This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

MonoLoT: Self-Supervised Monocular Depth Estimation in Low-Texture Scenes for Automatic Robotic Endoscopy

Qi He, Guang Feng, Sophia Bano, Danail Stoyanov, Siyang Zuo

***Abstract*— The self-supervised monocular depth estimation framework is well-suited for medical images that lack ground-truth depth, such as those from digestive endoscopes, facilitating navigation and 3D reconstruction in the gastrointestinal tract. However, this framework faces several limitations, including poor performance in low-texture environments, limited generalisation to real-world datasets, and unclear applicability in downstream tasks like visual servoing. To tackle these challenges, we propose MonoLoT, a self-supervised monocular depth estimation framework featuring two key innovations: point matching loss and batch image shuffle. Extensive ablation studies on two publicly available datasets, namely C3VD and SimCol, have shown that methods enabled by MonoLoT achieve substantial improvements, with accuracies of 0.944 on C3VD and 0.959 on SimCol, surpassing both depth-supervised and self-supervised baselines on C3VD. Qualitative evaluations on real-world endoscopic data underscore the generalisation capabilities of our methods, outperforming both depth-supervised and self-supervised baselines. To demonstrate the feasibility of using monocular depth estimation for visual servoing, we have successfully integrated our method into a proof-of-concept robotic platform, enabling real-time automatic intervention and control in digestive endoscopy. In summary, our method represents a significant advancement in monocular depth estimation for digestive endoscopy, overcoming key challenges and opening promising avenues for medical applications.**

***Index Terms*— Monocular depth estimation, automatic intervention, digestive endoscopy.**

I. Introduction

ENDOSCOPY has long served as the golden standard investigative procedure for clinical diagnosis, providing high-definition internal tissue images with a narrow scope. During the diagnostic process, clinicians record anatomical landmarks following specific guidelines, providing sources for developing an image recognition system [1]. On the other hand, computer graphics techniques, including 3D reconstruction and monocular depth estimation, provide an alternative approach to locating these landmarks [2], [3]. Furthermore, robot platforms based on traditional endoscopes or capsule endoscopes have shown more comfortable and less invasive alternative approaches [4], [5].

Depth estimation is crucial for downstream clinical diagnosis tasks, such as navigation [6], [7] and active stabilisation [8], [9], and therefore, it is the primary focus of this study. The monocular depth estimation method calculates the spatial distance between the viewpoint and corresponding 3D points using a single camera and associated algorithms. The monocular camera model is applicable to digestive endoscopy. However, previous methods face several challenges, including underperformance in low-texture scenes, limited adaptability to real-world datasets, and unclear applicability of depth estimation in visual servoing tasks.

Notably, this task lacks ground-truth (GT) depth in real-world scenarios due to the absence of depth sensors, such as depth cameras, in digestive endoscopy, making self-supervised training prevalent in this field. However, the low texture present in endoscopic images poses a significant challenge for self-supervised monocular depth estimation methods [10]–[12], as such scene result in inaccurate photometric reconstruction errors, thereby impeding the training process. Recent advances have explored the utilisation of Generative Adversarial Networks (GANs) to acquire depth benchmarks from alternative image domains like chromoendoscopy [2], and simulators [13]. But these methods demonstrate limited generalisability to real endoscopy images and rely on external chromoendoscopy images or synthetic data during training. Moreover, achieving real-time performance with minimal latency is crucial for visual servoing applications, making a lightweight model during inference more desirable than a cost volume-based model [14].

Recent efforts to address low-texture challenges, including the Transformer-based local feature matching method LoFTR [15], have shown potential in generating high-quality matches. We integrated this feature-matching method into our training process to enhance the self-supervised monocular depth estimation frameworks for low-texture scenes. Furthermore, the availability of publicly accessible endoscopic datasets, such as C3VD [16], SimCol [17], EndoSLAM [3], Endomapper [18], and upper gastrointestinal (GI) dataset [19], has expanded

This work was supported by the National Natural Science Foundation of China under Grant No. 62133010, Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), UK (203145/Z/16/Z), and the Engineering and Physical Sciences Research Council (EPSRC) (EP/P027938/1). *(Qi He and Guang Feng contribute equally to the article. Corresponding author: Siyang Zuo.)*

Qi He, Guang Feng, and Siyang Zuo are with the Key Laboratory of Mechanism Theory and Equipment Design of Ministry of Education, Tianjin University, Tianjin 300354, China (e-mail: howard@tju.edu.cn; fengguang@tju.edu.cn; siyang_zuo@tju.edu.cn).

Sophia Bano, Danail Stoyanov are with the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) and the Department of Computer Science, University College London, London W1W 7TY, UK. (e-mail: sophia.bano@ucl.ac.uk; danail.stoyanov@ucl.ac.uk).

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

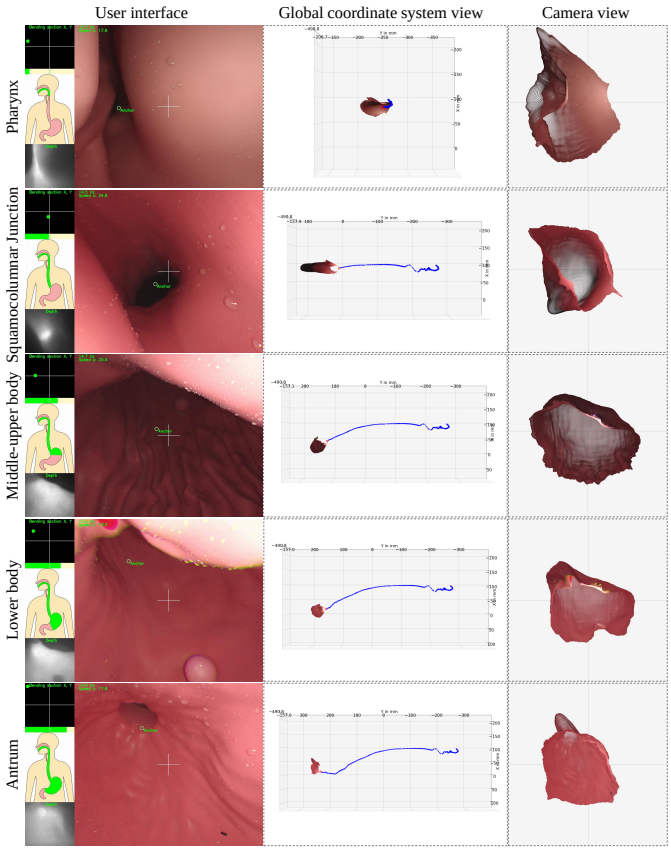


Fig. 1. Partial experimental frames of automatic robotic endoscopy on a realistic upper GI tract phantom. The camera poses and trajectories obtained from the Aurora tracker are depicted in red (current frame) and blue (previous frames) pyramid. Full details is provided in **Video S1**.

significantly. These datasets, encompassing phantom data, synthetic data, and real-world data, serve as valuable resources for training and validating models.

In this paper, we present a novel approach to address the challenges posed by inaccurate photometric loss during the training of self-supervised monocular depth estimation methods in low-texture scenes. We propose the point-matching loss, refining reprojected points of the monocular depth model through the feature correspondences derived from a feature-matching model (such as LoFTR). The efficiency of this step is improved by the adaptive weight and offline preprocessing. Additionally, the batch image shuffle loss is proposed to discriminate similar patches with different depths. Extensive experiments demonstrate the effectiveness of our proposed methods in overcoming these challenges. Furthermore, we present a proof-of-concept robotic platform to illustrate the feasibility of using monocular depth estimation models for automatic robotic endoscopy (see Fig. 1).

Our paper contributes significantly to the field of self-supervised monocular depth estimation in GI endoscope videos, with the key highlights as follows:

(i) We introduce a novel objective function that incorporates the point matching loss and batch image shuffle loss, effectively enhancing the performance of the self-supervised Monocular depth estimation in Low-Texture scenes (MonoLoT). Our methods resulted in improved accuracy ($\theta < 1.25$) from 4.7% to 18.8% on the C3VD dataset and from 4.4% to 5.7% on the SimCol dataset.

(ii) We present comprehensive quantitative and qualitative results on the C3VD dataset, SimCol dataset, upper-GI dataset, EndoSLAM dataset, and EndoMapper dataset, demonstrating the effectiveness and robustness of our proposed methods.

(iii) We have successfully integrated monocular depth estimation into a robotic platform designed for visual servoing. Our approach is thoroughly validated on a realistic GI phantom, confirming its robustness and practical applicability in real-world scenarios.

By combining these contributions, our work significantly advances the field of self-supervised monocular depth estimation, offering valuable insights for both researchers and practitioners in the field of visual servoing.

II. Related Work

A. Depth Estimation for Monocular Endoscope

Depth estimation of monocular endoscopic images has been an important research topic in recent years. Various approaches have been proposed to tackle this task. For example, Rau et al. [13] introduced a method based on GANs to estimate depth maps from monocular endoscopic images. Their method transfers the network's monocular depth estimation capabilities from the synthetic image domain to real endoscopic images. In another notable work by Widya et al. [20], the focus was on addressing the issue of balancing weights between the depth and pose loss terms by employing unsupervised photometric error as a general term of the supervised depth and pose error. Rodriguez-Puigvert et al. [21] presented a Bayesian deep network approach to explicitly model data uncertainty and model uncertainty for single-view depth estimation in colonoscopies. Their study involved ensembling multiple deep learning-based depth estimation models as a Bayesian teacher model, enabling the student model to learn from a similar distribution. In summary, the existing literature lacks comprehensive research on addressing the challenges posed by low-texture regions in endoscopic images, which is the primary focus of this paper.

