-the-art methods, we leverage the DenseCLIP [29] network to get comparatively better performance in semantic segmentation. In addition, a semantic guidance smoothness loss is designed. Therefore, we use semantic information to improve the results of depth estimation.

The framework is shown in Fig. 1. In the first stage, our network is trained on endoscopic images in a supervised manner from a pre-trained weight. We use prompts and pictures as the input and fed them into the segmentation network. The loss between the model's output and the manually annotated ground truth is calculated. The performance of medical endoscopy image segmentation is improved by the text diagnosis. In the second stage, the depth estimation network (DepthNet) takes a single endoscopic image as the input to generate corresponding depth information. The image encoders in DepthNet and the segmentation network share the weights. Adjacent endoscopic images are fed into the pose estimation network (PoseNet). Then PoseNet generates the pose matrix between the two images. The estimated pose and depth are combined with the camera's internal parameters and re-projected back onto the plane to generate a reconstructed image. In this way, the depth network and the pose network can be implicitly constrained based on the similarity between the reconstructed image and the original image. In particular, we use the encoder trained by the segmentation tasks as the initial value for the depth estimation network. The pose network and depth network are trained by photometric loss and smoothness loss simultaneously in a self-supervised manner. In addition, we utilize the segmentation mask as a condition to improve the performance of the depth estimation network. We introduce the methods utilized in these two stages below. According to the depth map and the segmentation mask, the 3D annotation with different colors is reconstructed.

3.2. Language prompt learning for segmentation

The CLIP model is originally applied to classification problems mainly by matching text and images to benefit from wider and richer data supervision. DenseCLIP [29] is the first method that employs the CLIP model for dense estimation tasks such as segmentation and detection. In stage one, the entire model includes a text encoder, an image encoder, a Transformer and a segmentation decoder. Among them, the text encoder, image encoder, and image decoder are all existing models and minor adjustments have been made to them. The core idea of the entire method is to match the pixels of the image with the text. Multi-head self-attention [23] in Transformer is utilized to the feature map of the image encoder to obtain language-compatible features. The product of image features and text features is used as a score map of similarity. This similarity score is passed to the decoder as a low-resolution feature cascaded with the features of the original encoder. The score map also represents a low-resolution result supervised through the segmentation task. In our framework, we use the names of some soft tissue structures as prompts. The prompts and endoscopic images are paired and fed into the model together to perform the segmentation task of specific soft tissue structures. The text encoder generates the text features and the image encoder obtains visual features. The similarity of text features and visual features enhances the information and is transferred to the decoder. An image encoder suitable for endoscopic images is obtained. Segmentation of soft tissue requires certain professional knowledge. Therefore, we utilize the strategy that combines a few annotations and a supervised learning network to generate pseudo labels. Specifically, we create pseudo labels through manual annotation and the existing network (U-Net [24]). Finally, we manually check all pseudo labels. The dimensions of the network are changed according to the labeled categories.

3.3. Semantic-guidance self-supervised depth estimation

DepthNet, PoseNet and brightness calibration module are included in stage two. DepthNet contains an image encoder and a depth decoder. The design of the DepthNet and PoseNet is the same as Monodepth2 [7]. The brightness calibration module is the pre-trained network presented in [31], including an appearance flow network and an optical flow network. We define the self-supervised depth estimation problem as minimizing the image reconstruction loss function between the target image and the re-projected image [12,7]. The image reconstruction loss is composed of the photometric loss ($\mathcal{L}_p$) and edge-aware loss. The photometric loss ($\mathcal{L}_p$) minimizes the image similarity function ($\mathcal{F}$) among two images with a visibility mask [31,7]. The use of photometric loss functions in endoscopic picture reconstruction problems may accumulate errors. The movement of cameras and reflections from smooth surfaces of soft tissue violate the photometric invariance assumption. Thus, the brightness calibration is applied to supplement the illumination.

The source image is defined as $\mathbf{I}^{\dagger}$. The reconstructed image ($\tilde{\mathbf{I}}$) is defined as follows:

$$\tilde{\mathbf{I}} = \pi(\mathbf{I}^{\dagger}, \mathbf{T}, \mathbf{D}, \mathbf{P}), \tag{1}$$

where $\mathbf{T}$ is the pose estimation, $\mathbf{P}$ represents the camera intrinsics, $\mathbf{D}$ is the predicted depth and re-projected function (π). After the brightness calibration, the modified image ($\hat{\mathbf{I}}$) is as follows:

$$\hat{\mathbf{I}} = \mathbf{I} + \mathbf{C}, \tag{2}$$

where $\mathbf{C}$ is the output of the pre-trained appearance flow network. The image similarity ($\mathcal{F}$) between the modified image ($\hat{\mathbf{I}}$) and the reconstructed image ($\tilde{\mathbf{I}}$) is calculated as follows:

$$\mathcal{F} = \beta \cdot \frac{1 - \mathrm{SSIM}(\hat{\mathbf{I}}, \tilde{\mathbf{I}})}{2} + (1-\beta) \cdot \left| \hat{\mathbf{I}} - \tilde{\mathbf{I}} \right|, \tag{3}$$

where SSIM is the structural similarity index [32] and $\beta = 0.85$ [17,33]. We also use the edge-aware loss following [17,12].

