Medical Image Analysis 99 (2025) 103379

Contents lists available at ScienceDirect

Medical Image Analysis

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

Multi-task learning with cross-task consistency for improved depth estimation in colonoscopy

Pedro Esteban Chavarrias Solano [a], Andrew Bulpitt [a], Venkataraman Subramanian [b,c], Sharib Ali [a,*]

[a] School of Computer Science, Faculty of Engineering and Physical Sciences, University of Leeds, Leeds, LS2 9JT, United Kingdom
[b] Department of Gastroenterology, Leeds Teaching Hospitals NHS Trust, Leeds, UK
[c] Division of Gastroenterology and Surgical Sciences Leeds Institute of Medical Research at St James's University of Leeds, Leeds, UK

ARTICLE INFO

Keywords:
Deep learning
Monocular depth estimation
Surface normal prediction
Multi-task learning
Cross-task consistency
3D colonoscopy

ABSTRACT

Colonoscopy screening is the gold standard procedure for assessing abnormalities in the colon and rectum, such as ulcers and cancerous polyps. Measuring the abnormal mucosal area and its 3D reconstruction can help quantify the surveyed area and objectively evaluate disease burden. However, due to the complex topology of these organs and variable physical conditions, for example, lighting, large homogeneous texture, and image modality estimating distance from the camera (*aka* depth) is highly challenging. Moreover, most colonoscopic video acquisition is monocular, making the depth estimation a non-trivial problem. While methods in computer vision for depth estimation have been proposed and advanced on natural scene datasets, the efficacy of these techniques has not been widely quantified on colonoscopy datasets. As the colonic mucosa has several low-texture regions that are not well pronounced, learning representations from an auxiliary task can improve salient feature extraction, allowing estimation of accurate camera depths. In this work, we propose to develop a novel multi-task learning (MTL) approach with a shared encoder and two decoders, namely a surface normal decoder and a depth estimator decoder. Our depth estimator incorporates attention mechanisms to enhance global context awareness. We leverage the surface normal prediction to improve geometric feature extraction. Also, we apply a cross-task consistency loss among the two geometrically related tasks, surface normal and camera depth. We demonstrate an improvement of 15.75% on relative error and 10.7% improvement on $\delta_{1.25}$ accuracy over the most accurate baseline state-of-the-art Big-to-Small (BTS) approach. All experiments are conducted on a recently released C3VD dataset, and thus, we provide a first benchmark of state-of-the-art methods on this dataset.

1. Introduction

Colorectal cancer (CRC) is among the third most common type of cancer in the world, imposing a healthcare burden globally. The estimated number of new CRC cases in 2023 will likely increase to 153,020 (Siegel et al., 2023). Optical endoscopy is the gold standard procedure for diagnosing and treating CRC (Rex et al., 2015). Despite its great potential, the colonoscopic procedure is subject to the clinician's experience as they have to deal with a complex anatomical environment, imaging artefacts, and a limited field of view. A retrospective analysis of clinical endoscopic video realised that 9.6% of the colon surface is never imaged during the screening procedure (McGill et al., 2018). Those missed regions can contribute to an estimated 22% of precancerous undetected lesions found by Van Rijn et al. (2006). The development of an intelligent system to reduce the missed detection rate and guide the clinician to potential regions of interest has caught the attention of the medical computer vision research community. Several methods have been developed to detect and segment polyp instances accurately. Unlike polyp detection and segmentation methods, 3D computer vision techniques in colonoscopy have not been widely explored. Some applications within this field cover lesion extent prediction (Abdelrahim et al., 2022; Ali et al., 2021), observational coverage of the colon (Armin et al., 2016; Bobrow et al., 2023), and 3D reconstruction (Zhang et al., 2021a).

The development of 3D computer vision applications in colonoscopy is limited due to the difficulty in acquiring ground truth labels. Compared to labelled datasets required for training detection and segmentation methods, acquiring datasets with accurate surface information (e.g., surface normal vectors and depth maps) for 3D scene understanding is far from practical during a clinical endoscopy procedure.

* Corresponding author.
E-mail address: s.s.ali@leeds.ac.uk (S. Ali).

https://doi.org/10.1016/j.media.2024.103379
Received 7 November 2023; Received in revised form 2 September 2024; Accepted 22 October 2024
Available online 4 November 2024

Hence, the development of commercial and computed tomography-derived silicone models has been suggested as an alternative method for data acquisition. The C3VD dataset leverages a novel technique for generating high-fidelity silicone phantom models of the colon with various textures and colours (Bobrow et al., 2023). The phantom model and a 2D-3D video registration algorithm were used to generate a ground truth dataset with pixel-level registration to a known 3D model. Unlike digital phantom datasets (Rau et al., 2019; Zhang et al., 2021a), silicone phantom models are closer to real-world clinical datasets and reflect more accurately to an actual colonoscopic procedure. Therefore, most methods that are trained on synthetically (digital) generated data (Rau et al., 2023; Mahmood and Durr, 2018; Jeong et al., 2024; Zhang et al., 2021b) suffer a relatively higher domain gap compared to those trained on a dataset acquired from a 3D printed phantom model.

Depth estimation is crucial for understanding geometry structure and a fundamental task in computer vision for 3D scene reconstruction. Most depth estimation approaches initially relied on stereo matching and triangulation methods to calculate the disparity of two 2D images. However, these binocular-based depth estimation methods require at least two fixed cameras. In addition, capturing enough features to match between images becomes challenging when the scene does not have enough texture (Ming et al., 2021). Due to the spatial constraint imposed by the lumen (e.g., size and complex non-uniform shape) of the gastrointestinal tract, monocular systems have been more attractive than stereo systems. Monocular depth estimators aim to learn a mapping between a single RGB image and their corresponding depth values by capturing features that represent geometric structures.

Single-view deep learning methods (Lee et al., 2021; Yuan et al., 2022; Piccinelli et al., 2023) for monocular depth estimation make use of monocular visual cues, e.g., texture gradients and lighting variations. These methods learn to incorporate scene priors without the need to compute camera motion. While this is an advantage, they usually perform well only in similar samples to those presented during the training stage, making these networks harder to generalise. Their accuracy is also affected by the essential ambiguity of the problem as an infinite number of world scenes and camera positions could have produced a given image (Eigen et al., 2014; Lee et al., 2021; Zhang et al., 2021a; Piccinelli et al., 2023). Therefore, recent approaches suggest using multi-task learning schemes as they leverage auxiliary tasks with depth-related features (Ming et al., 2021). Adopting an additional task aims to enhance the extraction of relevant geometrical cues during the encoding stage, leading to a better performance overall (Ming et al., 2021; Chen et al., 2020a; Qi et al., 2018). For example, joint learning of depth and optical flow (Zou et al., 2018; Chen et al., 2020a) and joint learning of depth and surface normal (Bae et al., 2022) have shown competitive performances on natural scene datasets. Multi-task learning approaches have been extensively studied in natural scenes, but their applicability in the colonoscopy domain has not been widely validated. In addition, consistency among tasks is desirable since both generate a particular domain representation of the same underlying reality (Zou et al., 2018; Zamir et al., 2020).

To this end, we propose Col3D-MTL, a novel multi-task learning with a cross-task consistency approach for joint monocular depth and surface normal prediction featuring attention mechanisms to improve global context awareness. We validate our study on a public colonoscopy dataset fully acquired using a silicone phantom model. Our proposed framework can be summarised on the following five main contributions:

1. **A multi-task learning approach with novel unit normal computational blocks:** We introduce Col3D-MTL, a new framework for joint estimation of monocular depth and surface normal maps on colonoscopy data. Our method consists of one shared encoder and two independent decoders specifically designed for each corresponding task. We present a novel unit normal computation block (UNC block) in our surface normal decoder to enable the accurate recovery of the geometrical orientation of the scene.
2. **Feature enhancement using attention modules at multiple scales:** We incorporate convolutional block attention modules (CBAM) in our depth estimation decoder at different scales to improve both local and global context awareness. Compared to the baseline network (Lee et al., 2021), the use of CBAM modules boosts the performance of the network.
3. **A weighted cross-task consistency loss with a novel depth-to-surface normal block:** We propose a weighted cross-task consistency loss between our predicted surface normal and the computed surface normal utilising depth image gradients (referred to as warped surface normal) to explicitly enforce equilibrium among the two geometrically related tasks. We introduce a depth-to-surface normal module (D2SN module) for learning the end-to-end mapping of depth-to-surface normal.
4. **A new benchmark of monocular depth estimation methods on colonoscopy dataset:** We provide a benchmark comprising SOTA monocular depth estimation methods on the publicly available C3VD dataset. This dataset was fully acquired using a clinical colonoscope on a realistic silicone phantom colon model with pixel-wise ground truth labelled data (Bobrow et al., 2023). We compare our approach against several monocular depth estimation methods. Furthermore, we qualitatively validate our proposed method on two publicly available real colonoscopy patient datasets.
5. **Improved generalisability using a self-supervised learning approach:** To improve the generalisability of our approach on real colonoscopy patient datasets, we propose to pre-train our encoder model using an architecture-agnostic masked image modelling technique (A^2MIM) (Li et al., 2023b).

The rest of the paper is organised as follows. Section 2 presents state-of-the-art methods on monocular depth estimation, multi-task learning, and cross-task consistency. In Section 3, we introduce the C3VD dataset used in this work and our proposed network. Section 4 describes the training and ablation study setups followed in this work. We also present the evaluation metrics and the corresponding quantitative and qualitative results. In Section 5, we discuss the findings of our approach. Finally, our conclusions are presented in Section 6.

2. Related work

This section introduces the most relevant technical aspects needed to understand our contribution. The structure of this section starts with a review of related works to monocular depth estimation in computer vision and endoscopy, followed by a discussion about multi-task learning approaches found in the literature covering the natural scene and endoscopy domains. Finally, we describe the cross-task consistency methods.

2.1. Monocular depth estimation in computer vision

Unlike most traditional stereo matching and triangulation approaches, monocular depth estimation methods only require a single camera to generate a depth map. Even though promising, it is still an ill-posed problem to regress depth from a single image (Ming et al., 2021). The success of deep learning in many computer vision tasks was also translated into the monocular depth estimation task. Learning-based approaches for monocular depth estimation were first introduced in 2014 by Eigen et al. (2014). Their proposed method consists of a convolutional coarse-scale network to predict depth at a global level, followed by a fine-scale network to incorporate finer details, such as object edges.

In Lee et al. (2021), an atrous spatial pyramid pooling (ASPP) module is used to leverage global context information, while the decoder applies local planar guidance at different resolutions to provide geometric guidance to the full-resolution depth map. Kim et al. (2020)

propose a convolution-based encoder–decoder scheme with attention mechanisms embedded in their skip connections to generate refined multi-scale features. A global context module is also introduced at the network's bottleneck to capture representative features on a global scale. In Ranftl et al. (2022), scale- and shift-invariant losses are proposed to mitigate the major sources of incompatibility between datasets. Furthermore, the authors explore optimal strategies for mixing multiple datasets during the training stage to enhance the robustness of their model. Patil et al. (2022) exploit the high degree of regularity in 3D scenes by using a piecewise planarity prior to leverage information from co-planar pixels to improve depth estimation.