B. Feature Extraction in Low-Texture Regions

Low-texture regions pose a common challenge in both indoor scenes and GI scenarios. To address this issue, several studies have investigated specific prior or global attention schemes. For instance, Li et al. [22] introduced a co-planar constraint specifically designed for low-texture regions in indoor images. Their method utilised geometric and colour dissimilarities to constrain the estimated depth, enabling reliable depth estimation in challenging regions lacking rich textures. Like Saunders et al. [23], our method also bypass the homography warping to input augmented images. However, our starting point is different because their baseline model was able to achieve highly accurate depth maps on augmented images. Therefore, they utilised the correspondence between unaugmented and augmented images, using the unaugmented depth to supervise the augmented depth, thereby improving the accuracy on augmented images. In contrast, our baseline

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

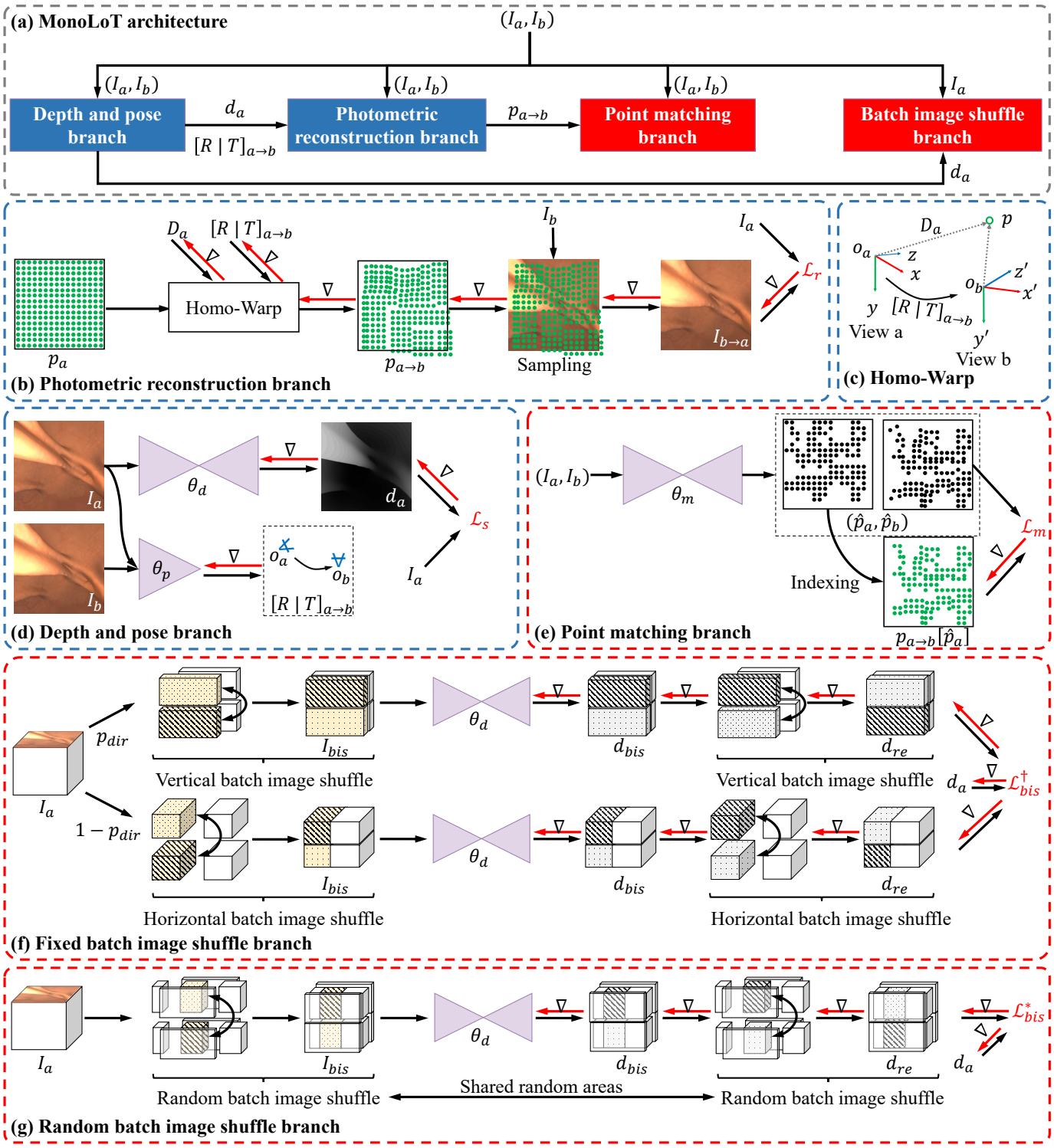


Fig. 2. Monocular depth estimation pipeline. The overall architecture is depicted in (a). The fundamental monocular depth estimation pipeline comprises (b) the photometric reconstruction branch, (c) homography warping, and (d) the depth and pose branch, highlighted within blue dash rectangles. The proposed methods, such as (e) the point matching branch, (f) the fixed batch image shuffle branch, and (g) the random batch image shuffle branch, are emphasised within red dash rectangles. Gradient back-propagation is depicted by the red arrows.

model performs poorly on unaugmented images. Zhao et al. [24] proposes the use of COLMAP to estimate preliminary poses, followed by neural network optimisation of the pose, specifically targeting the characteristic of large rotational movements between adjacent frames in indoor videos. However, GI videos do not exhibit such large rotations as seen in indoor scenes. Moreover, due to their weak texture characteristics, the performance of pose estimation by COLMAP tends to be poor for GI videos [17]. Unlike [25], which utilises the depth from the previous epoch model for self-

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

distillation of the current model, our method constructs a new regularisation term and does not block the gradient from back-propagating along the pseudo-label depth. Furthermore, Sun et al. [15] proposed a Transformer-based approach for generating dense matches in low-texture areas. By extracting coarse-to-fine interest points in global receptive field, their method improved the feature-matching performance in low-texture regions. Since the digestive tract is mostly composed of curved surfaces, the coplanar constraint is not suitable for the digestive tract scene in our study. Therefore, we instead focus on extracting feature points in weak textures to improve self-supervised training.

C. Transformer in Depth Estimation

In recent studies, Transformer-based models [26] have been explored for monocular endoscopy depth estimation, showcasing their potential in this domain [11], [12]. For instance, MonoViT [11] replaced the convolutional neural network (CNN)-based depth encoder of Monodepth2 [10] with MPViT [27], while Lite-Mono [12] employed consecutive dilated convolutions [28] and cross-covariance attention [29] to strengthen their depth encoder. In our study, we validate the limitations of these models on digestive endoscopy image data and construct suitable general modules to improve their performance.

III. Monocular Depth Estimation in Low-texture Endoscopic Videos

We observed that the poor performance of self-supervised monocular depth estimation baseline is primarily due to inaccurate photometric reconstruction errors. To address this issue, the point matching loss ($\mathcal{L}_m$) and batch image shuffle loss ($\mathcal{L}^{\dagger}_{bis}$ for fixed and $\mathcal{L}^{*}_{bis}$ for random) are proposed and introduced in the following sections.

A. Framework Overview

We present a summary of the state-of-the-art self-supervised monocular depth estimation pipeline [10]–[12], as seen in Fig. 2(b)(c)(d). Initially, PoseNet θ_p and DepthNet θ_d predict the viewpoint $[R \mid T]_{a\to b}$ and inverse depth d_a from the image pair (I_a, I_b) using the depth and pose branches (see Fig. 2(d)). Subsequently, the synthesised frame $I_{b\to a}$ is generated by the photometric reconstruction branch (see Fig. 2(b)). Their objective function consists of two key components:

$$\mathcal{L} = \mu\mathcal{L}_r + \lambda\mathcal{L}_s, \tag{1}$$

(i) $\mathcal{L}_r$ **(Minimum Reprojection Loss)**: this term aims to minimise the photometric reconstruction error $\mathrm{pe}(\cdot)$ and is defined as:

$$\mathcal{L}_r = \min_b \mathrm{pe}(I_a, I_{b\to a}). \tag{2}$$

(ii) $\mathcal{L}_s$ **(Edge-Aware Smoothness Loss)**: this term is designed to smooth the depth maps and is expressed as:

$$\mathcal{L}_s = |\partial_x d^*_a| e^{-|\partial_x I_a|} + |\partial_y d^*_a| e^{-|\partial_y I_a|}. \tag{3}$$

Here, $d^*_a = d_a/\overline{d_a}$ denotes the mean-normalised inverse depth [10].

Additionally, μ is used for "Auto-Masking" [10], calculated as:

$$\mu = \min_b \mathrm{pe}(I_a, I_{b\to a}) < \min_b \mathrm{pe}(I_a, I_b), \tag{4}$$

where λ plays a role in determining the relative importance of $\mathcal{L}_r$ and $\mathcal{L}_s$ in the overall loss.

In our proposed objective function, which extends the state-of-the-art pipeline in (1), we introduce two new components:

$$\mathcal{L} = \mu\mathcal{L}_r + \lambda\mathcal{L}_s + \mathcal{L}_m + \mathcal{L}_{bis}. \tag{5}$$

(i) $\mathcal{L}_m$ **(Point Matching Loss)**: this loss is formulated to refine the reprojected points by incorporating pseudo label supervision during training.

(ii) $\mathcal{L}_{bis}$ **(Batch Image Shuffle Loss)**: this loss is designed to enhance the model's ability to generalise to unseen data in low-texture scenes by training on synthetic images with similar textures.

The following paragraphs will provide a detailed explanation of the Point Matching Loss and the Batch Image Shuffle Loss, starting from the fundamentals of the minimum reprojection loss, also known as photometric reconstruction error.

B. Photometric Reconstruction Error

Given current image frame I_a, the image frame I_b of viewpoint $[R \mid T]_{a\to b}$, current monocular depth D_a, and the camera intrinsic matrix K, the photometric reconstruction error loss $\mathrm{pe}(\cdot)$ gathers the differences between I_a and synthesised frame $I_{b\to a}$ from $[R \mid T]_{a\to b}$:

$$\mathrm{pe}(I_a, I_{b\to a}) = (1-\alpha)\|I_a, I_{b\to a}\|_1 + \frac{\alpha}{2}(1-\mathrm{SSIM}(I_a, I_{b\to a})), \tag{6}$$

where $\mathrm{SSIM}(\cdot)$ is the structure similarity loss [30] and $\alpha = 0.85$ following the setting from [10]. Here $I_{b\to a}$ denotes the value obtained by sampling I_b using reprojected points $p_{a\to b}$ on the pixel coordinates system of I_b:

$$I_{b\to a} = I_b\langle p_{a\to b}\rangle, \tag{7}$$

$$\text{and} \quad p_{a\to b} = K \cdot (R_{a\to b} \cdot (K^{-1} \cdot p_a \cdot D_a) + T_{a\to b}). \tag{8}$$

Here $\langle\cdot\rangle$ denotes a sampling operator and p_a signifies the target location in the pixel coordinate system of I_a.

To summarise, the loss term $\mathcal{L}_r$ is back-propagated to update the parameters of $R_{a\to b}$, $T_{a\to b}$, and D_a (as depicted in Fig. 2(b)). These parameters are obtained through PoseNet θ_p and DepthNet θ_d (as illustrated in Fig. 2(d)), calculated as:

$$[R\,|\,T]_{a\to b} = \theta_p(I_a, I_b), \tag{9}$$

$$D_a = \frac{1}{(\frac{1}{\max_D} + (\frac{1}{\min_D} - \frac{1}{\max_D})d_a)}, \tag{10}$$

where, inverse depth d_a in the range of $[0, 1]$ is transformed into depth within the range of $[min_D, max_D]$. In this case, the parameters θ_p and θ_d are primarily obtained through self-supervised learning by minimising the photometric reconstruction error in (6).

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

However, there is a special case in which when the location of the reprojected point $p_{a\to b}$ lies in a low-texture area, the sampling operation in (7) will sample from pixels with very similar colours. This ultimately results in multiple minima for the term $\|I_a, I_{b\to a}\|_1$ of the photometric reconstruction errors in (6), thereby weakening its supervisory effect.