To emphasize consistency and smoothness within the same semantic mask, our semantic-guidance smooth loss ($\mathcal{L}_s$) is defined as:

$$\mathcal{L}_s = \mathbf{M}(|\partial_x d|e^{-|\partial_x \mathbf{I}|} + |\partial_y d|e^{-|\partial_y \mathbf{I}|}), \tag{4}$$

where d represents the mean-normalized inverse depth of $\mathbf{I}$ and $\mathbf{M}$ is the mask provided from the first stage. ∂d and $\partial \mathbf{I}$ are the gradients of disparity and image, respectively.

3.4. Surface reconstruction and annotation display

The segmentation results are combined with point clouds generated by the depth estimation network to form 3D annotations. 3D annotations show some physiological structures through different colors, such as the abdominal wall, liver and kidney, etc. 3D annotations can provide depth information more intuitively, and also relieve surgeons' visual fatigue. 3D surface reconstruction and annotation display can be completed together through our method. Firstly, masks of different colors are generated based on the segmentation results. We then overlay the semi-transparent masks onto the corresponding original endoscopic image. Finally, the point cloud with a specific color mask can be recovered using camera intrinsics and depth estimates to display the geometric structure. The truncated signed distance function (TSDF) [34] is applied to fuse multiple point clouds to extend the 3D model of the tissue surface. The implementation is developed by Open3d [35] according to [13].

4. Experiments

4.1. Experiment setup

We utilize the SCARED [36] dataset and SERV-CT [37] dataset to evaluate our methods' performance. The SCARED dataset contains 9 different sub-datasets collected from porcine cadavers and the SERV-CT dataset includes 16 image pairs and CT. We can evaluate the performance depth estimation methods using these datasets. Following [38, 31], 20,664 and 2991 images are used for training and validation, respectively. And 517 images are utilized for evaluation. The input pictures are uniformly scaled to the size of 320 × 256 and are collected from fresh porcine cadaver abdominal anatomy. To validate the generalization performance of the model, we test it on the laparoscopic dataset. Our laparoscopic dataset is collected under the guidance of doctors and complies with data privacy regulations and ethical standards. This dataset contains videos taken by laparoscopes without ground truth. The test data is not involved in the training stage. The model is not fine-tuned on the test dataset. For evaluation, following [7,31], we compute the five standard metrics: Abs Rel (absolute relative error), Sq Rel (square relative error), RMSE (root mean square error), RMSE log (root mean square logarithmic error), δ. These metrics are defined as follows:

$$\text{Abs Rel} = \frac{1}{|\mathbf{D}|} \sum_{d \in \mathbf{D}} |d^* - d| / d^* \tag{5}$$

$$\text{Sq Rel} = \frac{1}{|\mathbf{D}|} \sum_{d \in \mathbf{D}} |d^* - d|^2 / d^* \tag{6}$$

$$\text{RMSE log} = \sqrt{\frac{1}{|\mathbf{D}|} \sum_{d \in \mathbf{D}} |\log d^* - \log d|^2} \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{1}{|\mathbf{D}|} \sum_{d \in \mathbf{D}} |d^* - d|^2}, \tag{8}$$

$$\delta = \frac{1}{|\mathbf{D}|} \left| \left\{ d \in \mathbf{D} | max(\frac{d^*}{d}, \frac{d}{d^*}) < 1.25 \right\} \right| \times 100\% \tag{9}$$

where $\mathbf{D}$ is the set of the predicted depth. d and d^* denote the predicted depth and the ground truth, respectively. In the first stage, the model is trained with AdamW [39] optimizer. In the second stage, we train the model with a minibatch of 12 for 30 epochs. Image augmentation is applied during training, such as random horizontal flipping and random color augmentation with the settings form.

Table 1
DepthNet performance on SCARED dataset. 'M' means monocular dataset. 'S' means semantic dataset..

Method	Backbone	Strategy	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	δ ↑
Monodepth2 [7]	Resnet50	M	0.086	0.811	14.751	0.107	0.952
Endo-SfM [16]	Resnet50	M	0.084	0.672	6.102	0.103	0.959
AF-SfM [17]	Resnet50	M	0.083	0.651	6.058	0.102	0.964
Ours	Resnet50	M	**0.081**	**0.647**	**6.016**	**0.103**	**0.957**
SGD-Depth [5]	Resnet50	M+S	0.089	1.007	7.312	0.112	0.941
FSRE [6]	Resnet50	M+S	0.085	0.741	14.473	0.108	0.948
Ours	Resnet50	M+S	**0.081**	**0.625**	**5.941**	**0.102**	**0.961**

Table 2
DepthNet performance on SERV-CT dataset. 'M' means monocular dataset. 'S' means semantic dataset..