Leveraging the enhanced global context understanding of transformer-based architectures, Farooq Bhat et al. (2021) propose an adaptive bin-width estimator based on a mini vision transformer (mViT) network (Zhang et al., 2022). The idea behind this approach is to divide the depth range into several adaptive-width bins and predict the final depth map as a linear combination of the bin centres. All these methods have achieved state-of-the-art performance on popular natural scene depth prediction datasets (such as KITTI Geiger et al., 2012, NYU Nathan Silberman and Fergus, 2012).

2.1.1. Monocular depth estimation applied to endoscopy

In contrast to natural scenes, estimating depth from endoscopy data is highly affected by the lack of ground truth labelled data, low texture, variable lighting conditions and the presence of artefacts, e.g., specularities, saturation, and blurring effects (Ma et al., 2021; Rau et al., 2023). To address the lack of ground truth labelled data, Tukra and Giannarou (2022) present a novel randomly connected encoder–decoder network for self-supervised monocular depth estimation network on surgical data. The random connections within the encoder, which are generated by their cascade random search approach, increase the expressive capabilities of their feature extraction. Conditional generative adversarial networks (cGANs), e.g., pix2pix (Isola et al., 2017), have been used to estimate depth from monocular endoscopic images (Rau et al., 2019; Cheng et al., 2021). However, one major drawback of cGANs is the lack of realistic detail and texture in their representations.

Mahmood and Durr (2018) apply continuous conditional random fields (CRFs) and a convolutional neural network (CNN) to estimate depth from endoscopy images. However, one limitation of this approach is the generation of artefacts due to specular reflections. Yang et al. (2023) propose a geometry-aware monocular depth estimation network based on ManyDepth (Watson et al., 2021) and leverage a depth, smoothness, gradient, normal and geometric consistency losses to enhance depth predictions on endoscopy images. However, their normal loss only relies on the normal map generated from the predicted depth map, which does not faithfully represent the characteristics of the scene (Bae et al., 2022).

2.2. Multi-task learning in computer vision

The complementarity between depth and other geometrically -related features has recently been explored by computer vision researchers (Qi et al., 2018; Chen et al., 2020a; Long et al., 2021). According to a survey on monocular depth estimation (Ming et al., 2021), many approaches suggest incorporating joint multi-task training, in which the extracted features between tasks are projected from one to the other for improved performance (Zou et al., 2018; Ma et al., 2021; Bae et al., 2022). For example, Chen et al. (2020a) develop an architecture composed of two tightly coupled encoder–decoder networks to predict depth map and optical flow as primary and auxiliary tasks, respectively. The authors also introduce exchange blocks to effectively communicate between depth and optical flow networks and an epipolar layer that confines feature matching along the epipolar line. In GeoNet, Qi et al. (2018) propose jointly predicting the depth and surface normal maps from a single image. The method uses two stream-CNNs (ResNet-50 He et al., 2016 and VGG-16 Simonyan and Zisserman, 2015) to predict the initial depth and surface normal maps. It then applies the depth-to-normal and normal-to-depth modules to refine surface normal and depth maps. However, the depth-to-normal network solves a pre-trained least square equation from the initial depth map followed by a residual module to enhance the final prediction, which is not learned in an end-to-end fashion. Similarly, normal-to-depth also solves linear equations through a kernel regression module to infer depth from surface normal.

2.2.1. Multi-task learning applied to medical image analysis

Multi-task learning approaches have been studied for breast cancer segmentation and classification (Wang et al., 2023), left ventricle quantification (Xue et al., 2018), and CT-based identification and quantification (Goncharov et al., 2021). Islam et al. (2021) propose a spatio-temporal multi-task learning network with one shared-encoder and two spatio-temporal independent decoders for instrument segmentation and saliency on a robotic instrument segmentation dataset for endoscopy. Alistair et al. (2023) adopt a multi-task learning scheme for joint optimisation of depth and structured light projection on stereo pairs of surgical images. The disparity maps are generated by performing 2D cross-correlation over the epipolar lines of the predicted light patterns. The results show an improved capability of learning from small datasets while increasing its generalisability performance.

Other multi-task learning approaches on endoscopy have focused on monocular depth and motion estimation (Shao et al., 2022; Liu et al., 2022; Recasens et al., 2021). A 3D colon reconstruction approach suggested by Ma et al. (2021) incorporates a multi-task recurrent neural network (RNN) that estimates depth and camera pose (Wang et al., 2019) to improve the performance of a standard simultaneous localisation and mapping (SLAM) method. Zhang et al. (2021b) leverage surface normal estimation to enhance feature extraction and improve their depth estimation performance. The authors used a shared encoder and two independent decoders for depth and surface normal prediction. However, both decoders have almost the same architecture, with the number of output channels only varying. In this work, we propose task-specific architectures for each decoder.

2.3. Cross-task consistency loss

In visual perception, different domain representations of the same underlying reality or scene are not independent, i.e., a consistent factor between them should exist. A general fully computational method for augmenting training is proposed in Zamir et al. (2020). In this work, the authors introduce a loss for predicting domain y_1 from an input image, x, while imposing consistency with domain y_2. This approach compares prediction y_2 with the warped prediction of y_1 to domain y_2. An unsupervised framework leveraging geometric consistency for training single-view depth and optical flow networks on an unlabelled dataset was proposed in Zou et al. (2018). To enforce geometric consistency, the authors introduced a cross-task consistency loss to minimise the discrepancy between the estimated optical flow and a synthesised flow computed from the predicted depth map and an estimated 6D camera pose.

2.4. Domain gap minimisation for improved generalisability

Rau et al. (2023) propose domain gap reduction in endoscopy (DGRE) for monocular depth prediction. The authors trained a modified version of SharinGAN (Koutilya et al., 2020) to map task-specific meaningful information from synthetic and real-patient colonoscopy data to an intermediate domain. Cycle-consistent generative adversarial networks (CycleGAN) (Zhu et al., 2017) have also been used to map from real to synthetic colonoscopy domains (Masahiro et al., 2022; Jeong et al., 2024).

Wang et al. (2024) propose a framework based on SimCLR (Chen et al., 2020b), a contrastive self-supervised learning approach, to classify colorectal neoplasia based on the NICE classification. In Gan et al.

(2023), self-distillation-based contrastive learning is employed to enhance the detection of polyps. Contrastive approaches aim to pull similar features and push away dissimilar ones, hence their success in downstream discriminative tasks (Liu et al., 2023). Filiot et al. (2023) explore iBOT (Zhou et al., 2022), a generative self-supervised technique based on masked image modelling (MIM), on multiple downstream tasks using histopathological data. MIM leverages co-occurrence relationships among image patches to enhance its generalised feature extraction (Gui et al., 2024; Xie et al., 2022). Since most MIM approaches rely on vision Transformers (Xie et al., 2022; Bao et al., 2022; Minglan et al., 2023; Chen et al., 2023), A^2MIM (Li et al., 2023b) proposes a framework that is compatible with CNNs. A^2MIM masks the input image with the mean RGB value and places the mask token at intermediate feature maps. This method extracts more complex features, e.g., shape and edges, via middle-order interactions among patches and an additional loss in the Fourier domain (Li et al., 2023b).

3. Materials and method

In this section, we describe the datasets used in our work, and we also present the details of our proposed multi-task learning with a cross-task consistency framework.

3.1. The C3VD dataset

In this study, we use the new publicly available Colonoscopy 3D Video Dataset (C3VD) (Bobrow et al., 2023), which is the first video dataset containing 3D pixel-wise ground truth labelled data entirely recorded with a high-definition (HD) clinical colonoscope. The authors created a complete 3D phantom model of the colon, which was digitally sculpted by a board-certified anaplastologist. A 3D-printed phantom model was generated and coated with silicone, silicone pigments and silicone lubricants to mimic the specular appearance of the mucosa, tissue features and vascular patterns.

As described in Bobrow et al. (2023), the data acquisition is performed by mounting the tip of a colonoscope to the end-effector of a robotic arm with previously defined moving trajectories. Then, N, keyframes are extracted and used to generate target depth frames with a pretrained conditional generative adversarial network (Isola et al., 2017). Subsequently, pixel-level ground truth frames are rendered by moving a virtual camera along the recorded trajectory. Although the trajectory of the virtual camera is known, the location of the virtual phantom model relative to this trajectory is not. The pose of the phantom, which can be expressed as a single rigid body transformation, is estimated using a 2D-3D registration approach. The registration process iteratively samples the parameter space for a model transform prediction; then, at each keyframe, compares the geometric contours from the target and rendered depth frames of the current model transform; finally, an evolutionary optimiser called Matrix Adaptation Evolution Strategy (Hansen et al., 2003) updates the model transform aiming to maximise the overlap of the edge frames.

The dataset contains 10,015 frames with paired ground truth depth, surface normal, optical flow, occlusions, six-degrees-of-freedom (DoF) poses, coverage maps, and 3D models. The image resolution of all available data is 1080 × 1350 pixels. In total, 22 videos covering four different colon segments (caecum, transverse, descending, and sigmoid), four texture variations, and three predefined trajectories are publicly available. Fig. 1 shows sample images with their corresponding ground truth depth map, surface normal, and occlusion map.

The train, validation and test splits used in this work are detailed in Table 1. During the training stage, we followed a video-wise split in which we provided data from three colon segments (caecum, transverse, and sigmoid). Our models are validated on data collected from the same segments. To test our methods, we select one video from each colon segment. The caecum (C2V1) and transverse (T3V3) videos evaluate the methods on similar scenes and textures as the training set but from different viewpoints. The sigmoid (S3V2) sequence compares the ability of each method to generalise to different textures. Whereas, the descending (D4V2) colon assesses their generalisability to a completely unseen scene, not present in the training and validation set.

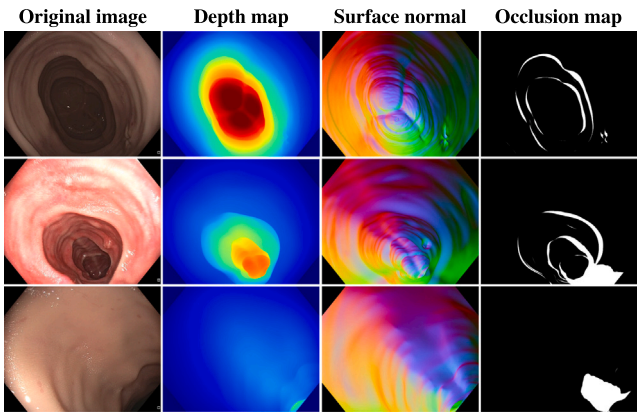

Fig. 1. The C3VD dataset. Sample data including the original RGB image with its corresponding ground truth depth map, surface normal, and occlusion map (Bobrow et al., 2023).

Table 1
Dataset split. Our dataset split follows a video-wise split. Each label corresponds to the colon segment, followed by the texture style and the predefined video trajectory. Here 'c' refers to caecum, 's' refers to sigmoid, 't' refers to transverse and 'd' for descending colon.

Split	Colonoscopy videos	No. of frames
Training	c1v1, c2v2, c2v3, c3v2, s1v3, s2v1, t1v1, t1v3, t2v1, t2v2, t3v2, t4v1, t4v3	6344
Validation	c4v2, c4v3, s3v1, t2v3	1738
Testing	c2v1, d4v2, s3v2, t3v3	1268

3.2. Method

This subsection presents our proposed multi-task learning with a cross-task consistency framework for improved colonoscopy depth estimation. We describe the baseline depth estimation network selected in this study, followed by a review of our multi-task learning scheme and the cross-task consistency approach we incorporate into the network.