C. Point Matching Loss

In response to this challenge, we have introduced the *Point Matching Loss* (PM) to improve the accuracy of reprojected points in low-texture regions.

1) Core mechanism: In addition to minimising the photometric reconstruction error, we can optimise θ_d and θ_p by reducing the distance between the reprojected points $p_{a\to b}$ and their corresponding feature points $\hat{p}_b$ on I_b, as determined by the local feature matching model (as shown in Fig. 2(e)). To elaborate, when given the image pair (I_a, I_b), the local feature matching model θ_m identifies corresponding feature points $(\hat{p}_a, \hat{p}_b)$ as follows:

$$(\hat{p}_a, \hat{p}_b) = \theta_m(I_a, I_b), \tag{11}$$

where θ_m was pretrained and remained frozen during our training process. In practice, we chose LoFTR [15] as our θ_m due to its outstanding feature point extraction ability on challenging scenarios such as texture-less wall and floor with repetitive patterns, which are commonly encountered in endoscopic videos. The point-matching process is illustrated in Fig. 3, highlighting the matching ability of LoFTR.

Subsequently, we substituted p_a in (8) with $\hat{p}_a$ and minimised the *Point Matching Loss* $\mathcal{L}_m$ between the reprojected feature point $\hat{p}_{a\to b}$ from I_a to I_b and the corresponding feature point $\hat{p}_b$ on I_b:

$$\mathcal{L}_m = \|\hat{p}_b, \hat{p}_{a\to b}\|_1. \tag{12}$$

Here, $\hat{p}_{a\to b}$ is computed as:

$$\hat{p}_{a\to b} = K \cdot (R_{a\to b} \cdot (K^{-1} \cdot \hat{p}_a \cdot D_a) + T_{a\to b}). \tag{13}$$

It's worth noting that $\hat{p}_b$ consists of subpixel (or floating-point coordinates), while $\hat{p}_a$ uses integer coordinates, which can also serve as indices into tensors, denoted as $[\hat{p}_a]$. Consequently, instead of passing $\hat{p}_a$ to the homography warping module, we directly employed $\hat{p}_a$ as the index for p_a and $p_{a\to b}$ (see Fig. 2(e)). Therefore, equation (13) can be simplified as:

$$\hat{p}_{a\to b} = p_{a\to b}[\hat{p}_a], \tag{14}$$

and $$p_{a\to b}[\hat{p}_a] = K \cdot (R_{a\to b} \cdot (K^{-1} \cdot p_a[\hat{p}_a] \cdot D_a) + T_{a\to b}). \tag{15}$$

2) Adaptive Point Matching Loss: While the point-matching loss is expected to decrease with the optimisation of the objective function, the inference error of the local feature-matching model can impede this process, especially when the photometric reconstruction error decreases below the point-matching loss. Specifically, the corresponding feature points $\hat{p}_b$ acquired through the local feature matching model θ_m may not be optimal, and we posit that the error between these points and the reprojected points $\hat{p}_{a\to b}$ remains within the predefined threshold δ. We define the *Adaptive Point Matching Loss* as:

$$\mathcal{L}_m = \begin{cases} \|\hat{p}_b, \hat{p}_{a\to b}\|_1 \cdot eps & \text{if } \|\hat{p}_b, \hat{p}_{a\to b}\|_1 < \delta \\ \|\hat{p}_b, \hat{p}_{a\to b}\|_1 & \text{otherwise} \end{cases} \tag{16}$$

where eps is set to 10^{-4}.

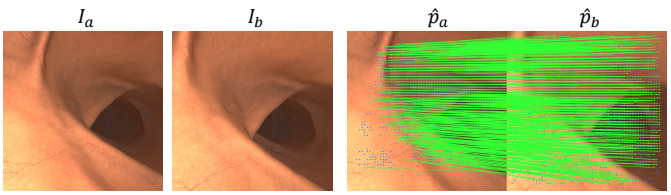


Fig. 3. Visualisation of LoFTR-based Point Matching. Compared with the general scene, the corresponding feature points ($\hat{p}_a$ and $\hat{p}_b$) are harder to detect (highlighted with blue dots) and pair (highlighted with green lines) in GI endoscopic images (I_a and I_b) due to the low-texture feature. Fortunately, this method can handle these task robustly.

3) Improving Computational Efficiency via Offline Preprocessing: In practice, we found that our point-matching loss module significantly increased training time, tripling it, leading to an approximately 9 GB increase in GPU memory usage. This was primarily due to the computationally intensive and GPU memory-demanding inference of (11). To address this issue, we now calculate the corresponding feature points for each image pair as a preprocessing step before training. In this revised approach, the point matching loss not only no longer extends the training time but also requires a minimal additional amount of GPU memory to execute.

D. Batch Image Shuffle Loss

Another challenge in self-supervised training is data augmentation, especially in low-texture endoscopic scenes, where pixel blocks may have uniform textures but represent varying depths. Specifically, if image frames I_a and I_b are related by the homography warping, considering the homography warping of augmented content is necessary when using augmented images to compute the photometric reconstruction error loss (Fig. 2(b)). However, bypassing the photometric reconstruction error allows us to perform image augmentation without considering the homography warping. We propose the *Batch Image Shuffle Loss*, which optimises the objective function through self-supervised learning by directly comparing the inverse depths before and after image augmentation. We regard the characteristic of low texture as having similar textures but different depths, and we need the model make correct predictions for such image areas. Through batch image shuffle, we can construct such regions at the junction of stitched images, and encourage the model to make correct predictions through self-supervision. As shown in Fig. 4, MonoLoT makes more accurate predictions in these kinds of junction areas compared to the baseline.

Input batch, denoted as $I_a \in \mathbb{R}^{F\times 3\times H\times W}$, is divided into upper and lower parts along the first dimension (F). Swap the top halves of the two half-batches along the third dimension

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

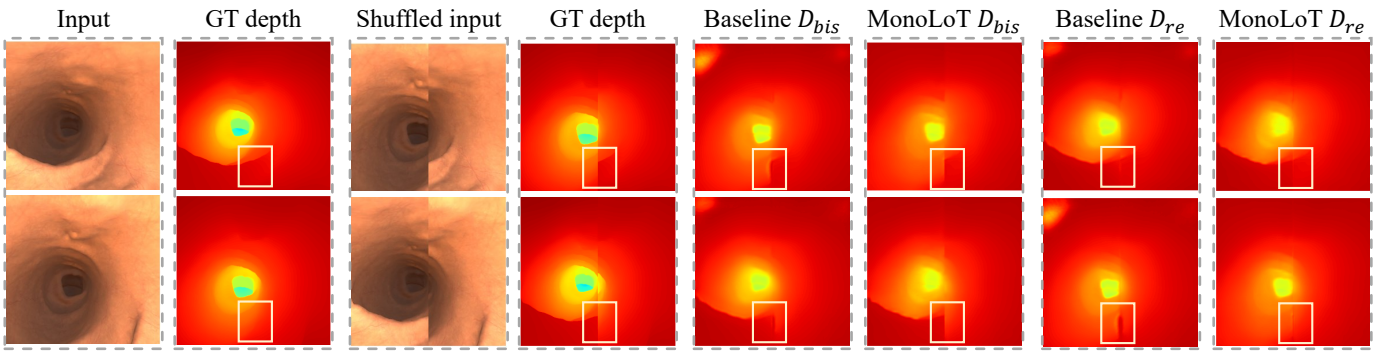


Fig. 4. Visualisation of Batch Image Shuffle. The first column illustrates the input (where batch size is 2 for exemplification). The second column shows the corresponding GT depths. Taking the horizontal batch image shuffle as an example, the third column shows the shuffled images ($\boldsymbol{I_{bis}}$) after the first shuffle, followed by the corresponding GT depths in the fourth column. The fifth and sixth columns show the shuffled depth generated by baseline and MonoLoT respectively. Predicted depths in the seventh and eighth columns shows MonoLoT outperforms the baseline by producing sharper, more detailed depth maps in junction areas (highlighted with yellow boxes).

(H) and merge them to form the shuffled input batch:

$$\text{shuffle}_{\text{ver}}(I_a) = \begin{pmatrix} I_a[\frac{F}{2}:, :, :\frac{H}{2}, :] & I_a[:\frac{F}{2}, :, \frac{H}{2}:, :] \\ I_a[:\frac{F}{2}, :, :\frac{H}{2}, :] & I_a[\frac{F}{2}:, :, \frac{H}{2}:, :] \end{pmatrix}. \quad (17)$$

Likewise, horizontal batch image shuffling, as represented by:

$$\text{shuffle}_{\text{hor}}(I_a) = \begin{pmatrix} I_a[\frac{F}{2}:, :, :, :\frac{W}{2}] & I_a[:\frac{F}{2}, :, :, \frac{W}{2}:] \\ I_a[:\frac{F}{2}, :, :, :\frac{W}{2}] & I_a[\frac{F}{2}:, :, :, \frac{W}{2}:] \end{pmatrix}, \quad (18)$$

allows for batch image shuffling along the fourth dimension (W). Introduce a probability term, allowing the model to use vertical batch image shuffling with a probability of p_{dir} (see Fig. 2(f)). Once the shuffled input batch is constructed, it is fed to the DepthNet, which estimates the shuffled inverse depth. This shuffled inverse depth is then rearranged to generate the reordered inverse depth d_{re}, denoted as:

$$d_{re} = \begin{cases} \text{shuffle}_{\text{ver}}(\theta_d(\text{shuffle}_{\text{ver}}(I))) & \text{w.p. } p_{dir} \\ \text{shuffle}_{\text{hor}}(\theta_d(\text{shuffle}_{\text{hor}}(I))) & \text{w.p. } 1 - p_{dir} \end{cases} \quad (19)$$

Minimise the batch image shuffle loss between the inverse depth of the original batch and the reordered batch:

$$\mathcal{L}_{bis} = \|d, d_{re}\|_1, \quad (20)$$

In addition to this fixed batch image shuffling approach, we have explored further randomisation of the swapped areas. As depicted in Fig. 2(g), random patches of the two half-batches are swapped before and after depth estimation. The randomly generated areas share the same location and size, ensuring consistent illumination. It's worth noting that we ensure the randomly generated coordinates are divisible by 8 at scale 0 to ensure precise alignment with the coordinates on scales 1, 2, and 3. We formally denote the Fixed Batch Image Shuffle loss (FBIS) as $\mathcal{L}_{bis}^{\dagger}$, and the Random Batch Image Shuffle loss (RBIS) as $\mathcal{L}_{bis}^{*}$.

IV. Robotic Gastrointestinal Intervention: Structure and Depth-Based Control

To demonstrate the robustness of the proposed MonoLoT on a realistic robotic endoscopy setup, we have implemented real-time visual servoing on a real GI endoscopy system using our proof-of-concept robotic platform. This achievement demonstrates the applicability of our proposed monocular depth estimation method and the potential for utilising depth-based control in automatic intervention.