Method	Backbone	Strategy	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	δ ↑
Monodepth2 [7]	Resnet50	M	0.098	1.442	10.419	0.123	0.915
Endo-SfM [16]	Resnet50	M	0.105	1.678	11.004	0.133	0.893
AF-SfM [17]	Resnet50	M	0.104	1.675	11.374	0.133	0.900
SGD-Depth [5]	Resnet50	M+S	0.123	1.988	12.123	0.156	0.847
FSRE [6]	Resnet50	M+S	0.118	2.007	12.744	0.157	0.850
Ours	Resnet50	M+S	**0.097**	**1.305**	**10.043**	**0.123**	**0.920**

4.2. Results

The performance of depth estimation is reported in this section, regarding the prompt pretraining model and the ablation experiment of semantic smoothness loss function. The performance of depth estimation on surgical laparoscopic images is also calculated. At the same time, we also conduct qualitative analysis, mainly including depth estimation, segmentation, and visualization results of point clouds. The results of augmented reality applications are collected. The limitations of our method are discussed.

4.2.1. Depth estimation

DepthNet network takes endoscopic images as input and predicts the depth value corresponding to each pixel. We perform error statistics between the predicted depth value and the ground truth. The accuracy of depth estimation results is reported on various metrics. Firstly, we evaluate the accuracy of our framework with three classic self-supervised learning methods, including Monodepth2 [7], Endo-SfM [16] and AF-SfM [17]. The pre-trained weights obtained from the segmentation task are utilized. Secondly, our framework is compared with the SoTA methods (FSRE [6] and SGD-Depth [5]) which also use semantic estimation. Here, our method employs explicit semantic guidance Loss. Existing segmentation tasks using CLIP do not provide pre-trained weights for Resnet18. To be fair, we use Resnet50 as the backbone for each method. In addition, other methods utilize weights pre-trained on ImageNet as initial values. For the ablation study, the performance of those models is collected according to different training strategies.

Table 1 shows the quantitative results of the comparative methods. The input for all comparison methods is 320 × 256 images. By analyzing the results of the first 4 rows in Table 1, we verify that the proposed encoder pre-trained using the segmentation method obtains more ideal results. We also show the visualization results of depth estimation in Fig. 3. Each column corresponds to a scene, which is the original image and the depth obtained by different comparison methods. The results obtained from Monodepth2 [7] are relatively low. We mainly analyze the differences between AF-SfM [17] and our method. As shown in Fig. 3(a) and (b), our method has better performance in terms of global consistency compared with the state-of-the-art method. Our method can obtain clearer depth results at the edges, as shown in Fig. 3(c) and (d). This allows the depth estimation results to display more details. The laparoscopic dataset does not include ground truth, so we only conduct quantitative experiments. As shown in Fig. 4, our method achieves smoother depth estimation results.

We also evaluate our method on the SERV-CT dataset in Table 2. Fig. 2 displays the Abs Rel error distribution on the SERV-CT dataset.

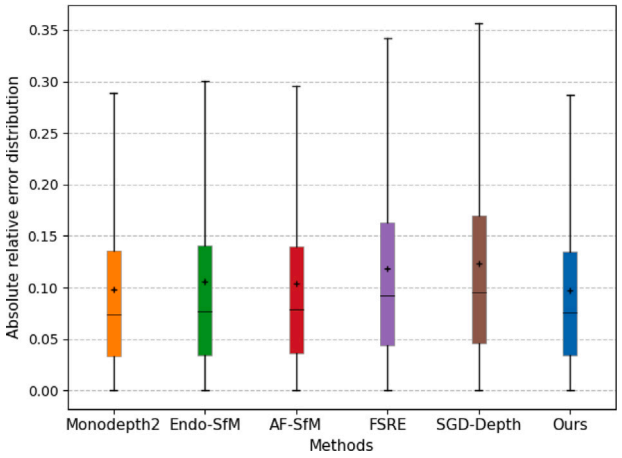


Fig. 2. Boxplot of absolute relative error distribution on SERV-CT dataset for Monodepth2, Endo-SfM, AF-SfM, FSRE, SGD-Depth and our method.

According to Fig. 2 and Table 2, our method also has lower mean and median errors on this dataset. Compared with other methods, the performance of AF-SfM is better because the photometric calibration module could alleviate the impact of lighting changes. Our method surpasses AF-SfM and achieves the smallest error on two datasets, which may benefit from the improvement of encoder performance in segmentation training. Other methods that use semantic information do not have the desired performance, which may be due to the significant difference between the endoscopic and natural scenes.