3.2.1. Depth estimation network

Our framework is inspired by the monocular depth estimation network, BTS, proposed by Lee et al. (2021). BTS has been referenced by several works (Bae et al., 2022; Yuan et al., 2022; Yang et al., 2023; Piccinelli et al., 2023; Patil et al., 2022; Farooq Bhat et al., 2021) as their state-of-the-art comparison method, achieving, in most of these works, the second best-performing network on the KITTI dataset (Geiger et al., 2012). This method follows an encoder–decoder scheme, in which the encoder performs dense feature extraction while the decoder aims to regress the depth values. While most SOTA methods use depth decoders based on simple bilinear interpolation, the BTS network incorporates local planar guidance (LPG) at different resolutions. LPG layers guide input feature maps to the desired depth map resolution. As a result, it incorporates multi-resolution features that are important in colonoscopy due to the limited texture of the data. The architecture of this network can be identified at the bottom of Fig. 2.

The network uses ResNet-50 (He et al., 2016) as its dense feature extractor, which outputs a feature map of $H/8$ resolution. The backbone is followed by an atrous spatial pyramid pooling module (Chen et al., 2018) to extract contextual information at multiple dilation rates. During the decoding phase, internal outputs are recovered to their

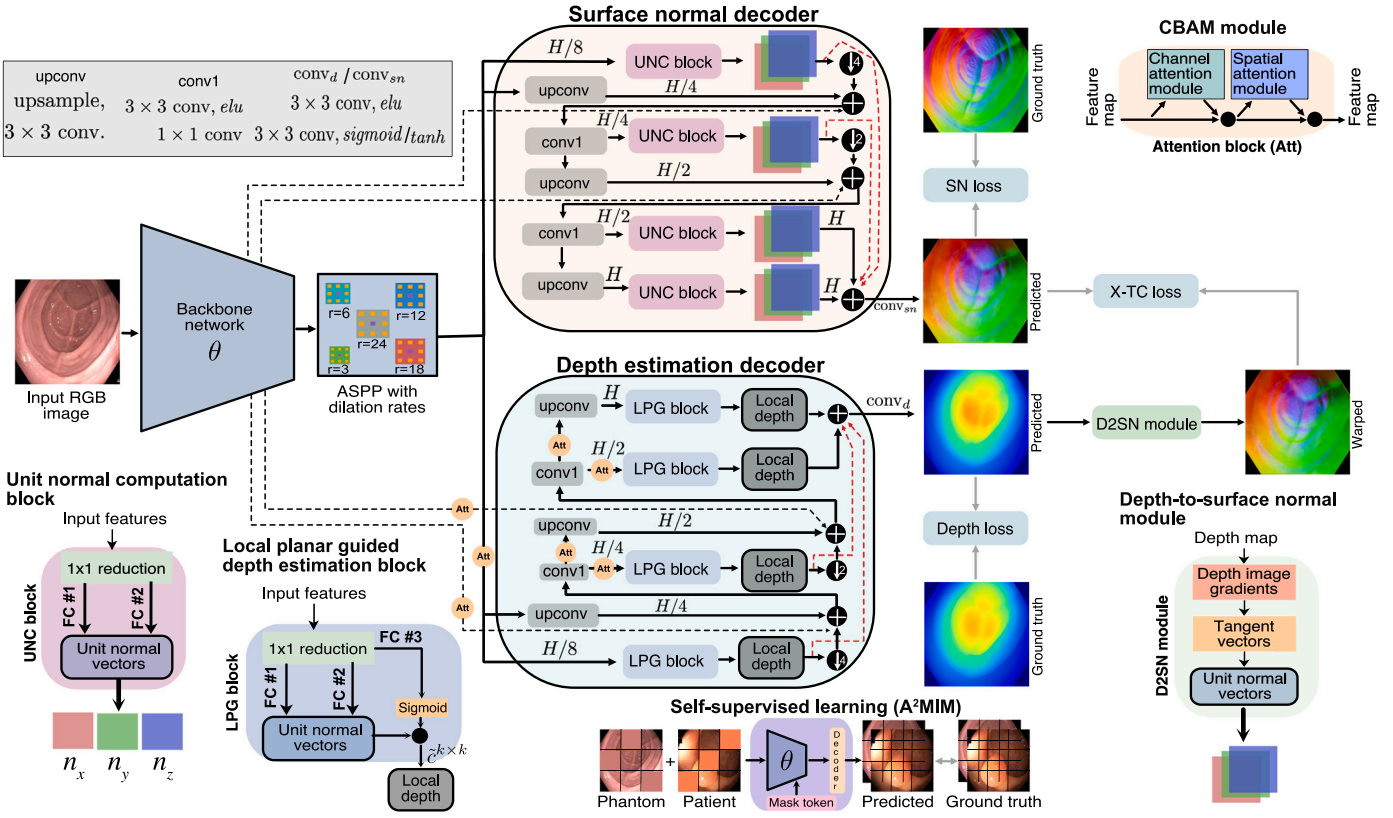

Fig. 2. Multi-task learning with cross-task consistency (Col3D-MTL). Our proposed framework follows the encoder–decoder scheme, in which the encoder consists of a shared backbone, θ, followed by an atrous spatial pyramid (ASPP) module to extract contextual information at different dilation rates. The decoder stage comprises a primary depth estimator decoder (bottom) and an auxiliary surface normal decoder (top). Our unit normal computation block (UNC block) uses two feature channels (FC) to compute the elements of unit normal vectors (n_x, n_y, and n_z). The local planar guided depth estimation block (LPG block) also uses a third FC to compute the perpendicular distance to the camera, which is incorporated together with the unit normals to provide local depth information $\tilde{c}^{k\times k}$ by the ray-plane intersection (see Eq. (2)). CBAM modules (Att) are introduced at the skip connections and after the convolutional layers of the depth decoder to enhance global context awareness. The depth-to-surface normal (D2SN) module receives the predicted depth and outputs a warped surface normal map, which is compared against the surface normal prediction to enforce consistency among tasks. The surface normal decoder and D2SN module can be easily combined with other depth estimators for an end-to-end MTL-X-TC. A^{2}MIM (see Section 3.3) is used to pre-train our encoder, θ, on phantom and patient colonoscopy data following a self-supervised learning approach based on masked image modelling.

original resolution H by a factor of 2 at each LPG block. The LPG block provides geometric guidance to the full-resolution depth map. A final 1×1 convolutional layer is also used to extract the finest estimation ($\tilde{c}^{1\times1}$) after the last *upconv* layer. All the estimated outputs ($\tilde{c}^{k\times k}$) are concatenated and processed through a convolutional layer to compute the final depth estimation. The model is optimised by minimising the scale-invariant logarithmic (SILog) error loss introduced by Eigen et al. (2014).

Multi-scale local planar guidance (LPG block). Most monocular depth estimation networks following an encoder–decoder architecture just apply simple nearest neighbour up-sampling to recover the original resolution of the input image (Kim et al., 2020; Yang et al., 2023). Unlike those methods, LPG blocks guide features to the full resolution leveraging the local planar assumption (Lee et al., 2021). The LPG block consists of a stack of 1×1 reduction layers, which iteratively decrease the number of channels by a factor of two until it reaches a channel dimension of three. The resulting feature map $(H/k, H/k, 3)$ is processed through two pathways to compute local plane coefficient estimations. The first pathway uses the first two channels (FC#1 and FC#2) to compute the (x, y, z) components of unit normal vectors, denoted by (n_x, n_y, n_z). A unit normal vector has only two degrees of freedom (DoF), described in spherical coordinates by polar (θ) and azimuthal (ϕ) angles from predefined axes. The two channels (FC#1 and FC#2) are regarded as polar (θ) and azimuthal (ϕ) angle maps, respectively, and converted to unit normal vectors in Cartesian coordinates by Eq. (1).

$$\begin{cases} n_x = sin(\theta)cos(\phi) \\ n_y = sin(\theta)sin(\phi) \\ n_z = cos(\theta) \end{cases} \tag{1}$$

The second pathway estimates the perpendicular distance (n_d) between the plane and the origin. This pathway computes a sigmoid function from the third channel and multiplies its output with the maximum depth value. Finally, these 4D plane coefficients (n_x, n_y, n_z, n_d) are used to compute $k \times k$ local depth cues using the ray-plane intersection Eq. (2).

$$\tilde{c}^{k\times k} = \frac{n_d}{n_x \cdot u_i + n_y \cdot v_i + n_z} \tag{2}$$

where (n_x, n_y, n_z, n_d) describe the previously estimated plane coefficients and u_i, v_i denote $k \times k$ patch-wise normalised coordinates of pixel i.

3.2.2. Attention mechanism

Convolutional neural networks have shown outstanding performance in enhancing local feature representations. However, focusing on relevant features while suppressing irrelevant ones further improves the process of capturing the visual structure of a scene. The convolutional block attention module (CBAM) proposed in Woo et al. (2018) sequential applies a channel attention module and a spatial attention

module to a given feature map. The process consists of two sequential element-wise multiplications. The first one multiplies the input feature map ($\mathcal{F}$) and the channel attention map (M_c). The second one is performed between the output of the first multiplication ($\mathcal{F}'$) and the output of the spatial attention module (M_s), resulting in a refined feature map $\mathcal{F}''$. The whole process can be summarised as shown in Eq. (3):

$$\begin{cases} \mathcal{F}' = \mathcal{F} \otimes M_c(\mathcal{F}) \\ \mathcal{F}'' = \mathcal{F}' \otimes M_s(\mathcal{F}') \end{cases} \tag{3}$$

This module has been used by Li et al. (2023a) in the skip connections and in the decoder of their network to recover meaningful global information at a low computational cost. We followed a similar approach to Li et al. (2023a) by incorporating CBAM modules in the skip connections and at each resolution level of our depth decoder. We aim to leverage the local feature representation of convolutional neural networks to extract monocular depth cues and the global context awareness of CBAM modules to relate them effectively. The incorporation of CBAM modules into our framework is shown in Fig. 2.

3.2.3. Multi-task learning network with UNC block

Our proposed framework is based on the geometric relationship between the depth and surface normal information of a 3D scene. Following this statement, our method aims to improve its depth estimation robustness by incorporating a geometrically related auxiliary task. The proposed architecture consists of a single shared encoder and two independent decoders. The purpose of the shared encoder is to extract meaningful geometric features that represent the 3D scene, while the two decoders are used to regress depth and surface normal maps.

Our framework extends the depth estimation network by adding a surface normal decoder located at the top of Fig. 2. Within our surface normal decoder, we introduce the unit normal computation block (UNC block) to compute surface normal maps at multiple resolutions. Our UNC block is based on the LPG blocks without the computation of the ray-plane intersection. Each surface normal map is represented by an RGB image, where each channel represents one particular axis: the red channel denotes the x-axis, the green channel represents the y-axis, and the blue channel represents the z-axis. The outputs of all UNC blocks undergo a channel concatenation followed by convolutional layers to compute the final surface normal prediction.