A. Robotic platform

The proposed automatic intervention method is implemented and validated using a novel robotic platform currently under development at Tianjin University laboratory. For completeness, we have included necessary details for the robotic platform sufficient to demonstrate the robustness of MonoLoT in practical application (see Fig. 5). It consists of two primary components:

6-DOF Serial Manipulator: This manipulator is responsible for adjusting the initial pose of the endoscopy. It provides six joints for precise control. The sixth joint (motor 4) of the manipulator allows rotation along the z-axis of the camera coordinates system.

End-effector: The end-effector is affixed to the sixth joint of the manipulator. Motor 1 and 2 control the rotation of the small knobs and large knobs separately, enabling the bending of the endoscope along the x-axis and y-axis of the camera coordinate system. Motor 3 is responsible for the insertion and retraction of the endoscope.

To simulate gastroscopy, we integrated the robotic platform with a GI endoscopy system (CV-290, Olympus, Japan) and utilised an upper GI tract phantom made of soft rubber material (see Fig. 5(b)). This phantom closely simulates the actual condition within the human GI tract, providing a realistic environment for testing and experimentation. To ensure precise pose collection for evaluation purposes, a real-time electromagnetic sensor (Aurora, NDI, Canada) was incorporated into the working channel of the endoscope.

B. Depth-Based Control

The control framework for automatic intervention consists of three parts (see Fig. 5(c)). The endoscope captures videos of the upper GI tract, transmitting them to image processing modules via a video capture card. Captured images are processed to convert them into depth maps using MonoLoT-enabled model. Within the depth map, the deepest point is selected as the

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

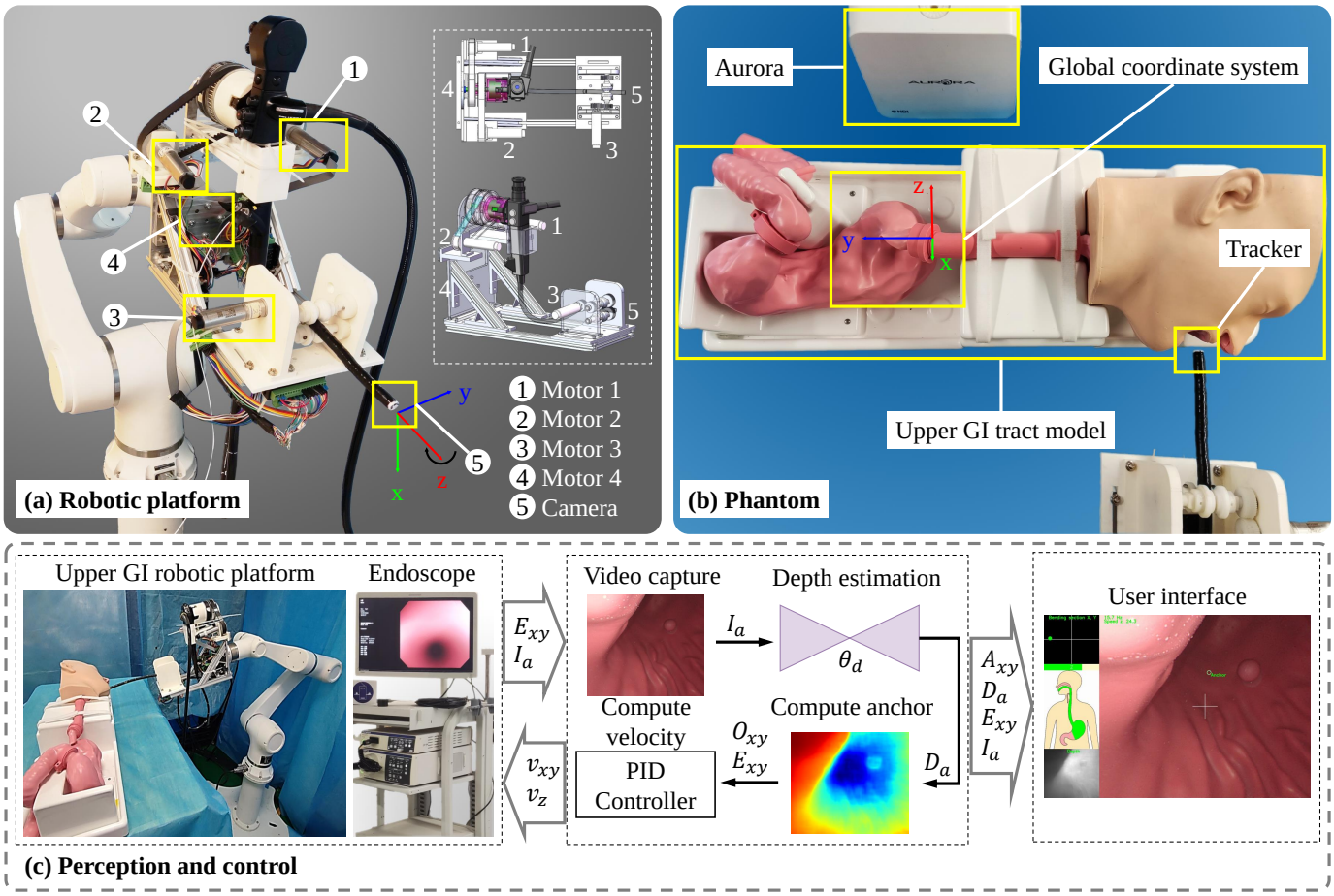


Fig. 5. Overview of Robotic Platform. (a) The robotic platform comprises four motors that manipulate the camera's pose. (b) The phantom is equipped with the Aurora sensor for pose tracking and recording. (c) Perception and control integrate our depth estimation into a PID controller, enabling real-time visual servoing. The global coordinate system is defined to maintain consistency with the Aurora system: +x points inward, +y points to the left, +z points upward.

TABLE I
C3VD SPLIT. THE INDEX OF EACH VIDEO IS FORMATTED AS "<TEXTURE><VIDEO>" [16].

	Cecum								Descending	Sigmoid				Transcending							
Index	1a	1b	2a	2b	2c	3a	4a	4b	4a	1a	2a	3a	3b	1a	1b	2a	2b	2c	3a	3b	4a	4b
Train	276	765	370	1,142	595	730				700	514			61	700	194	103	235	250	214		
Val							465					613									382	
Test								425	148				536									597

TABLE II
SIMCOL SPLIT. THE INDEX OF EACH VIDEO IS DERIVED FROM THE SIMCOL CHALLENGE [17].

	SyntheticColon_I									SyntheticColon_II									SyntheticColon_III
Index	S1-S3	S4	S5	S6-S8	S9	S10	S11-S13	S14	S15	B1-B3	B4	B5	B6-B8	B9	B10	B11-B13	B14	B15	O1-O3
Train	3,603			3,603			3,603			3,603			3,603			3,603			
Val		1,201			1,201			1,201			1,201			1,201			1,201		
Test			1,201			1,201			1,201			1,201			1,201			1,201	1803

anchor point, denoted as $A_{xy} = \arg\max_{xy}(D_a)$. The target direction $\overrightarrow{O}$ is calculated as:

$$\overrightarrow{O} = \frac{A_{xy} - C_{xy}}{C_{xy}}, \tag{21}$$

where $C_{xy} = \begin{pmatrix} W/2 & H/2 \end{pmatrix}$ represents the image centre. The anchor point coordinates, combined with the current state of motor 1 and 2, serve as target angles O_{xy} for the robotic platform's controller, calculated as:

$$O_{xy} = E_{xy} + s \cdot \overrightarrow{O}, \tag{22}$$

where, $E_{xy} = \begin{pmatrix} \text{Motor}_1 & \text{Motor}_2 \end{pmatrix}$ denotes the current angles of motor 1 and 2 from the motor encoder, while s signifies the scale of each move. Utilising the received O_{xy} from the image processing modules, the robotic platform adjusts the rotation of the large knob and the small knob to minimise $\overrightarrow{O}$. This adjustment is performed using the proportional-integral-

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

derivative (PID) controller:

$$v_{xy} = \mathrm{PID}(O_{xy}, E_{xy}, \mathrm{d}t), \tag{23}$$

where v_{xy} is the manipulated velocity of motor 1 and 2, $\mathrm{d}t$ is the time interval of our PID control. It's worth noting that our PID controller operates at a different frequency from the image processing modules, necessitating separate readings of the motor state. The speed control of insertion and retraction is determined as:

$$v_z = \frac{max_{v_z}}{min_{v_z}}^{-\frac{\|\vec{O}\|_2}{\sqrt{2}}} \cdot max_{v_z}. \tag{24}$$

Here, v_z denotes the velocity of motor 3, which falls within the predefined range of $[min_{v_z}, max_{v_z}]$. In implementation, position limits are set for motors 1, 2, and 3 to ensure safety throughout the entire intervention. Additionally, a moving average is applied on the anchor point to ensure a smooth transition between targets.

In summary, this control framework relies on real-time monocular depth estimation, enabling the robotic platform to accurately adjust the endoscope's position and reach the desired target within the upper GI tract.

V. Experiments

In this section, we will provide a detailed evaluation of the proposed objective function and the robotic platform.

A. Experimental Protocols

1) Datasets: The C3VD [16] and SimCol dataset [17] were used for training and evaluation. The upper-GI dataset [19], EndoSLAM dataset [3], and EndoMapper dataset [18] were utilised for generalisability evaluation, while the realistic upper GI tract phantom served as a validation platform for automatic intervention.

(i) The **C3VD dataset [16]** is a 3D video colonoscopy dataset acquired with a high-definition clinical colonoscope and high-fidelity colon phantoms. It primarily comprises 20 videos, each of which records segments of the lower GI tract, including categories such as the cecum, descending colon, sigmoid colon and transcending colon.

(ii) The **SimCol dataset [17]** are sourced from virtual environments built with Unity Engine. SimCol contains 33 videos, each recording the entire lower GI tract.

(iii) The **upper-GI dataset [19]** consists of 3704 images from 11 landmarks, such as pharynx, middle-upper body, antrum (see Fig. 9).

(iv) The **EndoSLAM dataset [3]** encompasses a wide range of synthetic and real endoscopic video data, capturing the diverse and challenging environments encountered within various parts of colons.

(v) The **EndoMapper dataset [18]** comprises 59 sequences totalling over 15 hours of video, providing benchmark for the evaluation of localisation and navigation of the endoscope.

(vi) A **realistic upper GI tract phantom** made of soft rubber material (see Fig. 5(b)) was used for qualitative evaluation of automatic intervention.

The data splits for our experiments, including training, validation, and test sets, are organised as follows:

(i) The **C3VD split** was arranged as shown in Table I, where the textures for testing would not be seen during training. Due to the slow camera movement, we set the triplet index to (-20, 0, 20). To further remove static frames, we downsample triplets to 320×256, perform feature matching [15] for consecutive frames in triplet, and remove those triplets whose average distance between feature points is less than 1 pixel. We finally got our C3VD split which has a total of 4,524 monocular triplets for training, 1,068 for validation and 1,261 for testing.