4.2.2. Ablation study

We verify the effectiveness of the proposed method through ablation experiments. Table 3 reports the experimental results using different pre-trained model weights and loss functions on the basic network structure. We still use the five metrics (Abs Rel, Sq Rel, RMSE, RMSE log, and δ) to reflect the performance of the method. The baseline uses the pre-trained model of ImageNet. In the header of Table 3, 'CLIP-Pretrain' represents that the CLIP-based pre-trained model weight is used. 'CLIP-Seg Pretrain' means we utilize the CLIP strategy to train the segmentation network. And 'L_s' indicates that the semantic segmentation loss function is used in the training stage. The difference between the second and the fourth row is that the second one utilizes the original CLIP model, while the fourth row employs the weight obtained from endoscopy segmentation.

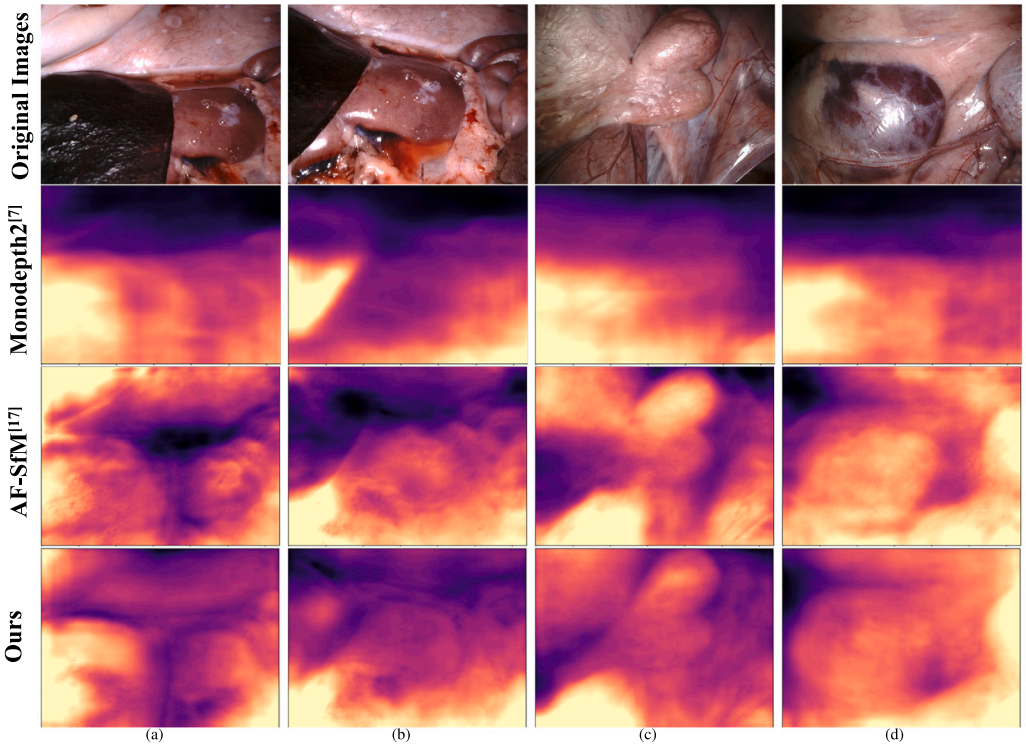


Fig. 3. Visualization results of depth estimation networks on SCARED dataset. (a), (b), (c), and (d) are four respective images and the predicted depth obtained from other SoTA methods and our framework.

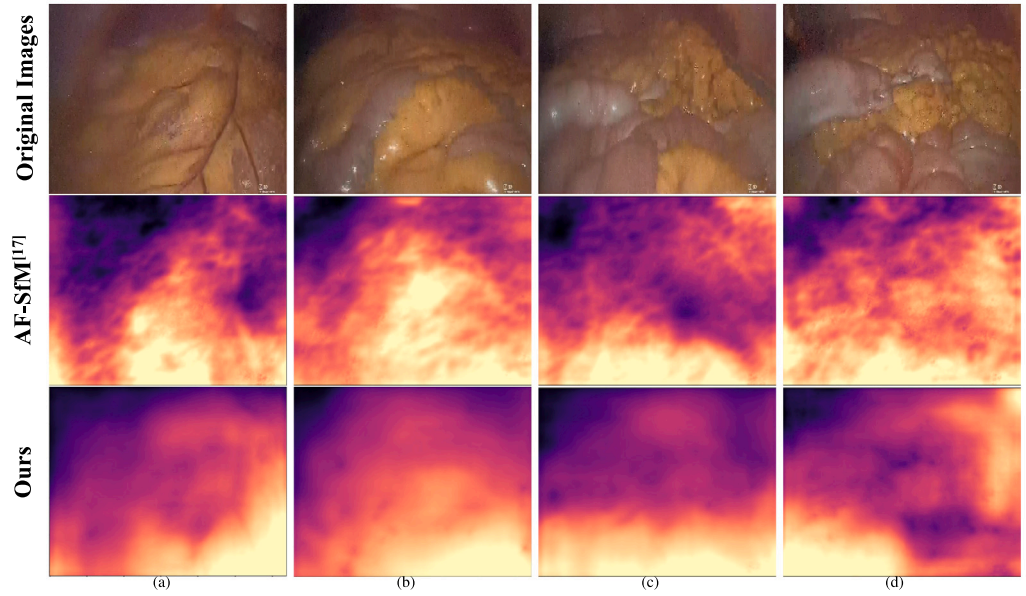


Fig. 4. Visualization results of depth estimation on laparoscopic images. (a), (b), (c), and (d) are four respective images and the predicted depth obtained from other SoTA methods and our framework.