The multi-task learning framework combines the losses of both tasks as described in Eq. (4).

$$\mathcal{L}_{MTL} = \lambda_1 \cdot \mathcal{L}_{depth} + \lambda_2 \cdot \mathcal{L}_{sn} \tag{4}$$

where λ_1, λ_2 are weighting factors equally set to 0.5, $\mathcal{L}_{depth}$ represents the scale-invariant logarithmic error (SILog) loss between the predicted and ground truth depth maps (Eigen et al., 2014), and $\mathcal{L}_{sn}$ symbolises the mean absolute error (MAE) loss between the estimated surface normal and its corresponding ground truth.

3.2.4. Cross-task consistency loss with D2SN module

We incorporate a cross-task consistency (X-TC) loss into our multi-task learning framework to enforce consistency among depth and surface normal predictions. To this end, we add a depth-to-surface normal (D2SN) warping module based on the mathematical method proposed by Nakagawa et al. (2015). Our warping module uses the output of our depth estimation decoder. The predicted depth map is processed within this module to generate a warped surface normal representation. The output of this module introduces a consistency constraint by being compared against the prediction of the surface normal decoder.

Depth-to-surface normal module (D2SN module). Since normal estimation is equivalent to fitting a plane to a local point cloud in the 3D space, several approaches leveraging optimisation techniques have been proposed. However, these methods are expensive in terms of computational resources (Nakagawa et al., 2015). Therefore, our D2SN module, which is shown in the right section of Fig. 2, computes a surface normal from depth image gradients (DIG) as shown in Nakagawa et al. (2015). The authors consider that adjacent 3D points in a depth image can be used to compute a local 3D plane whose orthogonal vector is equivalent to the normal vector at that particular pixel. Their proposed method consists of three steps:

1. Depth image gradients: Given the location of a pixel (x, y) and its depth value (Z), pixels can be projected to the 3D space $\mathrm{P}(X, Y, Z)$ by a transformation matrix with known camera intrinsic parameters. The generated point cloud is used to compute the partial directional derivatives as shown in Eq. (5):

$$\begin{cases} \frac{\partial Z(x,y)}{\partial x} = Z(x+1,y) - Z(x,y) \\ \frac{\partial Z(x,y)}{\partial y} = Z(x,y+1) - Z(x,y) \\ \frac{\partial X(x,y)}{\partial x} = \frac{Z(x,y)}{f} + \frac{(x-c_x)}{f}\frac{\partial Z(x,y)}{\partial x} \\ \frac{\partial X(x,y)}{\partial y} = \frac{(x-c_x)}{f}\frac{\partial Z(x,y)}{\partial y} \\ \frac{\partial Y(x,y)}{\partial x} = \frac{(x-c_y)}{f}\frac{\partial Z(x,y)}{\partial y} \\ \frac{\partial Y(x,y)}{\partial y} = \frac{Z(x,y)}{f} + \frac{(y-c_y)}{f}\frac{\partial Z(x,y)}{\partial y} \end{cases} \tag{5}$$

2. Tangent vectors: The x and y directional derivatives at a given 3D point, P, can be used as tangent vectors of the surface as shown in Eq. (6):

$$\begin{cases} \mathbf{v}_x(x,y) = \left(\frac{\partial X(x,y)}{\partial x}, \frac{\partial Y(x,y)}{\partial x}, \frac{\partial Z(x,y)}{\partial x} \right) \\ \mathbf{v}_y(x,y) = \left(\frac{\partial X(x,y)}{\partial y}, \frac{\partial Y(x,y)}{\partial y}, \frac{\partial Z(x,y)}{\partial y} \right) \end{cases} \tag{6}$$

3. Normal vector: The cross-product of the two tangent vectors described in the previous step is calculated to get the normal vector as in Eq. (7):

$$\mathbf{n}(x,y) = \mathbf{v}_x(x,y) \times \mathbf{v}_y(x,y) \tag{7}$$

While a mathematical formulation to map surface normals directly to absolute depth maps is non-trivial due to the inherent ambiguity in the relationship between surface orientation and depth. Some works that use surface normals to refine depth maps apply an iterative post-processing method requiring both the initial depth prediction and surface normal maps (Bae et al., 2022; Patil et al., 2022; Shao et al., 2023). However, even doing so a single surface normal can correspond to multiple absolute depth configurations creating ambiguity in the depth prediction. Thus, in contrast to these approaches, we aim to enhance the feature extraction and decoding processes in our network through cross-consistency loss utilising a mathematically feasible depth-to-surface normal configuration, which can also be trained in an end-to-end fashion.

Our multi-task learning with cross-task consistency network combines the losses from both tasks and the cross-task consistency loss into the final weighted loss function described in Eq. (8).

$$\mathcal{L}_{final} = \lambda_1 \cdot \mathcal{L}_{depth} + \lambda_2 \cdot \mathcal{L}_{sn} + \lambda_3 \cdot \mathcal{L}_{x-tc} \tag{8}$$

where λ_1, λ_2, and λ_3 are weighting factors, and $\mathcal{L}_{X-TC}$ represents the root mean squared error (RMSE) loss between the predicted and warped surface normal maps. Eq. (9) defines each of the loss functions used in our final loss.

$$\begin{cases} L_{depth} = \alpha\sqrt{D(g)};\ D(g) = \frac{1}{N}\sum_i g_i^2 - \frac{\lambda}{N^2}\left(\sum_i g_i\right)^2 \\ L_{sn} = \frac{1}{N}\sum_i |y_{i\_sn} - \hat{y}_{i\_sn}| \\ L_{x-tc} = \sqrt{\frac{1}{N}\sum_i (y^*_{i\_sn} - \hat{y}_{i\_sn})^2} \end{cases} \tag{9}$$

where $\alpha = 10$ is a scaling parameter of the range of the loss function to improve convergence (Lee et al., 2021) and $\lambda = 0.85$ is a trade-off parameter between element-wise l_2 error ($\lambda = 0$) and the exact scale invariant error ($\lambda = 1$) (Eigen et al., 2014). The variable y_i represents the ground truth map, $\hat{y}_i$ denotes the prediction map, and $y^*_{i\_sn}$ the warped surface normal map. The variable g_i denotes the difference between the predicted and the ground truth depths in logarithmic scale for each sample i, i.e., $g_i = log\left(y_{i\_depth}\right) - log\left(\hat{y}_{i\_depth}\right)$.

3.3. Self-supervised learning using masked image modelling

To improve the generalisability of our network to real colonoscopy patient data, we leverage A^{2}MIM to pre-train our encoder on patient and 3D phantom model data. A^{2}MIM computes the mean RGB value of the input image and uses it to perform patch masking (see Fig. 2, bottom right). While most existing mask image modelling (MIM) frameworks apply the mask token in the input space, following this approach can affect the context extraction capabilities of CNNs. Additionally, this technique limits the feature extraction capabilities to local texture features learned by low-order interactions among patches. Therefore, a learnable mask token is added on intermediate feature maps of the encoder architecture, where semantic and spatial features are available. Furthermore, a Fourier loss is incorporated to enable CNNs to model features of medium frequencies (middle-order interactions) for more generalised feature extraction (Li et al., 2023b). To this extent, we create a combination of colonoscopy data from the publicly available PolypGen (Ali et al., 2023) (clinical colonoscopy videos from 6 different medical centres acquired) and C3VD (Bobrow et al., 2023) (see Section 3.1) datasets.

4. Experiments and results

4.1. Training setup

Each model presented in this study is trained using a single NVIDIA V100 GPU. All models are trained for up to 50 epochs (pix2pix and MonoDepth+FPN are trained for 200 epochs) using a batch size of 8, with an initial learning rate of $1e^{-4}$, and a weight decay of $1e^{-2}$. The input images are resized from their original resolution to 320×320 pixels. Only random rotation is performed as a data augmentation technique to avoid the loss of structural information and visual cues. Random cropping, which is suggested by baseline methods, is discarded because we observed that it could lead to heavy close-ups towards the walls of the colon, drastically reducing contextual information.

4.2. Ablation study setup

We perform an ablation study to analyse the proposed network components that lead to the design of our multi-task learning approach with cross-task consistency featuring attention mechanisms. We select the BTS architecture proposed by Lee et al. (2021) as our baseline method. Before incorporating our multi-task learning approach, we add CBAM attention modules at different stages of the depth decoder and its skip connections. We further extend this method following a multi-task learning scheme. Experimentally, we define the best loss function to optimise our auxiliary task. Additionally, we implement a cross-task consistency module into our multi-task learning approach. Finally, we conduct a hyperparameter study on our multi-task learning with a cross-task consistency framework to determine the set of λ values in our loss function that leads to the best trade-off performance among both tasks. In Table 2, we describe the different network configurations trained and evaluated in our ablation study.

Table 2
Model configurations. Our ablation study setup is constituted by four different model configurations.

Model ID	CBAM	MTL	X-TC
BTS (baseline) (Lee et al., 2021)			
BTS-CBAM	✓		
BTS-CBAM-MTL	✓	✓	
BTS-CBAM-MTL-X-TC (Col3D-MTL)	✓	✓	✓

4.3. Metrics and assessment

To evaluate our methods, we follow standard depth estimation metrics described by Eigen et al. (2014). These include five error metrics: absolute relative error (Abs Rel), squared relative error (Sq Rel), logarithmic error (log10), root mean squared error (RMSE), root mean squared logarithmic error (RMSE_{log}), scale-invariant logarithmic error (SILog); and three accuracy metrics that are described in Eq. (10) : $\delta_{1.25}$, $\delta_{1.25^2}$, and $\delta_{1.25^3}$. We used standard surface normal evaluation metrics, which include two error metrics: average angular error (AAE) and median angular error (Med. AE), and three accuracy metrics: $\delta_{11.25^\circ}$, $\delta_{22.5^\circ}$, and δ_{30°.

$$\%\ \text{of}\ \hat{y}_i\ \text{s.t.}\ \max\left(\frac{y_i}{\hat{y}_i}, \frac{\hat{y}_i}{y_i}\right) = \delta_{threshold} < threshold \tag{10}$$

4.4. Results

In this section, we provide the quantitative and qualitative results from our network configuration ablation study, followed by our hyperparameter study on the λ weighting factors of our final loss function. Finally, we compare our approach against other state-of-the-art methods.

4.4.1. Quantitative results

In Table 3, we provide the results of our ablation study on the proposed network configurations. We compare the performances of BTS, BTS-CBAM, BTS-CBAM-MTL, and BTS-CBAM-MTL-X-TC for all depth and surface normal evaluation metrics on the validation set of the C3VD dataset. Our BTS-CBAM network yields a relative improvement over the BTS method by 8.9% and 2.4% in terms of SILog and $\delta_{1.25}$, respectively.

Following the incorporation of attention mechanisms, the middle section of Table 3 illustrates the effect of the L_1 and L_2 loss functions to optimise the surface normal decoder of our BTS-CBAM-MTL approach. The use of the L_1 loss function leads to a relative improvement of 2.6% on SILog metric but a relative decrease of 1.5% on $\delta_{1.25}$ regarding the use of the L_2 loss. Furthermore, our surface normal decoder optimised through the L_1 loss demonstrates a relative improvement of 1% and 26.5% on mean angular error and $\delta_{11.25^\circ}$, respectively. The bottom section of Table 3 assesses the performance of our BTS-CBAM-MTL-X-TC employing the L_1 and L_2 loss functions to optimise our surface normal decoder. Initially, the λ weighting factors in our final loss function (Eq. (8)) are set to $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.2$. Our proposed network configuration optimising our surface normal decoder with the L_1 loss outperforms all previous network configurations in all depth and surface normal evaluation metrics.