(ii) The **SimCol split** was organised as depicted in Table II, where videos from various sets may share the same texture. Since the camera speed is fast, we set the triplet index to (-3, 0, 3). We finally got a total of 21,510 monocular triplets for training, 7,161 for validation and 8,951 for testing.

2) Implementation Details: The training and inference of our models are supported by two NVIDIA RTX 3090 GPUs and four Titan Xp GPUs. We summarise our hyperparameters in Table III following the setting from Monodepth2 [10], Lite-mono [12], and MonoViT [11]. Unless otherwise specified, all models on the SimCol dataset have an input/output resolution of 352×352, and all models on the C3VD dataset have an input/output resolution of 320×256. The objective function of GT depth-supervised baselines is defined as:

$$\mathcal{L} = (1-\alpha)\|D_a - D_{gt}\|_1 + \frac{\alpha}{2}(1 - \mathrm{SSIM}(D_a, D_{gt})), \tag{25}$$

where D_{gt} denotes the GT depth. Lite-mono with FBIS and PM trained on C3VD is used to generate depth maps for automatic robotic endoscopy.

TABLE III
Detail hyperparameters.

Method	Epoch	Batch size	Optimiser	Scheduler	Initial LR	Last LR
Monodepth2 [10]	25	12	Adam	StepLR(20, 0.1)	10^{-4}	10^{-5}
Lite-mono [12]	25	12	AdamW	CosineAnnealing(·)	10^{-4}	10^{-5}
MonoViT [11]	25	10	AdamW	ExponetialLR(0.9)	5×10^{-5}	-

3) Evaluation Metrics: The performance is evaluated by seven metrics proposed in prior works [31], including mean absolute relative error (Abs Rel), squared relative error (Sq Rel), root mean squared error (RMSE), root mean squared log error (RMSE log), and the accuracies under threshold ($\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$). Following the evaluation protocol established in [10], we present results using per-image median GT depth scaling. Additionally, we perform five training runs for each model, calculating average and standard deviation values for every evaluation metric. Subsequently, qualitative evaluation is conducted on the model with median performance from these five runs (see Figs. 6 and 7). Unless otherwise noted, all qualitative results are colour-mapped based on their absolute depth.

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

TABLE IV
COMPARISON WITH BASELINE METHODS ON THE C3VD DATASET. "D": GROUND TRUTH DEPTH SUPERVISION. "M": SELF-SUPERVISED MONO SUPERVISION. THE BEST RESULTS ARE IN **BOLD**.

Method	Train	Depth Error(↓)				Depth Accuracy(↑)		
		Abs rel	Sq rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [10] (baseline)	D	0.119±0.001	0.006±0.000	0.039±0.001	0.141±0.001	0.877±0.003	0.976±0.001	0.998±0.000
Lite-mono [12] (baseline)	D	0.097±0.000	0.004±0.000	0.037±0.000	0.120±0.000	0.901±0.000	0.996±0.000	**1.000**±0.000
MonoViT [11] (baseline)	D	**0.078**±0.000	**0.003**±0.000	**0.033**±0.000	**0.096**±0.000	**0.941**±0.001	**0.999**±0.000	**1.000**±0.000
Monodepth2 [10] (baseline)	M	0.160±0.004	0.014±0.000	0.080±0.001	0.207±0.004	0.746±0.017	0.945±0.004	0.988±0.004
Lite-mono [12] (baseline)	M	0.102±0.009	0.007±0.001	0.062±0.006	0.139±0.013	0.897±0.021	0.984±0.006	0.997±0.002
MonoViT [11] (baseline)	M	0.114±0.008	0.009±0.001	0.070±0.007	0.156±0.015	0.864±0.029	0.978±0.008	0.996±0.003
Monodepth2 + $\mathcal{L}^{*}_{bis}$ + $\mathcal{L}_m$ (MonoLoT)	M	0.096±0.001	0.005±0.000	**0.043**±0.001	0.118±0.001	0.934±0.002	**0.997**±0.001	**1.000**±0.000
Lite-mono + $\mathcal{L}^{\dagger}_{bis}$ + $\mathcal{L}_m$ (MonoLoT)	M	**0.088**±0.001	**0.004**±0.000	0.044±0.001	**0.111**±0.001	**0.944**±0.001	**0.997**±0.001	**1.000**±0.000
MonoViT + $\mathcal{L}^{\dagger}_{bis}$ + $\mathcal{L}_m$ (MonoLoT)	M	0.090±0.002	0.005±0.000	0.046±0.002	0.114±0.002	0.942±0.003	**0.997**±0.001	**1.000**±0.000

TABLE V
COMPARISON WITH BASELINE METHODS ON THE SIMCOL DATASET. "D": GROUND TRUTH DEPTH SUPERVISION. "M": SELF-SUPERVISED MONO SUPERVISION. GROUND TRUTH DEPTH VALUES RESCALED TO [**0, 20**] CM. THE BEST RESULTS ARE IN **BOLD**.

Method	Train	Depth Error(↓)				Depth Accuracy(↑)		
		Abs rel	Sq rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [10] (baseline)	D	0.030±0.000	0.005±0.000	0.118±0.001	0.041±0.000	0.996±0.000	**0.999**±0.000	**1.000**±0.000
Lite-mono [12] (baseline)	D	0.055±0.000	0.041±0.000	0.194±0.000	0.064±0.000	0.987±0.000	0.997±0.000	0.999±0.000
MonoViT [11] (baseline)	D	**0.025**±0.000	**0.004**±0.000	**0.108**±0.002	**0.035**±0.000	**0.997**±0.000	**0.999**±0.000	**1.000**±0.000
Monodepth2 [10] (baseline)	M	0.160±0.022	1.008±0.306	0.724±0.084	0.169±0.016	0.907±0.008	0.953±0.004	0.970±0.004
Lite-mono [12] (baseline)	M	0.171±0.008	1.318±0.119	0.742±0.026	0.167±0.004	0.913±0.003	0.951±0.001	0.967±0.001
MonoViT [11] (baseline)	M	0.143±0.014	0.562±0.182	0.627±0.050	0.171±0.008	0.901±0.006	0.951±0.003	0.969±0.002
Monodepth2 + $\mathcal{L}^{*}_{bis}$ + $\mathcal{L}_m$ (MonoLoT)	M	0.074±0.002	0.066±0.000	**0.370**±0.006	**0.104**±0.002	**0.959**±0.002	**0.988**±0.000	0.995±0.001
Lite-mono + $\mathcal{L}^{*}_{bis}$ + $\mathcal{L}_m$ (MonoLoT)	M	0.075±0.000	0.069±0.000	0.394±0.003	0.107±0.001	0.957±0.001	0.987±0.000	0.994±0.000
MonoViT + $\mathcal{L}^{\dagger}_{bis}$ + $\mathcal{L}_m$ (MonoLoT)	M	**0.073**±0.002	**0.065**±0.001	0.394±0.008	0.105±0.002	0.958±0.001	**0.988**±0.000	**0.995**±0.000

TABLE VI
ABLATION STUDIES OF FBIS, RBIS AND PM ON THE C3VD DATASET. THE BEST RESULTS ARE IN **BOLD**. THE SECOND BEST ARE UNDERLINED.

DepthNet	$\mathcal{L}_{bis}$†	$\mathcal{L}_{bis}$*	$\mathcal{L}_m$	Depth Error(↓)				Depth Accuracy(↑)		
				Abs rel	Sq rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [10]		(baseline)		0.160±0.004	0.014±0.000	0.080±0.001	0.207±0.004	0.746±0.017	0.945±0.004	0.988±0.004
	✓			0.103±0.002	0.006±0.000	0.053±0.001	0.132±0.003	0.905±0.008	0.993±0.002	**1.000**±0.000
		✓		0.098±0.002	**0.005**±0.000	0.046±0.001	0.121±0.002	0.928±0.005	0.996±0.000	**1.000**±0.000
			✓	0.147±0.009	0.012±0.002	0.069±0.004	0.187±0.011	0.791±0.019	0.947±0.010	0.989±0.006
	✓		✓	0.098±0.003	0.006±0.000	0.051±0.001	0.124±0.002	0.920±0.004	0.995±0.001	**1.000**±0.000
		✓	✓	**0.096**±0.001	**0.005**±0.000	**0.043**±0.001	**0.118**±0.001	**0.934**±0.002	**0.997**±0.001	**1.000**±0.000
Lite-mono [12]		(baseline)		0.102±0.009	0.007±0.001	0.062±0.006	0.139±0.013	0.897±0.021	0.984±0.006	0.997±0.002
	✓			0.093±0.003	0.005±0.000	0.048±0.001	0.118±0.002	0.935±0.004	0.996±0.000	**1.000**±0.000
		✓		0.102±0.002	0.006±0.000	0.051±0.001	0.126±0.002	0.920±0.005	0.996±0.000	**1.000**±0.000
			✓	**0.086**±0.002	0.005±0.000	0.046±0.001	0.112±0.002	0.933±0.007	0.995±0.002	**1.000**±0.000
	✓		✓	0.088±0.001	**0.004**±0.000	**0.044**±0.001	**0.111**±0.001	**0.944**±0.001	0.997±0.001	**1.000**±0.000
		✓	✓	0.092±0.001	0.005±0.000	0.046±0.001	0.116±0.001	0.938±0.002	**0.997**±0.000	**1.000**±0.000
MonoViT [11]		(baseline)		0.114±0.008	0.009±0.001	0.070±0.007	0.156±0.015	0.864±0.029	0.978±0.008	0.996±0.003
	✓			0.116±0.004	0.008±0.001	0.062±0.002	0.148±0.005	0.872±0.011	0.990±0.001	0.999±0.001
		✓		0.134±0.004	0.011±0.001	0.071±0.002	0.169±0.004	0.819±0.011	0.986±0.002	0.998±0.001
			✓	0.092±0.007	0.006±0.001	0.057±0.005	0.126±0.010	0.919±0.014	0.984±0.003	0.996±0.002
	✓		✓	**0.090**±0.001	**0.005**±0.000	**0.046**±0.002	**0.114**±0.002	**0.943**±0.003	**0.997**±0.001	**1.000**±0.000
		✓	✓	0.120±0.010	0.009±0.002	0.064±0.008	0.153±0.014	0.862±0.031	0.988±0.004	0.998±0.001

B. Quantitative Results

Results comparing with the self-supervised baseline and the supervised baseline are reported in Tables IV and V. To further validate the effectiveness of MonoLoT, an ablation study was conducted to assess the significance of various design components within the architecture (refer to Table VI and Table VII). The results are reported for the C3VD and SimCol datasets.