Table 3
DepthNet ablation.

Method	Pretrain	L_s	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	δ ↑
Baseline			0.083	0.651	6.058	0.102	0.964
With CLIP Pretrain	✓		0.087	0.759	6.518	0.110	0.948
With CLIP Pretrain and L_s	✓	✓	0.085	0.729	6.371	0.109	0.948
With CLIP-Seg Pretrain	✓		0.081	0.647	6.016	0.103	0.957
With L_s and CLIP-Seg Pretrain	✓	✓	0.081	0.625	5.941	0.102	0.961

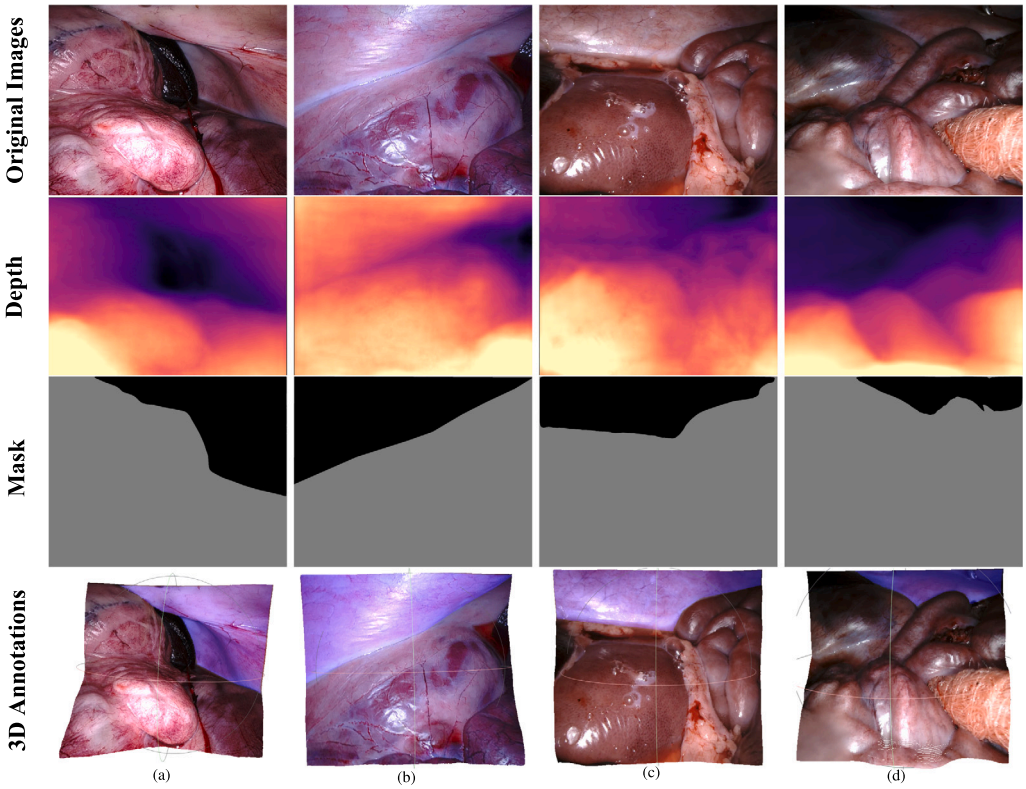


Fig. 5. Visualization results of depth estimation, semantic segmentation masks, and point clouds with annotations. (a), (b), (c), and (d) show the process of generating four point clouds with different color annotations.

By comparing the results of the first and second rows, we find that directly using CLIP's pre-trained model does not directly improve performance. On the contrary, all metrics have declined to a certain extent. The data set sources of the CLIP model are mostly natural scenes and this could result in a significant decrease in performance. By comparing the results of the first and fourth rows, we verify that using the segmentation-trained encoder for depth estimation improves the performance of the model on most metrics. On this basis, our method improves the accuracy of the model by explicitly using the semantic guidance loss function.

4.2.3. Qualitative evaluation

In qualitative experiments, our experimental goal is to enhance the position of the abdominal wall in the image with purple for display. This goal has practical significance in surgery. During surgery, surgeons generally use instruments to manipulate organs. Displaying the abdominal wall with another color could reduce accidental touching. In addition, the abdominal cavity environment is similar, and large areas with similar textures will cause visual fatigue for surgeons. By marking the abdominal wall purple, we allow the surgeon to focus more on the surgical area. Fig. 5 is a visual display of the intermediate results at each stage of our method and the final point cloud augmented with annotated information. Each column in Fig. 5 represents a scenario, and each row represents the visualization results of different stages. The last line shows the point cloud reconstructed from depth estimation results after labeling the abdominal wall position with a purple label.