Table 4 includes the results of our hyperparameter study on different sets of λ weighting factors in the loss function (Eq. (8)) of our BTS-CBAM-MTL-X-TC network. Based on our experimental results, the best λ configuration set consists of $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.2$. This configuration leads to the best performance among all evaluated models, achieving a relative improvement of 10.9%, 11.9%, and 7% on SILog, RMSE, and $\delta_{1.25}$ metrics over our baseline method. Furthermore, we outperform the surface normal prediction of our BTS-CBAM-MTL configuration by a relative improvement of 45.3%, 39.5%, and 37.4% on mean angular error, median angular error, and $\delta_{11.25^\circ}$.

Table 3
Quantitative results for various network configurations on validation set (ablation study). Each network uses the learned weights that lead to a better performance during the training stage. **First** and second best performing methods for each evaluation metric are formatted.

Method	Losses	Depth (in mm)									Surface normals (in *degrees*)				
		Abs Rel ↓	Sq Rel ↓	log 10 ↓	RMSE ↓	$RMSE_{log}$ ↓	SILog ↓	$\delta_{1.25}$ ↑	$\delta_{1.25^2}$ ↑	$\delta_{1.25^3}$ ↑	AAE ↓	Med. AE ↓	$\delta_{11.25^\circ}$ ↑	$\delta_{22.5^\circ}$ ↑	δ_{30° ↑
BTS (baseline) (Lee et al., 2021)	Depth: SILog	0.179	1.088	0.073	5.667	0.208	14.066	0.756	0.959	0.992	–	–	–	–	–
BTS-CBAM	Depth: SILog	0.170	0.979	0.070	5.503	0.198	12.814	0.774	0.969	0.994	–	–	–	–	–
BTS-CBAM-MTL	Depth: SILog SN: L_1	0.166	0.934	0.070	5.422	0.195	12.872	0.785	0.969	0.996	43.892	35.363	17.637	40.340	51.814
	Depth: SILog SN: L_2	0.163	0.900	0.067	5.298	0.190	13.219	0.797	0.972	0.995	44.337	35.142	13.940	37.288	49.594
	Depth: SILog SN: AAE	0.164	0.946	0.072	5.426	0.205	14.095	0.755	0.970	0.995	36.362	27.732	**24.699**	52.089	62.485
BTS-CBAM-MTL-X-TC (Ours: Col3D-MTL)	Depth: SILog SN: L_1 X-TC: L_2	**0.156**	**0.805**	**0.065**	**4.994**	**0.186**	**12.530**	**0.809**	**0.975**	**0.997**	**23.999**	**21.382**	24.232	**55.883**	**71.758**
	Depth: SILog SN: L_2 X-TC: L_2	0.176	1.057	0.073	5.723	0.211	14.822	0.731	0.968	0.996	40.664	32.630	15.974	40.219	52.394
	Depth: SILog SN: AAE X-TC: AAE	0.184	1.182	0.074	5.902	0.216	16.560	0.735	0.957	0.992	39.831	32.366	16.008	39.360	52.721

Table 4
Hyperparameter study on validation set. Proposed BTS-CBAM-MTL-X-TC networks are evaluated in the validation set to avoid λ weighting factors to be adjusted based on the testing data. Each network uses the learned weights that lead to a better performance during the training stage. **First** and second best performing methods for each evaluation metric are formatted.

Method	λ_1	λ_2	λ_3	Depth (in mm)									Surface normals (in *degrees*)				
				Abs Rel ↓	Sq Rel ↓	log 10 ↓	RMSE ↓	$RMSE_{log}$ ↓	SILog ↓	$\delta_{1.25}$ ↑	$\delta_{1.25^2}$ ↑	$\delta_{1.25^3}$ ↑	AAE ↓	Med. AE ↓	$\delta_{11.25^\circ}$ ↑	$\delta_{22.5^\circ}$ ↑	δ_{30° ↑
BTS-CBAM-MTL-X-TC (Col3D-MTL)	0.4	0.3	0.3	0.178	1.061	0.072	5.639	0.204	13.632	0.769	0.963	0.994	**22.957**	**20.405**	**25.028**	**57.662**	**73.668**
	0.4	0.2	0.2	0.187	1.116	0.077	5.664	0.213	13.672	0.726	0.965	0.990	30.100	24.999	21.081	49.877	63.892
	0.5	0.3	0.2	**0.156**	**0.805**	**0.065**	4.994	**0.186**	**12.530**	**0.809**	**0.975**	**0.997**	23.999	21.382	24.232	55.883	71.758
	0.5	0.4	0.1	0.161	0.837	0.066	**4.901**	0.189	13.036	0.795	0.974	**0.997**	34.013	24.606	21.812	50.571	63.335
	0.6	0.2	0.2	0.222	1.470	0.089	6.269	0.248	16.163	0.620	0.939	0.987	29.338	25.449	17.860	45.908	62.025
	0.6	0.3	0.1	0.187	1.083	0.075	5.598	0.209	13.773	0.742	0.957	0.994	45.448	35.227	15.080	36.828	48.007
	0.7	0.2	0.1	0.178	1.028	0.070	5.128	0.202	14.329	0.764	0.961	0.991	33.879	28.175	17.835	43.982	58.190
	0.8	0.1	0.1	0.180	1.078	0.074	5.753	0.210	13.573	0.739	0.963	0.995	23.880	21.434	23.147	55.364	71.800

Table 5
Benchmark results on test set. Evaluation of three state-of-the-art methods, our baseline method, our proposed framework, and our proposed framework with SSL pre-training on the C3VD dataset. All methods are trained and evaluated on the same data distributions. **First** and second best performing methods for each evaluation metric are formatted.

Method	# parameters	Depth (in mm)								
		Abs Rel ↓	Sq Rel ↓	log 10 ↓	RMSE ↓	$RMSE_{log}$ ↓	SILog ↓	$\delta_{1.25}$ ↑	$\delta_{1.25^2}$ ↑	$\delta_{1.25^3}$ ↑
pix2pix (Isola et al., 2017)	14.2 M	0.157 ± 0.103	0.730 ± 0.647	0.062 ± 0.034	3.721 ± 1.385	0.237 ± 0.144	21.062 ± 14.118	0.801 ± 0.181	0.956 ± 0.078	0.986 ± 0.029
MonoDepth+FPN (Ali et al., 2021)	61.8 M	0.204 ± **0.052**	1.494 ± 0.832	0.090 ± 0.025	6.762 ± 1.812	0.342 ± 0.078	31.737 ± 6.25	0.628 ± 0.176	0.949 ± 0.046	0.991 ± 0.010
NeWCRFs (Chen et al., 2020a)	270.4 M	0.133 ± 0.097	0.557 ± 0.595	0.048 ± 0.033	3.058 ± 1.419	0.151 ± 0.082	11.817 ± 4.714	0.854 ± 0.197	0.973 ± 0.066	0.997 ± 0.006
NDDepth (Shao et al., 2023)	348.4 M	0.316 ± 0.076	2.651 ± 1.691	0.111 ± 0.027	7.043 ± 3.317	0.322 ± **0.043**	24.255 ± 6.497	0.498 ± 0.218	0.873 ± 0.031	0.956 ± 0.018
BTS (Lee et al., 2021)	50.3 M	0.127 ± 0.089	0.622 ± 0.543	0.055 ± 0.034	3.823 ± 2.024	0.150 ± 0.089	11.400 ± 5.614	0.812 ± 0.198	0.979 ± 0.051	0.998 ± 0.003
Col3D-MTL (ours)	50.3 M	0.109 ± 0.064	0.386 ± **0.316**	0.046 ± **0.024**	3.052 ± 1.143	0.131 ± 0.065	11.035 ± 4.856	0.896 ± **0.140**	**0.989** ± **0.021**	0.998 ± 0.003
Col3D-MTL + SSL (ours + SSL)	50.3 M	**0.107** ± 0.075	**0.346** ± 0.415	**0.042** ± 0.026	**2.729** ± **1.036**	**0.128** ± 0.062	**10.072** ± **3.164**	**0.899** ± 0.174	**0.989** ± 0.026	**0.999** ± **0.001**

In Table 5, we compare the performance of our baseline and our Col3D-MTL methods against other state-of-the-art networks on supervised monocular depth estimation: NeWCRFS (Chen et al., 2020a), pix2pix (Isola et al., 2017), MonoDepth+FPN (Ali et al., 2021), and NDDepth (Shao et al., 2023). MonoDepth+FPN (Ali et al., 2021) is used to estimate depth maps on oesophageal endoscopy, while pix2pix (Isola et al., 2017) is applied to a synthetic colonoscopy dataset to generate depth predictions (Rau et al., 2019). All methods are trained on the same data distribution and tested on the same held-out testing data, setting a new benchmark on the C3VD dataset. Our proposed frameworks (with and without SSL pre-training, respectively) achieve the first and second-best performances on all depth evaluation metrics. Our Col3D-MTL + SSL improves our baseline method by 11.7%, 28.6%, and 10.7% on SILog, RMSE and $\delta_{1.25}$, respectively. Moreover, it outperforms the state-of-the-art method NeWCRFs by 14.8%, 10.8%, and 5.3% on the same evaluation metrics.

Table 6 analyses the performance of all the evaluated methods on each colon segment of the testing set separately. On the caecum and transverse segments, which respectively represent 43% and 36% of our training data, our baseline and Col3D-MTL achieve the first and second best-performing methods in terms of SILog, RMSE, and $\delta_{1.25}$. Only NeWCRFs accomplish a lower RMSE value on the caecum segment, achieving a relative improvement of 4.9% over our baseline.

Table 6
Quantitative results for per colon segment. Evaluation of all evaluated methods on each colon segment in the testing set is provided. **First** and second best performing method for each evaluation metric on each colon segment is formatted. In '(.)' we include the percentage of training samples used from each segment.