C. Qualitative Results

A comprehensive qualitative evaluation of MonoLoT is organised as follows:

(i) Results for the high-definition phantom (**C3VD**) [16] and virtual colonoscopy (**SimCol**) [17] are presented in Figs. 6 and 7, respectively.
(ii) The **generalisation capability** of MonoLoT on **real GI endoscopy** is evaluated using the upper-GI dataset [19] (see Fig 9), the EndoSLAM dataset, and the EndoMapper dataset (see Fig 8).
(iii) **Steering capabilities** consists linear trajectories along

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

TABLE VII
ABLATION STUDIES OF FBIS, RBIS AND PM ON THE SIMCOL DATASET. THE BEST RESULTS ARE IN **BOLD**. THE SECOND BEST ARE UNDERLINED.

DepthNet	$\mathcal{L}_{bis}$†	$\mathcal{L}_{bis}$*	$\mathcal{L}_m$	Depth Error(↓)				Depth Accuracy(↑)		
				Abs rel	Sq rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Monodepth2 [10]		(baseline)		0.160±0.022	1.008±0.306	0.724±0.084	0.169±0.016	0.907±0.008	0.953±0.004	0.970±0.004
	✓			**0.074**±0.000	**0.060**±0.001	**0.365**±0.001	**0.103**±0.000	**0.962**±0.001	**0.989**±0.000	**0.995**±0.000
		✓		0.078±0.001	0.068±0.001	0.370±0.005	0.106±0.001	0.957±0.001	0.987±0.000	0.994±0.000
			✓	0.124±0.001	0.613±0.013	0.522±0.005	0.140±0.001	0.937±0.000	0.976±0.001	0.987±0.000
	✓		✓	0.084±0.002	0.098±0.020	0.403±0.009	0.115±0.002	0.952±0.001	0.985±0.001	0.992±0.001
		✓	✓	0.074±0.002	0.066±0.000	0.370±0.006	0.104±0.002	0.959±0.002	0.988±0.000	0.995±0.001
Lite-mono [12]		(baseline)		0.171±0.008	1.318±0.119	0.742±0.026	0.167±0.004	0.913±0.003	0.951±0.001	0.967±0.001
	✓			0.080±0.001	0.073±0.001	0.400±0.003	0.112±0.001	0.954±0.001	0.985±0.000	0.993±0.000
		✓		0.084±0.003	0.079±0.004	0.406±0.006	0.115±0.003	0.950±0.003	0.984±0.000	0.993±0.000
			✓	0.117±0.002	0.640±0.045	0.491±0.003	0.133±0.001	0.943±0.001	0.979±0.000	0.988±0.000
	✓		✓	0.078±0.000	0.069±0.001	0.401±0.003	0.110±0.000	0.954±0.001	0.986±0.000	**0.994**±0.000
		✓	✓	**0.075**±0.000	**0.069**±0.000	**0.394**±0.003	**0.107**±0.001	**0.957**±0.001	**0.987**±0.000	**0.994**±0.000
MonoViT [11]		(baseline)		0.143±0.014	0.562±0.182	0.627±0.050	0.171±0.008	0.901±0.006	0.951±0.003	0.969±0.002
	✓			0.085±0.002	0.074±0.001	0.402±0.006	0.113±0.002	0.952±0.002	0.986±0.000	0.994±0.000
		✓		0.093±0.002	0.081±0.002	0.424±0.012	0.121±0.002	0.943±0.003	0.984±0.001	0.993±0.000
			✓	0.128±0.002	0.796±0.069	0.533±0.011	0.139±0.001	0.941±0.001	0.977±0.001	0.986±0.000
	✓		✓	**0.073**±0.002	**0.065**±0.001	**0.394**±0.008	**0.105**±0.002	**0.958**±0.001	**0.988**±0.000	**0.995**±0.000
		✓	✓	0.076±0.001	0.070±0.000	0.401±0.005	0.108±0.001	0.956±0.001	0.987±0.000	0.994±0.000

the X axis, Y axis, and two diagonal lines with angles ranging from ±40°to ±240°, and circular trajectories with radii ranging from 40°to 160°, all in 40° increments (see Fig 10).

(iv) The **automatic robotic endoscopy on the realistic upper GI tract phantom** showcases the user interface, GT pose in 3D space, and the appearance of the robotic platform during insertion and retraction (see **Video S1**). The insertion or retraction process encompasses five stages (see Fig. 1).

VI. DISCUSSION

This study aimed to improve the performance of self-supervised monocular depth estimation methods in real-world endoscopic scenarios with low-texture regions. Furthermore, we designed and implemented a robotic platform for controlling digestive endoscopy and ultimately validated the feasibility of using the proposed monocular depth estimation method for automatic robotic endoscopy. In this section, we will provide a detailed analysis of the proposed objective functions and the robotic platform.

A. Performance of the Monocular Depth Estimation

Typically, there exists a performance gap between self-supervised methods and GT depth-supervised methods [10], as well as between CNN-based methods and Transformer-based methods [11], [32]. However, our experiments on low-texture scenes suggest an exception, as they demonstrate that self-supervised methods can achieve competitive performance comparable to GT depth-supervised methods. Moreover, the disparity between CNN-based methods and Transformer-based methods can be mitigated.

1) C3VD Dataset: As demonstrated in Table IV, the "baselines (M)" exhibited notably lower accuracy compared to the "baselines (D)", particularly when evaluating the "$\delta < 1.25$" metric. However, the performance of MonoLoT-enabled models, namely "MonoLoT (M)", improved significantly, even surpassing the "baselines (D)". For instance, the "$\delta < 1.25$" metric of "Monodepth2 (M)" increased from 0.746 (baseline) to 0.934 (MonoLoT), surpassing that of "Monodepth2 (D)" (0.877). A similar trend is observable for Lite-mono and MonoViT. In this case, the performance gap of the "$\delta < 1.25$" between "Monodepth2 (M)" and "Lite-mono (M)" reduced from 0.118 (baseline) to 0.01 (MonoLoT).

Comprehensive ablation studies are reported in Table VI, highlighting that a majority of the combinations involving the proposed batch image shuffle and point matching loss yield significant performance improvements for the baselines. Notably, the combination of RBIS and PM demonstrated superior performance on Monodepth2, while the combination of FBIS and PM outperformed other combinations on Lite-mono and MonoViT.

Additional qualitative results for the C3VD dataset (see Fig. 6) reveal a notable side effect of FBIS on Monodepth2. This effect is characterised by the presence of a borderline running from top to bottom and another border line from left to right in the depth estimation. Fortunately, this issue was effectively mitigated by employing RBIS. Such an effect was not observed in Transformer-based methods, such as Lite-mono and MonoViT. Experimental results showed that, in most case, FBIS brought more benefit to Transformer-based methods compared to RBIS.

2) SimCol Dataset: A similar trend of improved performance is observed in Table V on the SimCol dataset. By using the proposed loss function, the "RMSE" metric of all the "baseline (M)" models was reduced by half, while the "$\delta < 1.25$" metric increased by around 5%. We also conduct extensive experiments of ablation studies on the SimCol dataset as reported in Table VII, which shows all the combinations of proposed loss functions brought increasing performance. The combination with the best performance are selected as our final models for comparison with the baseline in Table V, except the "Monodepth2 + $\mathcal{L}_{bis}^{\dagger}$" model because of the side effect discussed before. In alternative, the second best results of Monodepth2 in Table VII, namely "Monodepth2 +

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

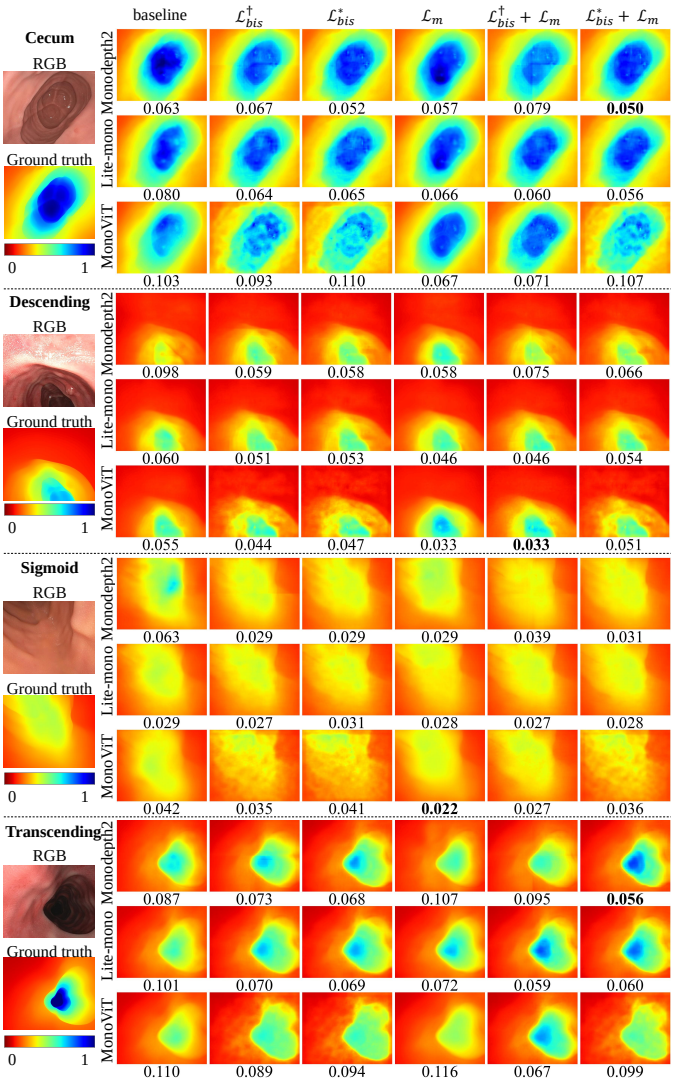


Fig. 6. Qualitative C3VD [16] results. Monodepth2 + $\mathcal{L}_{bis}^{*}$ + $\mathcal{L}_{m}$, Lite-mono + $\mathcal{L}_{bis}^{\dagger}$ + $\mathcal{L}_{m}$, and MonoViT + $\mathcal{L}_{bis}^{\dagger}$ + $\mathcal{L}_{m}$ exhibit equally excellent results, outperforming the others. The RMSE values are presented below each depth map, with the best results highlighted in **bold**.