4.2.4. Augmented reality application

We propose a self-supervised learning framework that combines depth estimation and semantic segmentation. An important application of this multi-task learning framework is AR navigation. We perform experiments on endoscopic images to demonstrate the potential of our approach. Fig. 6 are two examples of effectively combining depth estimation and segmentation results. Fig. 6(a) is a picture taken of the kidney, and Fig. 6(b) is a picture taken of the liver. If we use traditional annotation methods, we can only obtain 2D annotations, as shown in the two pictures in the second column. 3D annotations with depth information can be generated using a multi-task learning framework, as shown in the last two columns of Fig. 6. Our visualization results demonstrate that a multi-task framework could provide surgeons with more informative interactions.

4.2.5. Discussion

Our framework improves the performance of depth estimation methods with the help of segmentation tasks that easily obtain a large number of ground-truth values. Existing methods have proven that text and image comparison learning can significantly improve model performance, allowing the network to benefit from more supervised data. Based on the above implementation, our inspiration is to improve the performance of segmentation methods by exploiting the correspondences existing in cases during existing endoscopic inspections. We believe that the proposed framework has the potential to achieve greater improvements with the accumulation of matching data in real medical settings. Our experiments also provide preliminary verification of this.

Our proposed method can alleviate the dilemma of lack of ground-truth values. Our experiments show that depth estimation performance can be improved through segmentation annotation of endoscopic images. Using endoscopic images from the Vivo environment, the models trained by our method can be more easily transferred and applied in real surgeries. Because only images are used for input, there is no change to the entire surgical process and it can be easily integrated into existing surgical systems.

However, our approach is not without limitations. The proposed framework is a two-stage model which may result in the accumulation

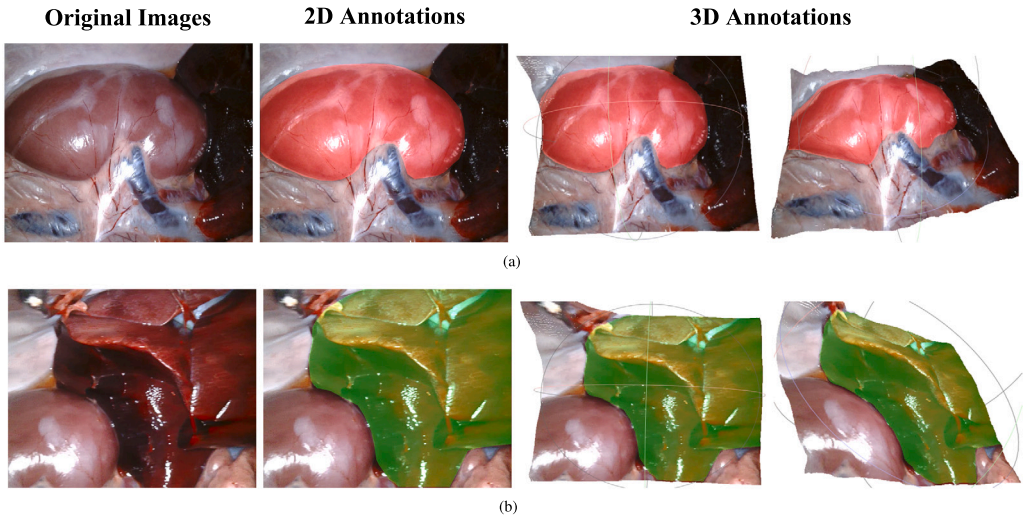


Fig. 6. Visualization of AR applications using our proposed framework. (a) The kidney is marked in red. (b) The liver is labeled in green.

of errors. Furthermore, there is a time gap between the semantic segmentation task and the depth estimation task while updating the data. In the future, we will expand the data scale and build a standard large-scale endoscopic data set. The complementary advantages of the two tasks will be further utilized in designing a new network. We will improve the model on a broader data set and evaluate it in animal experiments.

5. Conclusion

To overcome the weakness of lack of ground truth, we propose a self-supervised framework that leverages semantic information to improve depth estimation performance. In addition, the correspondence between text and pictures that naturally exists in medical cases is regarded as a starting point to improve the performance of segmentation tasks by using CLIP models. In this way, our framework is expected to obtain a large number of human endoscopic images based on existing case data and solve the problem of collecting ground truth of depth estimation. It improves the generalization ability of the model and reduces the gap between the experimental model and the actual surgical environment. Finally, we present a semantically guided smoothness loss function. Experiments confirm the effectiveness of our method. Our method assists doctors in obtaining depth perception and visual cues simultaneously and demonstrates the application potential in the clinic.