Colon segment	Method	Depth (in mm)		
		SILog ↓	RMSE ↓	$\delta_{1.25}$ ↑
Caecum (43%)	pix2pix(Isola et al., 2017)	6.805 ± 1.345	2.649 ± 1.342	0.956 ± 0.083
	MonoDepth+FPN(Ali et al., 2021)	28.834 ± 4.958	7.143 ± 1.405	0.822 ± 0.052
	NeWCRFs(Chen et al., 2020a)	7.494 ± 0.696	**1.567** ± 0.552	0.996 ± 0.004
	NDDepth(Shao et al., 2023)	17.027 ± 1.111	11.548 ± 1.946	0.342 ± 0.161
	BTS(Lee et al., 2021)	**5.171** ± 1.065	1.649 ± 0.447	**0.997** ± 0.003
	Ours	5.738 ± 1.232	2.069 ± 0.547	0.996 ± 0.004
	Ours + SSL	6.749 ± **0.557**	1.894 ± **0.407**	0.996 ± **0.002**
Transverse (36%)	pix2pix(Isola et al., 2017)	13.464 ± 1.304	2.374 ± 0.100	0.929 ± 0.017
	MonoDepth+FPN(Ali et al., 2021)	25.409 ± 0.695	4.894 ± 0.279	0.759 ± 0.026
	NeWCRFs(Chen et al., 2020a)	8.858 ± 0.917	2.598 ± 0.416	0.992 ± 0.004
	NDDepth(Shao et al., 2023)	21.466 ± 0.422	3.677 ± 0.123	0.772 ± 0.007
	BTS(Lee et al., 2021)	**6.243** ± **0.265**	1.840 ± 0.145	0.997 ± 0.002
	Ours	6.412 ± 0.490	**1.648** ± **0.056**	**0.997** ± **0.001**
	Ours + SSL	7.164 ± 0.466	2.927 ± 0.143	0.990 ± 0.002
Sigmoid (21%)	pix2pix(Isola et al., 2017)	32.429 ± 12.481	4.618 ± 0.593	0.745 ± 0.068
	MonoDepth+FPN(Ali et al., 2021)	36.365 ± 4.780	6.422 ± 0.546	0.459 ± 0.072
	NeWCRFs(Chen et al., 2020a)	13.135 ± 1.600	3.664 ± 1.052	0.809 ± 0.158
	NDDepth(Shao et al., 2023)	30.938 ± 2.369	5.654 ± 1.352	0.493 ± 0.191
	BTS(Lee et al., 2021)	16.276 ± 1.526	5.838 ± 0.498	0.715 ± 0.048
	Ours	15.460 ± 1.240	4.020 ± 0.353	0.877 ± **0.023**
	Ours + SSL	**12.515** ± **1.106**	**2.602** ± **0.351**	**0.917** ± 0.029
Descending (0%)	pix2pix(Isola et al., 2017)	26.52 ± 6.868	5.095 ± 0.599	0.432 ± 0.121
	MonoDepth+FPN(Ali et al., 2021)	31.629 ± 6.907	11.405 ± 1.479	0.523 ± 0.109
	NeWCRFs(Chen et al., 2020a)	22.137 ± 2.915	5.468 ± **0.437**	0.464 ± **0.098**
	NDDepth(Shao et al., 2023)	22.156 ± 5.466	5.675 ± 1.482	0.511 ± 0.201
	BTS(Lee et al., 2021)	15.647 ± 2.705	5.121 ± 0.627	0.396 ± 0.138
	Ours	14.156 ± 2.325	**4.279** ± 0.584	**0.524** ± 0.144
	Ours + SSL	**12.201** ± **0.772**	5.611 ± 0.885	0.399 ± 0.166

Considering the sigmoid segment, which only constitutes 21% of the training set, Col3D-MTL + SSL is the best-performing method. Col3D-MTL achieves the second best $\delta_{1.25}$ and improves our baseline method by 5%, 31.1%, and 22.7% on SILog, RMSE, and $\delta_{1.25}$, respectively. The most remarkable improvement occurs in the descending segment of the colon, which is not given in the training stage. On this colon segment, Col3D-MTL ranks first among all the evaluated methods on RMSE and $\delta_{1.25}$. Col3D-MTL + SSL achieves the best SILog evaluation metric among all the evaluated methods. Moreover, it yields a relative improvement of 22% and 0.8% on SILog and $\delta_{1.25}$ over our baseline.

Table 6 also compares the standard deviation of all evaluated methods on each colon segment of the testing set. Col3D-MTL + SSL achieves the lowest standard deviation on all evaluation metrics on the caecum. Additionally, on the sigmoid segment, it achieves the best standard deviation on SILog and RMSE, only surpassed by Col3D-MTL on $\delta_{1.25}$ by 26%. On the transverse segment, Col3D-MTL yields the lowest RMSE and $\delta_{1.25}$ standard deviations, outperforming our baseline by 61.4% and 50%, respectively. Considering the descending colon segment, Col3D-MTL + SSL is the best-performing method in terms of SILog, surpassing the second-best-performing method (Col3D-MTL) by 66.8%. The worst performance of our methods is observed on the $\delta_{1.25}$ standard deviation on the descending colon segment, in which they only outperform NDDepth by a relative improvement of 28.4% (Col3D-MTL) and 17.4% (Col3D-MTL + SSL).

4.4.2. Qualitative results

Fig. 3 contains sample input images with their corresponding ground truth annotations and the predictions of our baseline and our proposed method. We can observe that our baseline properly recovers a global depth map of the 3D scene; however, it has a tendency to generate smooth transitions between anatomical structures. Our proposed method addresses these cases by recovering geometrical information about the scene and enforcing cross-task consistency between our depth and surface normal predictions, leading to sharper boundaries and reduced visual artefacts, e.g., specular reflection. We include an absolute error map for each channel pair among prediction and ground truth maps from both tasks. From the absolute error maps, we can observe that our baseline generates brighter regions than Col3D-MTL, i.e., our approach recovers the 3D information of the scene with less absolute error than its baseline method.

By analysing the most challenging samples, we can observe that the cases in which our methods fail to recover an accurate depth estimation are small regions with low lighting conditions. Even though our surface normal decoder recovers the overall geometry of small structures, such as polyps, it does not compute a detailed representation of the surface orientation within these small regions. Looking at the areas denoted with red arrows, we can observe that regions with higher errors in our depth maps are consistent with regions with higher errors in our surface normal maps. Despite these focalised errors, the overall depth estimation of our proposed method performs better than the one from its baseline.

In Fig. 4, we compare the absolute error maps from all the proposed network configurations of our ablation study to observe the effect of each approach. Lower absolute error maps are achieved after each incorporated module, emphasising their positive impact towards an enhanced depth estimation. Our baseline method inaccurately estimates the depth values at transition zones, e.g., on the folds of the colon and occluded regions. The addition of CBAM modules reduces the absolute depth error at areas corresponding to the folds of the colon but not at occlusion zones where depth variation can be higher. Incorporating our surface normal predictor diminishes the absolute error of our depth estimation at these zones. However, the highest absolute errors within each depth map remain in these areas. Our BTS-CBAM-MTL-X-TC approach addresses these cases by explicitly enforcing consistency among both predictions. Our proposed framework leads to an overall lower absolute error map but also achieves a sharper depth prediction with fewer visual artefacts.

Fig. 5 shows a qualitative comparison between all the state-of-the-art methods evaluated in this study: BTS, pix2pix, NeWCRFs, MonoDepth+FPN, NDDepth, and our Col3D-MTL frameworks. From the

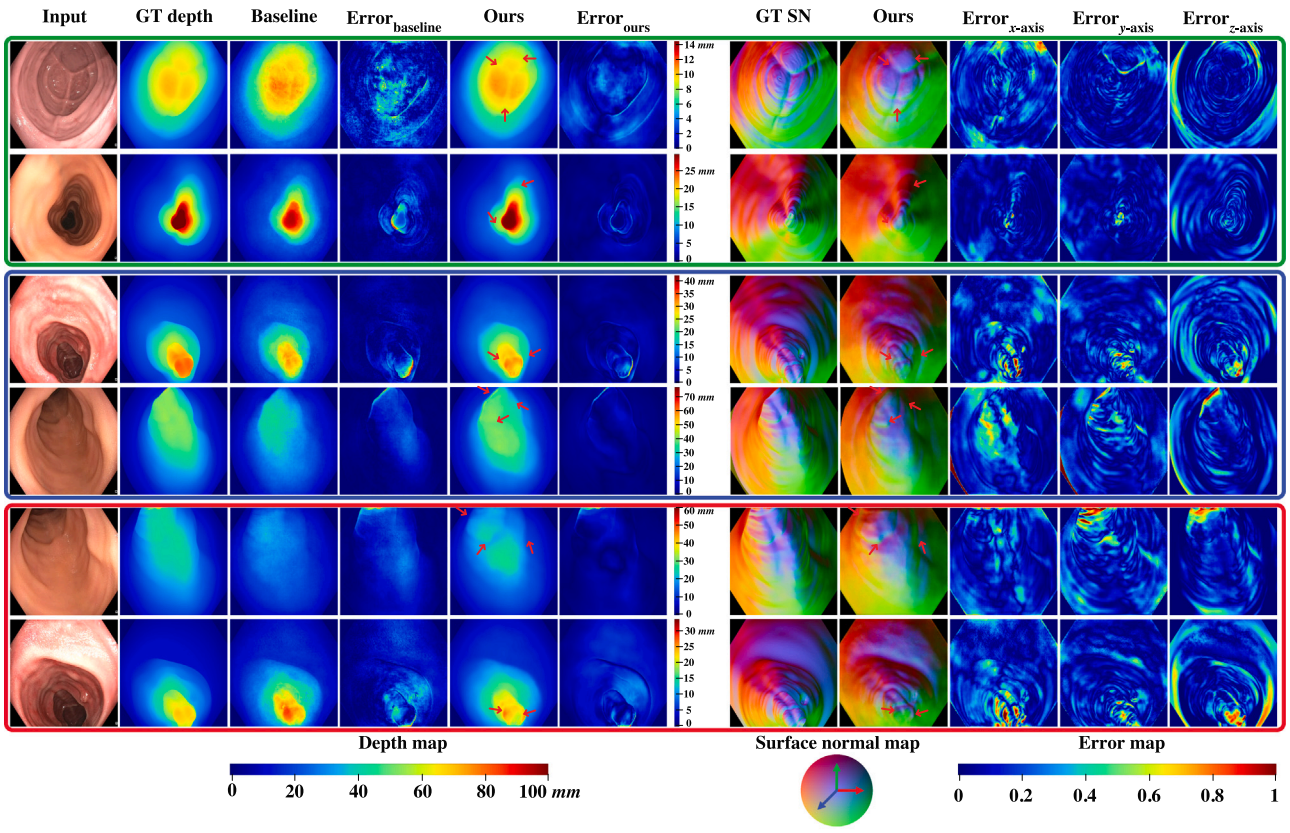

Fig. 3. Qualitative comparison between our baseline and our proposed framework on best, average, and worst performing cases. We show absolute error maps for both methods to observe the most challenging regions and the impact of our BTS-CBAM-MTL-X-TC. In the first two rows, we can observe the best-performing cases (green) in which both methods lead to low absolute error maps. The third and fourth rows show average-performing cases (blue) in which the lack of texture and regions with high-depth variability affecting our baseline method are addressed by our BTS-CBAM-MTL-X-TC framework. Consistency can be observed on the predicted surface normal maps, which help to recover the shape of the scene, e.g., folds of the colon and small protuberances like polyps. The last two rows represent challenging cases (red) in which both methods generate less accurate depth estimations. Low-lighting regions, usually located at the furthest section of the scene, represent challenging cases for both methods.

testing set, one sample from each colon segment with its corresponding ground truth and depth prediction is given. The highest absolute error maps are generated by NDDepth, followed by MonoDepth+FPN and pix2pix, which recovered an overall depth representation of the scene but not as sharp as our baseline method. Our baseline method improves pix2pix but generates smooth changes in transition zones. NeWCRFs surpass our baseline method in regions with low texture but cannot accurately estimate depth in regions with high depth variability. Our proposed frameworks show similar predictions, leading to the lowest absolute error maps among all the previous methods, addressing low texture and regions with increased depth variability. Overall, Col3D-MTL + SSL leads to lower absolute error maps, except on the last sample corresponding to the transverse segment.