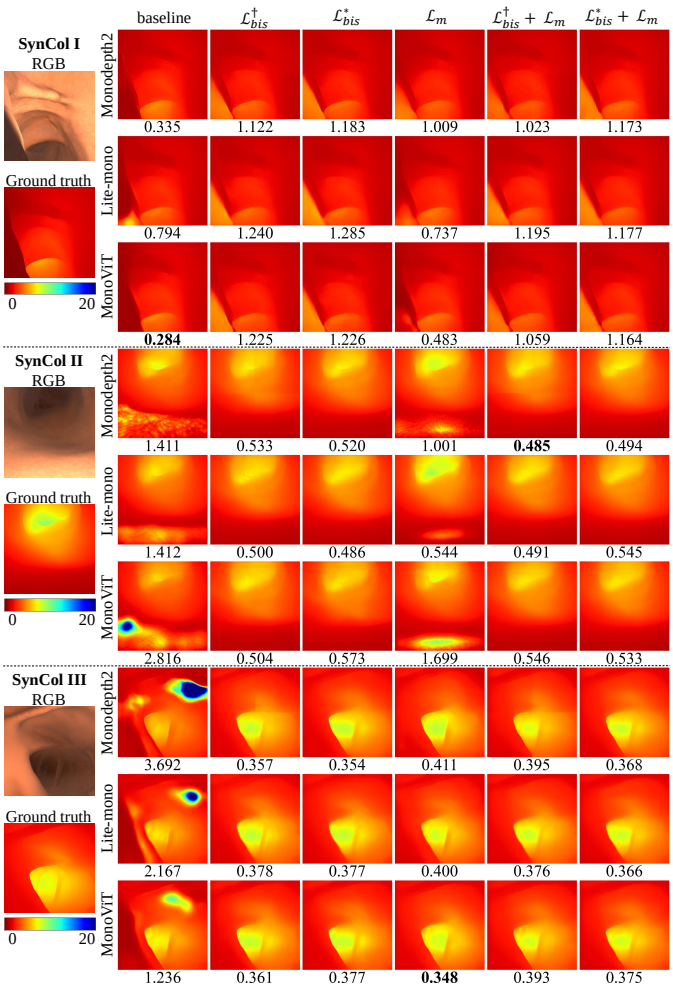


Fig. 7. Qualitative SimCol [17] results. GT depth values rescaled to [0, 20] cm. Monodepth2 + $\mathcal{L}_{bis}^{*}$ + $\mathcal{L}_{m}$, Lite-mono + $\mathcal{L}_{bis}^{*}$ + $\mathcal{L}_{m}$, and MonoViT + $\mathcal{L}_{bis}^{\dagger}$ + $\mathcal{L}_{m}$ exhibit equally excellent results, outperforming the others. The RMSE values are presented below each depth map, with the best results highlighted in **bold**.

$\mathcal{L}_{bis}^{*}$ + $\mathcal{L}_{m}$", showed competing performance to "Monodepth2 + $\mathcal{L}_{bis}^{\dagger}$" was taken as the best model.

The qualitative results depicted in Fig. 7 provide instances where the baseline models encountered failures. However, "SynCol I" in Fig. 7 represents a distinct failure case unrelated to the baseline models, attributable to the lighting setup. Specifically, the lower-left part of the image, where the digestive tract obstructs one of the light sources, results in a dark region. It's noteworthy that the SimCol dataset employs two light sources, one on each side of the camera. Consequently, when one light source is obstructed, the other remains work. In contrast, real endoscopes typically feature a single light source. If this light source becomes obstructed, it leads to two potential outcomes: firstly, a strong reflection may emerge, causing the obstructed area to appear bright. Secondly, most of the light rays may be entirely blocked, resulting in reduced overall brightness in the image. Consequently, we posit that the lighting setup in the SimCol dataset could potentially impact the model's generalisation to real endoscopic images, as the lighting condition is a critical factor in determining depth.

3) Generalisability of the Model: In general, models trained on the C3VD dataset exhibited greater robustness to reflected light compared to those trained on the SimCol dataset, as the illumination setting and texture of the C3VD dataset closely resemble real endoscopy conditions (see Fig. 9). The predicted depth maps of Monodepth2 are similar to those of Lite-mono and MonoViT when models are trained with the proposed objective function, thus validating that MonoLoT has reduced the performance gap between CNN-based methods and Transformer-based methods. Moreover, MonoLoT-enabled methods trained on the C3VD dataset produced more accurate range of depth than the others. For example, MonoLoT-enabled self-supervised Monodepth2, Lite-mono, and MonoViT trained on C3VD produced accurate depth maps showed that the middle-upper body in Fig. 9 covers a wider range of depth than the others. "Lite-mono + $\mathcal{L}_{bis}^{\dagger}$ + $\mathcal{L}_{m}$" model trained on C3VD demonstrated strong generalisability and was subsequently integrated to the robotic platform. Ad-

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

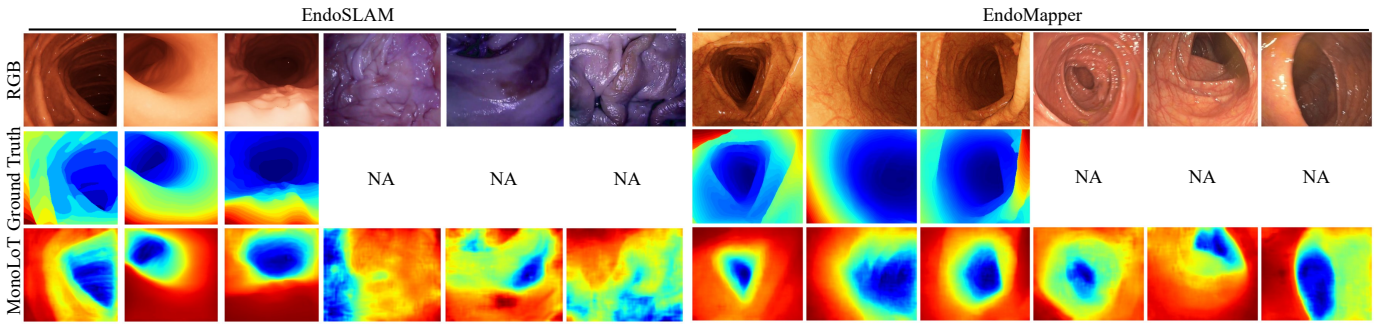


Fig. 8. Zero-shot evaluation on EndoSLAM and EndoMapper datasets demonstrates MonoLoT's capability to generate depth maps that accurately capture spatial structures and exhibit robustness to reflections for unseen synthetic and real images.

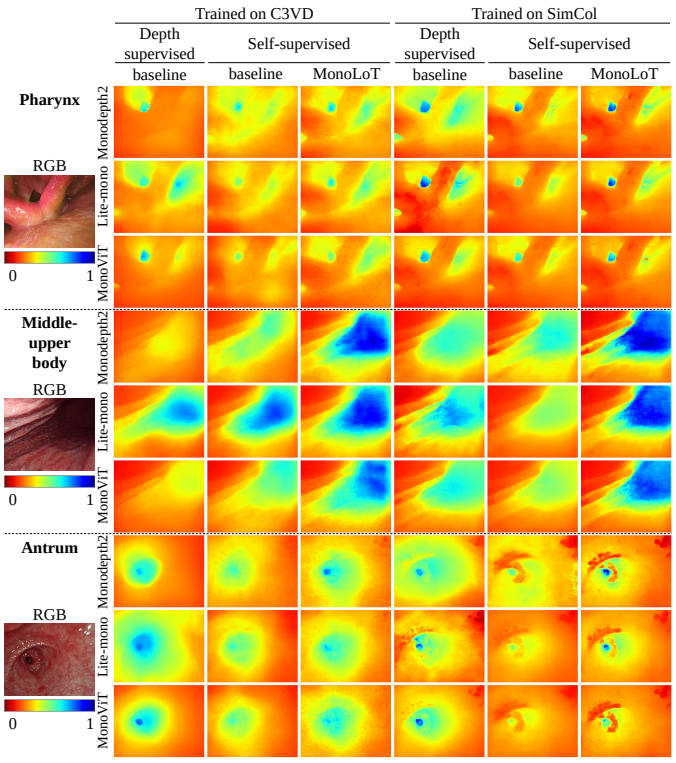


Fig. 9. Qualitative results showing the prediction of selected anatomical landmarks in upper-GI images [19]. MonoLoT-enabled methods trained on C3VD showed an accurate range of depth and robustness to reflected light. Predicted depths of models trained on the SimCol dataset are clipped to the range [0, 10] cm and then resized to the range [0, 1] for alignment with the C3VD dataset.

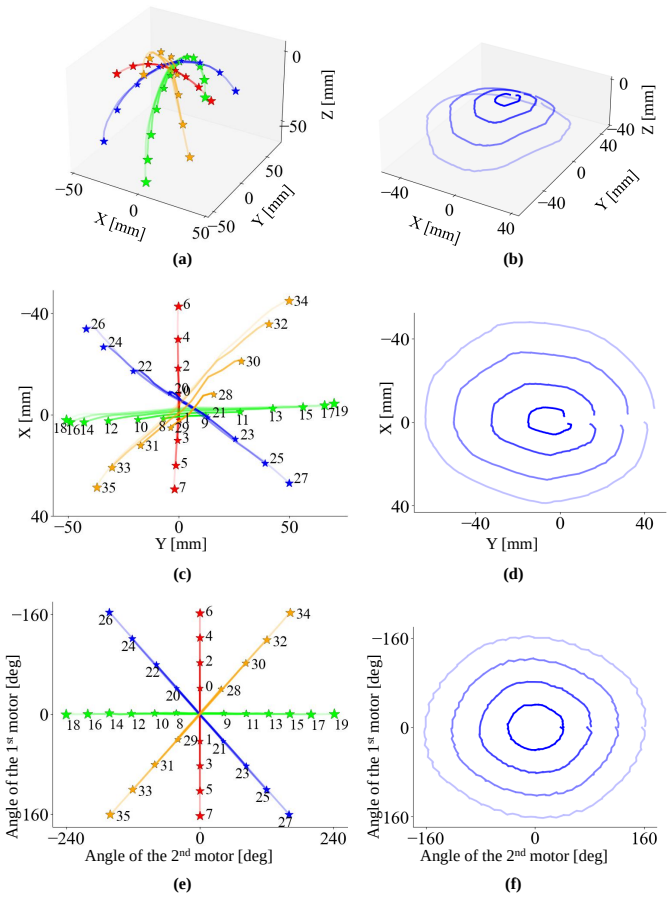


Fig. 10. Visualisation of linear trajectories (a, c) and circular trajectories (b, d) in 3D space (a, b) and the X-Y plane (c, d), accompanied by the corresponding motor degrees (e, f). The initiation and termination points of each linear trajectory are marked with timestamps and ⋆ symbols.

ditionally, we evaluated MonoLoT on the EndoSLAM dataset and the EndoMapper dataset under a zero-shot setting, as shown in Fig. 8. Since the intrinsic matrix of the training set and test set are different, the predicted depths are inevitably different from the ground truth. Due to discrepancies in the intrinsic matrices between the training and test sets, the predicted depths diverge from the ground truth. However, the qualitative outcomes indicate that our approach successfully captures accurate details across a wide range of endoscopic videos.

B. Performance of the Robotic Platform

1) Mechanical Structure: The proposed robotic platform provides a concise interface to the Olympus endoscopy system and is portable enough to be adapted for use with similar endoscopy systems due to its simplified design (see Fig. 5 (a)). Considering the potential motion errors introduced by the soft endoscope mechanism and forces between the endoscope and tissues, it is essential to assess their impact on endoscopic kinematics. Fortunately, we have found that PID control is a viable alternative. Experimental results for linear and circular

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

motions using PID control in an open space have shown that the trajectories of the endoscopic tip closely follow the motion patterns of the motors (see Fig. 10). Though relocation error is unavoidable, the depth-based PID control can effectively address this issue.