CRediT authorship contribution statement

Zhuoyue Yang: Writing – original draft, Visualization, Methodology, Data curation, Conceptualization. **Junjun Pan:** Writing – review & editing, Conceptualization. **Ju Dai:** Writing – review & editing, Conceptualization. **Zhen Sun:** Data curation. **Yi Xiao:** Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research is supported by National Science and Technology Major Project in China (No. 2022ZD0115902), National Natural Science Foundation of China (Nos. U20A20195, 62272017, 62172437, 62102208), Beijing Natural Science Foundation (L232065, L232135). The authors are grateful to Dr. Max Allen in Intuitive Inc. for his assistance in data annotation.

References

[1] Mirza Awais Ahmad, et al., Towards in-utero navigational assistance: A multi task neural network for segmentation and pose estimation in fetoscopy, in: 2023 International Symposium on Medical Robotics, ISMR, 2023, pp. 1–6, http://dx.doi.org/10.1109/ISMR57123.2023.10130205.

[2] Adrito Das, et al., A multi-task network for anatomy identification in endoscopic pituitary surgery, in: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention, MICCAI, 2023, pp. 472–482, http://dx.doi.org/10.1007/978-3-031-43996-4_45.

[3] Rich Caruana, Multitask learning, Mach. Learn. 28 (1997) 41–75, http://dx.doi.org/10.1023/A:1007379606734.

[4] Masahiro Oda, et al., Depth estimation from single-shot monocular endoscope image using image domain adaptation and edge-aware depth estimation, Comput. Methods Biomech. Biomed. Eng. Imaging Vis. 10 (3) (2022) 266–273, http://dx.doi.org/10.1080/21681163.2021.2012835.

[5] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, Tim Fingscheidt, Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance, in: Proc. Eur. Conf. Comput. Vis., ECCV, 2020, pp. 582–600, http://dx.doi.org/10.1007/978-3-030-58565-5_35.

[6] Hyunyoung Jung, Eunhyeok Park, Sungjoo Yoo, Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation, in: Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 12642–12652, http://dx.doi.org/10.1109/ICCV48922.2021.01241.

[7] Clément Godard, Oisin Mac Aodha, Michael Firman, Gabriel J Brostow, Digging into self-supervised monocular depth estimation, in: Proc. IEEE Int. Conf. Comput. Vis., ICCV, 2019, pp. 3828–3838, http://dx.doi.org/10.1109/ICCV.2019.00393.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, CVPR, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[9] Anita Rau, et al., Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy, Int. J. Comput. Assist. Radiol. Surg. 14 (7) (2019) 1167–1176, http://dx.doi.org/10.1007/s11548-019-01962-w.

[10] Cheng Wang, et al., Depth-based branching level estimation for bronchoscopic navigation, Int. J. Comput. Assist. Radiol. Surg. 16 (10) (2021) 1795–1804, http://dx.doi.org/10.1007/s11548-021-02460-8.

[11] Clément Godard, Oisin Mac Aodha, Gabriel J. Brostow, Unsupervised monocular depth estimation with left-right consistency, in: IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2017, pp. 6602–6611, http://dx.doi.org/10.1109/CVPR.2017.699.

[12] Tinghui Zhou, Matthew Brown, Noah Snavely, David G. Lowe, Unsupervised learning of depth and ego-motion from video, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2017, pp. 6612–6619, http://dx.doi.org/10.1109/CVPR.2017.700.

[13] David Recasens, et al., Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints, IEEE Robot. Autom. Lett. 6 (4) (2021) 7225–7232, http://dx.doi.org/10.1109/LRA.2021.3095528.

[14] Mehmet Turan, et al., Unsupervised odometry and depth learning for endoscopic capsule robots, in: IEEE Int. Conf. Intell. Rob. Syst., IROS, 2018, pp. 1801–1807, http://dx.doi.org/10.1109/IROS.2018.8593623.

[15] Xingtong Liu, et al., Dense depth estimation in monocular endoscopy with self-supervised learning methods, IEEE Trans. Med. Imaging 39 (5) (2020) 1438–1447, http://dx.doi.org/10.1109/TMI.2019.2950936.

[16] Kutsev Bengisu Ozyoruk, et al., Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos, Med. Image Anal. 71 (2021) 102058, http://dx.doi.org/10.1016/j.media.2021.102058.

[17] Shuwei Shao, et al., Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue, Med. Image Anal. 77 (2022) 102338, http://dx.doi.org/10.1016/j.media.2021.102338.

[18] Ling Li, et al., Unsupervised-learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery, IEEE Trans. Ind. Inf. 17 (6) (2021) 3920–3928, http://dx.doi.org/10.1109/TII.2020.3011067.

[19] Hayato Itoh, et al., Binary polyp-size classification based on deep-learned spatial information, Int. J. Comput. Assist. Radiol. Surg. 16 (10) (2021) 1817–1828, http://dx.doi.org/10.1007/s11548-021-02477-z.