In Fig. 6, we can visualise the 3D projection of each colon segment from the depth predictions of each state-of-the-art method evaluated in this work. Each 3D projection is computed from 50 depth map predictions following the ground truth camera poses provided by the C3VD dataset. Overall, the 3D projections computed from the depth estimation maps of NDDepth and pix2pix contain more artefacts and distortions than the ones from the other methods (first and fourth row of Fig. 6). Another significant drawback of these methods is their inability to recover the overall shape of the colon segments. For example, the smoothed high-depth values predicted by MonoDepth+FPN on the caecum segment (first row of Fig. 5) lead to a protruded 3D projection in which the folds of the colon are not recovered properly (first row of Fig. 6). Additionally, the polyp projected from the depth map predictions of MonoDepth+FPN does not recover the rounded shape of the polyp. NeWCRFs achieves a more realistic 3D model of the scene, considerably reducing the generation of artefacts and distortions. However, the predictions NeWCRFs and our baseline generate smaller and narrower projections of the polyp than the one generated from ground truth depth maps. Our Col3D-MTL networks overcome these drawbacks, generating an overall better representation of the shape of each colon segment with sharper boundaries at the folds of the gastrointestinal tract (first and fourth rows of Fig. 6). Col3D-MTL + SSL also generates a projection of the polyp whose shape and extent resemble the one computed from the ground truth depth maps (second row of Fig. 6).

Fig. 7 qualitatively compares our approach against our baseline method on real colonoscopy patient samples extracted from the CVC-ColonDB-300 (Bernal et al., 2012) and the PolypGen (Ali et al., 2023) datasets. The depth predictions on real colonoscopy frames demonstrate the enhanced generalisability of our proposed Col3D-MTL, even though our methods are trained on 3D phantom-based data. Col3D-MTL shows improved robustness against specularities and bubbles (see red arrows in the first and second samples from both datasets in Fig. 7) compared to our baseline. Additionally, it estimates smoother depth maps with sharper edges on the boundaries of polyps and gastrointestinal folds (see red arrows in the third sample from both datasets in Fig. 7) than our baseline method. Some limitations of this method are close, convex objects (e.g., big polyps in the fourth and fifth samples from PolypGen in Fig. 7), in which Col3D-MTL is able to estimate sharper boundaries but can generate higher depth values in its surface. To improve the generalisability of our method to unseen domains, we have compared three different techniques: CycleGAN (Zhu et al., 2017) for domain translation, domain gap reduction in endoscopy (Rau et al., 2023), and A^{2}MIM (Li et al., 2023b) as a pre-training task. CycleGAN and domain gap reduction in endoscopy are used to pre-process the

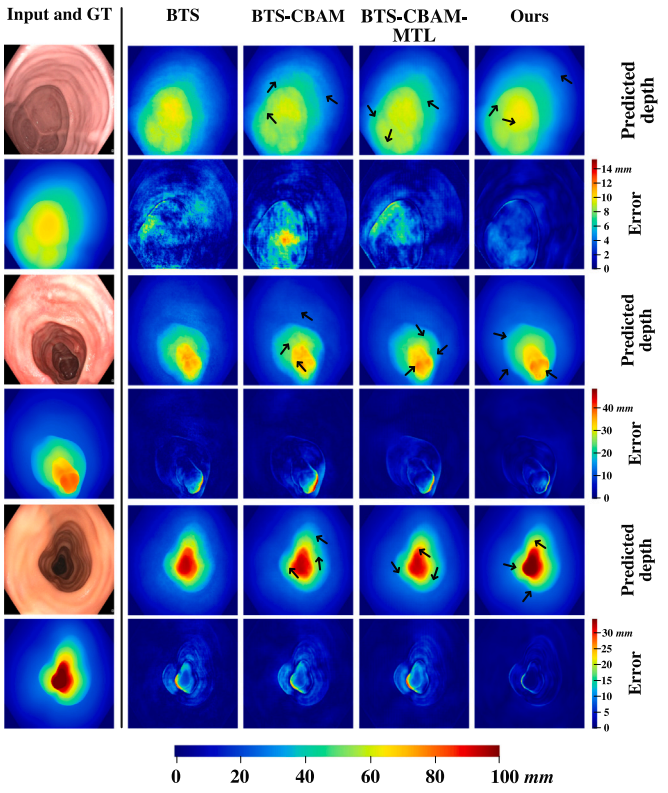

Fig. 4. Qualitative comparison between our different network configurations. We show the input image, its corresponding ground truth depth maps, the depth predictions for each network configuration, and its corresponding absolute error maps. White arrows show the positive impact of each network configuration concerning the previous one. The addition of the CBAM modules partially reduces the smooth transitions of our baseline method. Leveraging our MTL approach reduces the absolute error at transition zones but does not recover an accurate estimation at areas with low texture. Explicitly enforcing consistency among tasks achieves an enhanced depth estimation.

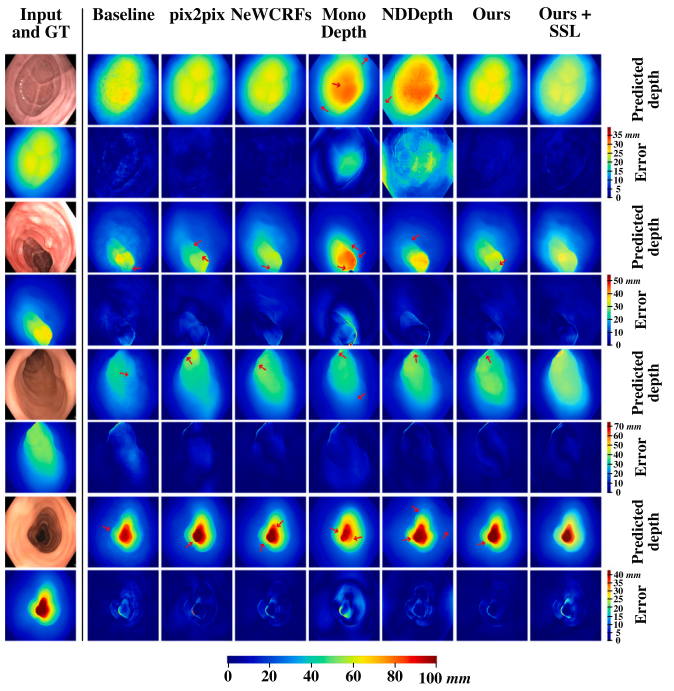

Fig. 5. Qualitative comparison between all evaluated methods in this study. One sample from each colon segment and its corresponding ground truth depth map is given from the testing set. The depth prediction of each method is provided with its corresponding absolute error map. Red arrows specify challenging regions for each method, e.g., folds of the colon, polyps, occlusion zones, low-texture regions, and areas with low lighting conditions.

input image (we will refer to them as Ours + cGAN and Ours + DGRE, respectively). A^{2}MIM is used to pre-train the feature extractor of our proposed network (we will refer to it as Col3D-MTL + SSL). While the translated images by CycleGAN seem to follow the distribution of the 3D phantom model dataset, it leads to flattened polyps and distorted shapes that affect depth estimation (see red arrows in the fourth column of Fig. 7). The use of domain gap reduction in endoscopy preserves the shape of the scene and emphasises useful features for depth prediction. This approach improves depth estimation on big polyps not given in the training set. However, it yields blurry images and generates image artefacts that affect depth prediction (red arrows in the sixth column of Fig. 7). The pre-training of our encoder using A^{2}MIM helps to mitigate the limitations of the previous methods. The predictions show sharper depth maps and improved generalisability to big polyps and unseen patient colonoscopy data.

5. Discussion

Monocular depth estimation methods rely on the representation of visual cues to recover the depth information from a single image. Convolutional neural networks have shown an outstanding performance in extracting local feature representations but need more contextual information to relate them properly. Therefore, we explored the use of CBAM modules at each multi-scale stage of the decoder and skip connections of our baseline method to incorporate attention mechanisms and leverage global context awareness. Our experiments show an improved performance on all evaluation metrics (see top of Table 3) and a refined depth map with respect to our baseline (second and third columns of Fig. 4). However, our BTS-CBAM configuration shows a tendency to create blurry regions. To further enhance the extraction of salient features and leverage the orientation of the scene, we integrate a geometrically related task, namely surface normal estimation, into our previous network with attention mechanisms. We have designed an independent surface normal decoder with novel unit normal computation blocks and incorporated it into our baseline method with CBAM modules. Our BTS-CBAM-MTL approach using an L_1 loss function to optimise the surface normal decoder yields the best trade-off performance considering all evaluation metrics for both tasks (see middle of Table 3). The extraction of salient geometrical features leads to a better scene representation, partially reducing the generation of blurry regions of our BTS-CBAM network (third and fourth columns of Fig. 4). However, without an explicit constraint to enforce cross-task consistency, we do not achieve a refined depth prediction at regions with high-depth variability, such as the folds of the gastrointestinal tract and occlusion zones. To explicitly enforce consistency among both predictions, we implement a cross-task consistency scheme. In order to evaluate consistency among depth and surface normal predictions, we implement a warping module based on DIG to generate a warping surface normal map from our depth estimation. Our cross-task consistency scheme aims to minimise the RMSE between the warped surface normal and the surface normal prediction of our decoder. We define our final loss function as a weighted sum of each loss function that optimises our depth estimation, surface normal prediction, and cross-task consistency module. The results of our ablation study show that the best set of λ weighting factors consists of $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.2$ (Table 4). Our proposed Col3D-MTL framework surpasses all previous approaches in our network configuration ablation study (see bottom of Table 3). We can observe that our approach leads to the most refined depth maps with the lowest absolute error maps (fifth column of Fig. 4). Moreover, we can notice an improved accuracy on transition zones, e.g., on the folds of the colon or in regions containing polyps. Furthermore, the

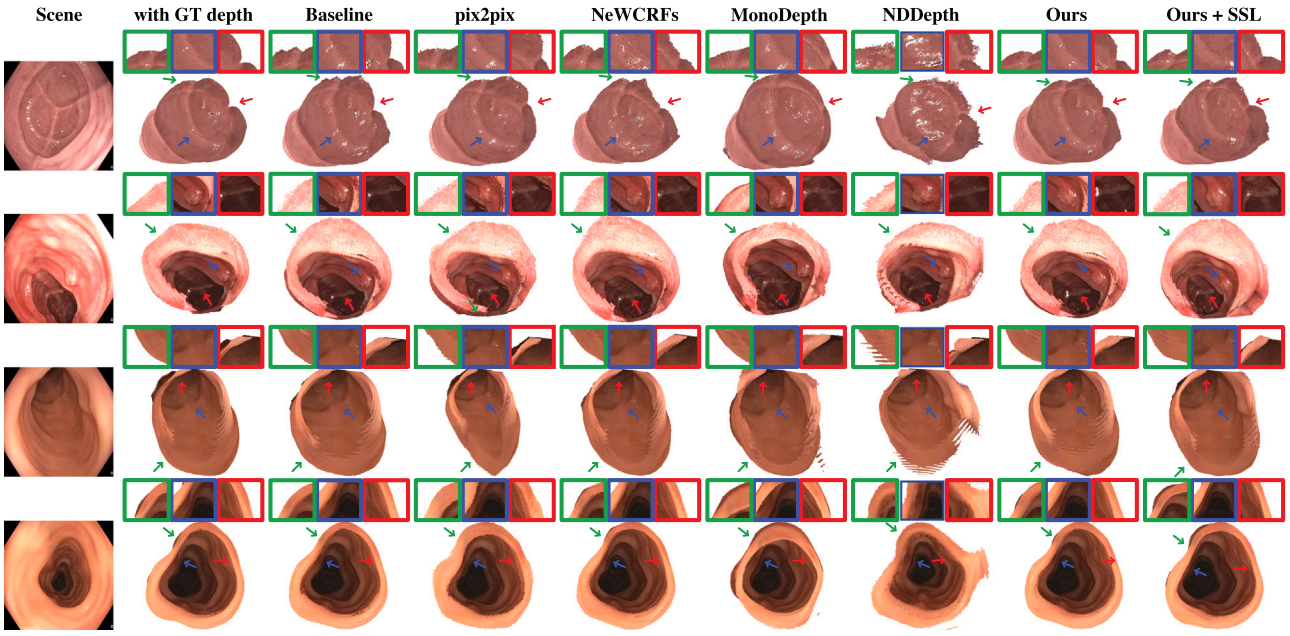

Fig. 6. Qualitative comparison between the 3D projections generated from evaluated methods for different colon segments. Each 3D projection consists of 50 depth map predictions projected into the world-coordinate space following the ground truth camera trajectory provided by the C3VD dataset. In the first row (caecum), our approaches achieve the sharpest boundaries at the folds of the gastrointestinal tract (red arrows). The second row (descending) shows that the shape of the polyp is better projected by our methods in comparison to the other analysed networks (blue arrows). In the third row (sigmoid), the overall geometry of the colon scene is less distorted by our frameworks (green arrows). The fourth row (transverse) demonstrates that our networks recover inner structures with more realistic details while maintaining the overall shape of the scene and reducing the presence of artefacts (green, blue, and red arrows). Visualisations performed with the Open3D library (Zhou et al., 2018).

enhanced, detailed regions in our depth estimation show consistency with the accurate surface normal predictions (see Fig. 3).