2) Depth-based PID Control for Automatic Intervention: To illustrate the real-time applicability of depth-based PID control in our robotic platform, we present a test video showcasing the insertion and retraction procedures in a realistic upper GI phantom. Full details can be seen in the supplementary materials (**Video S1**). MonoLoT has demonstrated excellent generalisability on a realistic upper GI phantom (see Fig. 1), providing effective support for the anchor point computation. Unlike [7] and [8], our controller prototype is designed to be adaptable to existing endoscopes and robotic platforms, regardless of their structure, as long as they offer appropriate motion interfaces. We utilised the PID controller for the movement of the bending section because this model-free controller is robust and widely used in the same application, such as [33]. The PID controller is more suitable in this scenario than the Jacobian controller because the commercial endoscope already provides a control interface, and it is not possible to directly drive the sheath to control the endoscope kinematically. Therefore, we ultimately use the PID controller to control the movement of the endoscope. Guided by an anchor point determined through depth estimation, the PID controller systematically navigates the endoscope from the pharynx to the antrum. Our experimental findings robustly demonstrate the feasibility and effectiveness of depth-based PID control for autonomous interventions in digestive endoscopy. Consequently, this proposed method holds significant promise for clinical application in automated gastrointestinal examinations, thereby liberating endoscopists to concentrate more on critical decision-making processes. Furthermore, intricate endoscopic procedures such as endoscopic retrograde cholangiopancreatography (ERCP) and endoscopic submucosal dissection (ESD), which typically require multiple operators for manual endoscopic interventions and handling of surgical instruments, could benefit from the introduction of an automated robotic system. This system has the potential to mitigate issues related to miscommunication, the allocation of costly medical resources, and extended durations of ERCP and ESD procedures by assisting surgeons in the execution of these endoscopic tasks.

C. Limitations and Future Works

Our experimental results expose some shortcomings of our robot design. As the intervention distance increases, the deformation of endoscope also increases due to the increasing contact between the endoscope and the gastrointestinal tract, leading to difficulty in inserting the endoscope further. This can be solved with custom-made endoscopes with variable stiffness and additional manipulators [34]. Furthermore, autonomous endoscopic intervention requires safe interaction with the GI tract, thus an additional interactive force measurement helps ensure the intervention's safety [35]. On the other hand, our vision-based algorithm overcomes the low-texture challenge, but still shares some limitations similar to the other self-supervised depth estimation algorithm, incapable of predicting the absolute depth of images in various scenes. Fortunately, the latest trends based on few-shot learning and pre-trained vision language model [36] shows a potential to address this issue. Moreover, the predicted depth around the vanishing point of the GI image shows a counter-intuitive closed surface that deserves to be explored in future work.

VII. Conclusion

Our research has introduced an innovative approach that addresses the challenges associated with self-supervised monocular depth estimation in digestive endoscopy. We have addressed two critical aspects: the limitations of self-supervised depth estimation in low-texture scenes and the application of depth estimation in visual servoing for digestive endoscopy.

Our investigation has revealed that the struggles of self-supervised depth estimation in low-texture scenes stem from inaccurate photometric reconstruction losses. To overcome this, we have introduced the point-matching loss, which refines the reprojected points. Furthermore, during the training process, data augmentation is achieved through batch image shuffle loss, significantly improving the accuracy and generalisation capability of the depth model. The combined contributions of the point matching loss and batch image shuffle loss have boosted the baseline accuracy by a minimum of 5% on both the C3VD and SimCol datasets, surpassing the generalisability of ground truth depth-supervised baselines when applied to upper-GI datasets. Moreover, the successful implementation of our robotic platform for automatic intervention in digestive endoscopy demonstrates the practical and impactful application of monocular depth estimation technology.

In conclusion, our work not only advances the field of endoscopic image analysis but also holds promise for enhancing clinical procedures and ultimately improving patient outcomes in gastrointestinal healthcare.

References

[1] Q. He, S. Bano, O. F. Ahmad, B. Yang, X. Chen, P. Valdastri, L. B. Lovat, D. Stoyanov, and S. Zuo, "Deep learning-based anatomical site classification for upper gastrointestinal endoscopy," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 7, pp. 1085–1094, Jul. 2020.

[2] A. R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, "Whole stomach 3d reconstruction and frame localization from monocular endoscope video," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 7, pp. 1–10, 2019.

[3] K. B. Ozyoruk, G. I. Gokceler, T. L. Bobrow, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira *et al.*, "Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos," *Medical image analysis*, vol. 71, p. 102058, 2021.

[4] G. Ciuti, K. Skonieczna-Żydecka, W. Marlicz, V. Iacovacci, H. Liu, D. Stoyanov, A. Arezzo, M. Chiurazzi, E. Toth, H. Thorlacius *et al.*, "Frontiers of robotic colonoscopy: a comprehensive review of robotic colonoscopes and technologies," *Journal of clinical medicine*, vol. 9, no. 6, p. 1648, 2020.

[5] F. Chadebecq, L. B. Lovat, and D. Stoyanov, "Artificial intelligence and automation in endoscopy and surgery," *Nature Reviews Gastroenterology & Hepatology*, vol. 20, no. 3, pp. 171–182, 2023.

[6] A. Pore, M. Finocchiaro, D. Dall'Alba, A. Hernansanz, G. Ciuti, A. Arezzo, A. Menciassi, A. Casals, and P. Fiorini, "Colonoscopy navigation using end-to-end deep visuomotor control: A user study," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 9582–9588.

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2024.3423791

[7] T. Yang, Y. Yang, P. Wang, Y. Cao, Z. Yang, and H. Liu, "A lumen-adapted navigation scheme with spatial awareness from monocular vision for autonomous robotic endoscopy," *Robotics and Autonomous Systems*, vol. 165, p. 104444, 2023.

[8] L. Ott, F. Nageotte, P. Zanne, and M. De Mathelin, "Robotic assistance to flexible endoscopy by physiological-motion tracking," *IEEE Transactions on Robotics*, vol. 27, no. 2, pp. 346–359, 2011.

[9] L. Barducci, B. Scaglioni, J. Martin, K. L. Obstein, and P. Valdastri, "Active stabilization of interventional tasks utilizing a magnetically manipulated endoscope," *Frontiers in Robotics and AI*, vol. 9, p. 854081, 2022.

[10] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[11] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, "Monovit: Self-supervised monocular depth estimation with a vision transformer," in *2022 International Conference on 3D Vision (3DV)*, 2022, pp. 668–678.

[12] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 537–18 546.

[13] A. Rau, P. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka, L. B. Lovat, and D. Stoyanov, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *International journal of computer assisted radiology and surgery*, vol. 14, pp. 1167–1176, 2019.

[14] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman, "The temporal opportunist: Self-supervised multi-frame monocular depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1164–1174.

[15] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[16] T. L. Bobrow, M. Golhar, R. Vijayan, V. S. Akshintala, J. R. Garcia, and N. J. Durr, "Colonoscopy 3d video dataset with paired depth from 2d-3d registration," *arXiv preprint arXiv:2206.08903*, 2022.

[17] A. Rau, B. Bhattarai, L. Agapito, and D. Stoyanov, "Bimodal camera pose prediction for endoscopy," *IEEE Transactions on Medical Robotics and Bionics*, vol. 5, no. 4, pp. 978–989, 2023.

[18] P. Azagra, C. Sostres, Á. Ferrández, L. Riazuelo, C. Tomasini, O. L. Barbed, J. Morlana, D. Recasens, V. M. Batlle, J. J. Gómez-Rodríguez *et al.*, "Endomapper dataset of complete calibrated endoscopy procedures," *Scientific Data*, vol. 10, no. 1, p. 671, 2023.

[19] Q. He, S. Bano, O. F. Ahmad, B. Yang, X. Chen, P. Valdastri, L. B. Lovat, D. Stoyanov, and S. Zuo, "Deep learning-based anatomical site classification for upper gastrointestinal endoscopy," *International journal of computer assisted radiology and surgery*, vol. 15, pp. 1085–1094, 2020.

[20] A. R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, "Learning-based depth and pose estimation for monocular endoscope with loss generalization," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 3547–3552.

[21] J. Rodriguez-Puigvert, D. Recasens, J. Civera, and R. Martinez-Cantin, "On the uncertain single-view depths in colonoscopies," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*. Springer, 2022, pp. 130–140.

[22] B. Li, Y. Huang, Z. Liu, D. Zou, and W. Yu, "Structdepth: Leveraging the structural regularities for self-supervised indoor depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 663–12 673.

[23] K. Saunders, G. Vogiatzis, and L. J. Manso, "Self-supervised monocular depth estimation: Let's talk about the weather," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8907–8917.

[24] C. Zhao, M. Poggi, F. Tosi, L. Zhou, Q. Sun, Y. Tang, and S. Mattoccia, "Gasmono: Geometry-aided self-supervised monocular depth estimation for indoor scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 209–16 220.

[25] Z. Liu, R. Li, S. Shao, X. Wu, and W. Chen, "Self-supervised monocular depth estimation with self-reference distillation and disparity offset refinement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7565–7577, 2023.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[27] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "Mpvit: Multi-path vision transformer for dense prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7287–7296.

[28] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[29] A. Ali, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek *et al.*, "Xcit: Cross-covariance image transformers," *Advances in neural information processing systems*, vol. 34, pp. 20 014–20 027, 2021.

[30] Z. Wang, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[31] D. Eigen, C. Puhrsch, and R. Fergus, "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[32] Y. Liu and S. Zuo, "Self-supervised monocular depth estimation for gastrointestinal endoscopy," *Computer Methods and Programs in Biomedicine*, p. 107619, 2023.

[33] S. Zuo, T. Chen, X. Chen, and B. Chen, "A wearable hands-free human-robot interface for robotized flexible endoscope," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3953–3960, 2022.

[34] T. Liang, C. Zhang, Y. Wang, K. Kong, X. Chen, B. Wei, S. Wang, and S. Zuo, "A novel miniature flexible robotic system for endoscopic mucosal dissection: an animal experimental study," *Journal of Robotic Surgery*, vol. 18, no. 1, pp. 1–7, 2024.

[35] Z. Zhang, C. Zhang, and S. Zuo, "A novel bioinspired whisker sensor for gastrointestinal endoscopy," *IEEE/ASME Transactions on Mechatronics*, 2023.

[36] X. Hu, C. Zhang, Y. Zhang, B. Hai, K. Yu, and Z. He, "Learning to adapt clip for few-shot monocular depth estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5594–5603.

Authorized licensed use limited to: University College London. Downloaded on July 18,2024 at 15:31:36 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.