[20] Yaqub Jonmohamadi, et al., 3D semantic mapping from arthroscopy using out-of-distribution pose and depth and in-distribution segmentation training, in: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention, MICCAI, 2021, pp. 383–393, http://dx.doi.org/10.1007/978-3-030-87196-3_36.

[21] Numan Celik, Sharib Ali, Soumya Gupta, Barbara Braden, Jens Rittscher, Endouda: A modality independent segmentation approach for endoscopy imaging, in: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention, MICCAI, 2021, pp. 303–312, http://dx.doi.org/10.1007/978-3-030-87199-4_29.

[22] Jieneng Chen, et al., Transunet: Transformers make strong encoders for medical image segmentation, 2021, http://dx.doi.org/10.48550/arXiv.2102.04306, arXiv preprint arXiv:2102.04306.

[23] Ashish Vaswani, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 6000–6010, http://dx.doi.org/10.48550/arXiv.1706.03762.

[24] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, in: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention, MICCAI, 2015, pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[25] Dimitrios Psychogyios, Evangelos Mazomenos, Francisco Vasconcelos, Danail Stoyanov, MSDESIS: Multitask stereo disparity estimation and surgical instrument segmentation, IEEE Trans. Med. Imaging 41 (11) (2022) 3218–3230, http://dx.doi.org/10.1109/TMI.2022.3181229.

[26] Reza Mahjourian, Martin Wicke, Anelia Angelova, Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints, in: IEEE Conf. Comput. Vis. Pattern Recognit., CVPR, 2018, pp. 5667–5675, http://dx.doi.org/10.1109/CVPR.2018.00594.

[27] Baoru Huang, Jian-Qing Zheng, Anh Nguyen, Chi Xu, Ioannis Gkouzionis, Kunal Vyas, David Tuch, Stamatia Giannarou, Daniel S Elson, Self-supervised depth estimation in laparoscopic image using 3D geometric consistency, in: Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention, MICCAI, 2022, pp. 13–22, http://dx.doi.org/10.1007/978-3-031-16449-1_2.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., Learning transferable visual models from natural language supervision, in: Proc. Int. Conf. Mach. Learn., ICML, 2021, pp. 8748–8763, http://dx.doi.org/10.48550/arXiv.2304.07039.

[29] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, Jiwen Lu, Denseclip: Language-guided dense prediction with context-aware prompting, in: Proc. IEEE Conf. Comput. Vis. Pattern Reconit., CVPR, 2022, pp. 18082–18091, http://dx.doi.org/10.1109/CVPR52688.2022.01755.

[30] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, Ziwei Liu, Learning to prompt for vision-language models, Int. J. Comput. Vis. 130 (9) (2022) 2337–2348, http://dx.doi.org/10.1007/s11263-022-01653-1.

[31] Shuwei Shao, et al., Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue, Med. Image Anal. (ISSN: 1361-8415) 77 (2022) 102338, http://dx.doi.org/10.1016/j.media.2021.102338.

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612, http://dx.doi.org/10.1109/TIP.2003.819861.

[33] Hang Zhao, Orazio Gallo, Iuri Frosio, Jan Kautz, Loss functions for image restoration with neural networks, IEEE Trans. Comput. Imaging 3 (1) (2016) 47–57.

[34] Brian Curless, Marc Levoy, A volumetric method for building complex models from range images, in: Annu. Conf. Comput. Graph. Interactive Techn., SIGGRAPH, 1996, pp. 303–312, http://dx.doi.org/10.1145/237170.237269.

[35] Qian-Yi Zhou, Jaesik Park, Vladlen Koltun, Open3D: A modern library for 3D data processing, 2018, http://dx.doi.org/10.48550/arXiv.1801.09847, arXiv:1801.09847.

[36] Max Allan, et al., Stereo correspondence and reconstruction of endoscopic data challenge, 2021, http://dx.doi.org/10.48550/arXiv.2101.01133, arXiv:2101.01133.

[37] P.J. Eddie Edwards, Dimitris Psychogyios, Stefanie Speidel, Lena Maier-Hein, Danail Stoyanov, SERV-CT: A disparity dataset from cone-beam CT for validation of endoscopic 3D reconstruction, Med. Image Anal. 76 (2022) 102302, http://dx.doi.org/10.1016/j.media.2021.102302.

[38] David Eigen, Christian Puhrsch, Rob Fergus, Depth map prediction from a single image using a multi-scale deep network, in: Proc. Inter. Conf. Neural Inf. Process. Syst., NIPS, 2014, pp. 2366–2374, http://dx.doi.org/10.48550/arXiv.1406.2283.

[39] Ilya Loshchilov, Frank Hutter, Decoupled weight decay regularization, Int. Conf. Learn. Represent. (2018) arXiv.1711.05101.