We set a new benchmark on the C3VD dataset, in which our proposed Col3D-MTL + SSL achieves the best performance among the evaluated state-of-the-art methods on monocular depth estimation (see Table 5). Col3D-MTL + SSL yields a relative improvement of 11.7% on SILog, 28.6% on RMSE, and 10.7% on $\delta_{1.25}$ over our baseline. We also qualitatively compare each of the evaluated methods in Fig. 5, in which we can observe that NDDepth generates higher absolute depth error maps. We consider that the piecewise planarity assumption leveraged by NDDepth can provide geometric guidance to the model in 3D natural scenes. However, this assumption might affect the performance of the model in colonoscopy scenes because of their low degree of regularity and complex topology. Based on our qualitative analysis, our Col3D-MTL + SSL approach estimates depth maps with lower absolute error maps than the other state-of-the art methods.

We assess the impact of similar and distinct scene data in all colon segments by using a careful split between training and test sets of the provided data as discussed in Section 3.1. For example, the sigmoid segment in the test set, which has a different texture from the training samples, generates a significant drop in performance when compared to the caecum or transverse segments which have some resemblance to the training data (Table 6). Additionally, we analyse the dependence of the performance of the model on the structure of the 3D scene. For example, the transverse colon has more folds along the lumen and a broader depth range than the caecum segment (first (caecum) and seventh (transverse) row of Fig. 5). Even though there are more training samples related to the caecum, Col3D-MTL achieves a lower RMSE on the transverse segment (Table 6). Notably, on all colon segments, except for the caecum, our proposals achieve the best $\delta_{1.25}$ value. However, the most remarkable achievement relates to the unseen descending colon segment, in which Col3D-MTL surpasses all the networks presented in this study (only outperformed by Col3D-MTL + SSL on SILog evaluation metric) (see Table 6).

We evaluate temporal consistency in terms of standard deviation across sequential frames for each colon segment. A lower standard deviation denotes higher temporal consistency. We observe that our methods achieve lower RMSE standard deviation values, below one millimetre and within the top two best-performing methods in all colon segments, suggesting stable depth estimation across the video frames (see Table 6). Moreover, we notice that the lack of texture in the caecum decreases the temporal consistency of our approach in terms of SILog and RMSE evaluation metrics.

The disparity maps generated by NDDepth result in distorted point clouds that do not recover a realistic topology of the colon segments, e.g., at the folds of the gastrointestinal tract (Fig. 6). Unlike NDDepth, the support of our auxiliary surface normal task guides our network towards an enhanced and detailed 3D point cloud. In contrast to pix2pix and MonoDepth, Col3D-MTL decreases the generation of artefacts and distortions during the 3D projection of the colonoscopy scene. Among the evaluated methods, NeWCRFs qualitatively yields detailed 3D projections on each colon segment. However, our approach addresses the cases in which NeWCRFs does not lead to a proper 3D representation by generating sharper boundaries at the folds of the colon and recovering the geometry of small protuberances, such as polyps (see Fig. 6).

Our proposed Col3D-MTL qualitatively outperforms our baseline on two real colonoscopy datasets (see Fig. 7). Overall, Col3D-MTL results in smoother depth maps with sharper edges on the boundaries of polyps and gastrointestinal folds. However, we noticed that our method generates wrong depth estimations within the surface of big polyps, as indicated by the red arrows on the last two samples from the PolypGen dataset (see Fig. 7). We consider that the convex shape and prominent size of the polyp affect the performance of our method because similar samples are not available in our training dataset. We observed that self-supervised pre-training with real patients and 3D phantom-based colonoscopy data improves the generalisability of our method to unseen patient colonoscopy datasets with minimal performance drop on the

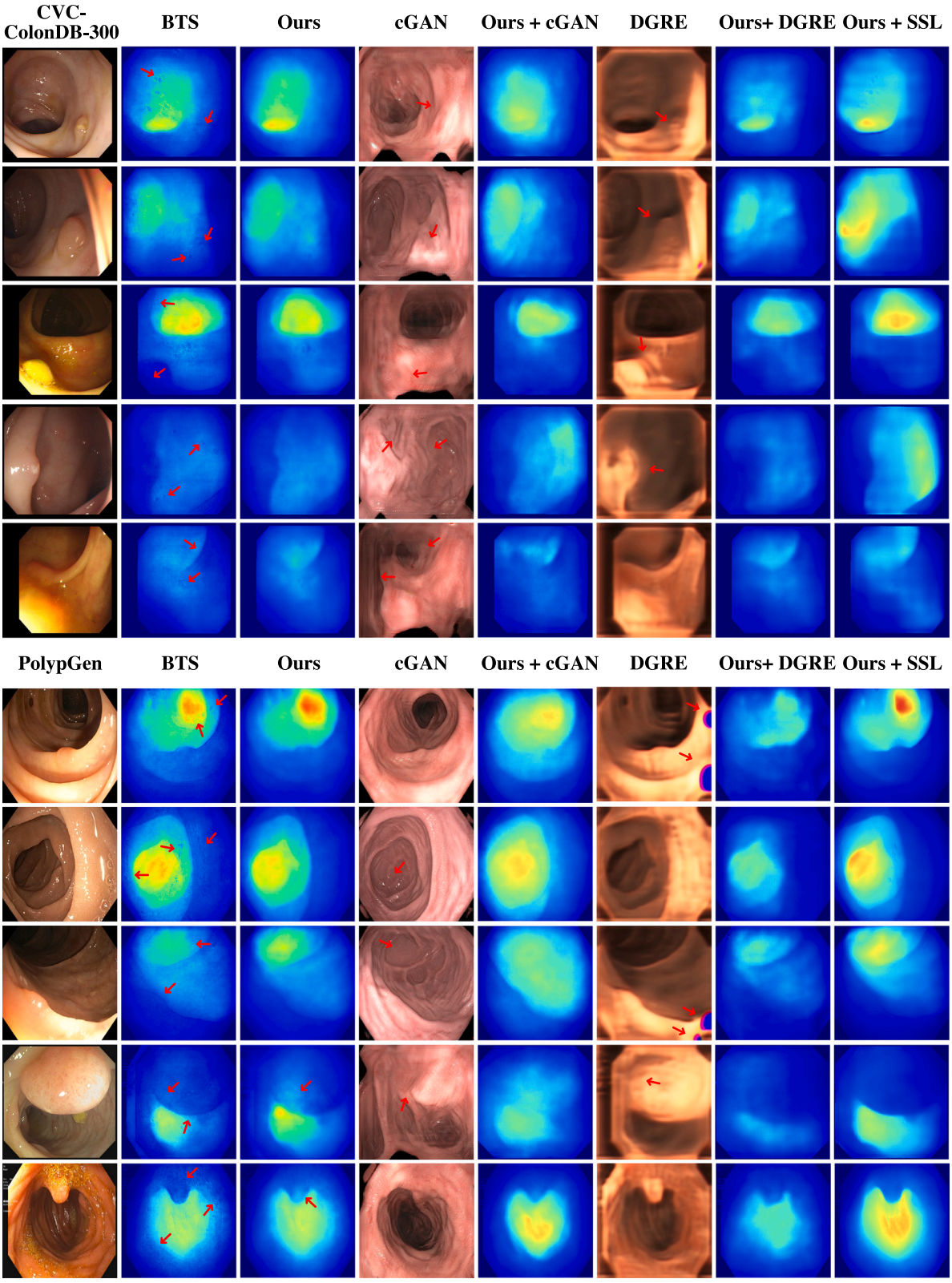

Fig. 7. Qualitative evaluation on real patient colonoscopy samples. We evaluate our baseline method and our proposed approaches on five samples from the CVC-ColonDB-300 (Bernal et al., 2012) and the PolypGen (Ali et al., 2023) datasets. In the first two rows from both datasets, we show the robustness of our Col3D-MTL methods when tested on frames with different photometric properties, e.g., contrast variabilities and specularities. The third sample of each dataset examines the performance of all approaches on different texture and colour details, e.g., boundaries of the polyp and shape of the folds. In the last two samples of each dataset, we compare the predictions on frames with occlusion regions caused by polyps and folds. Notably, Col3D-MTL+SSL achieves refined boundaries of the occlusion regions and mitigates the wrong depth predictions of Col3D-MTL within the protuberances.

source phantom data (see Fig. 7, Tables 4, and 6). While domain gap reduction achieves better results than domain translation using CycleGAN (see Fig. 7), we observed that mapping colonoscopy data from one domain to another is complex due to inherent geometrical changes of the scene that distorts the transformed data. In contrast, we noticed that learning middle-order interactions from pretraining using real patient data and 3D phantom data through a self-supervised MIM approach yields enhanced generalisability to unseen real patient colonoscopy data without such geometrical distortions and hence not affecting the depth predictions.

6. Conclusion

Colonoscopy screening remains the gold standard for diagnosing and treating inflammatory bowel diseases. However, due to its challenging anatomical environment and variable conditions, it is a highly operator-dependent procedure, which usually leads to a high missed-detection rate. Although several approaches have been proposed to detect and segment instruments and polyps, recovering the 3D scene information to perform a quantitative assessment has not been widely studied. Recovering the depth information of a scene is the first step in a 3D reconstruction pipeline. We have identified the current challenges of monocular depth estimation methods and developed our proposed framework towards its applicability in the colonoscopy domain. We selected BTS as our baseline monocular depth estimation method. Given the outstanding local feature representation of convolutional neural networks, our proposed method leverages CBAM attention mechanisms to improve global context awareness and to relate our extracted local features, a surface normal decoder with novel unit normal computation blocks to enhance the 3D representation of the scene, and a cross-task consistency scheme to explicitly enforce consistency among depth and surface normal predictions. To demonstrate the impact of each module, we have provided a comprehensive experimental setup which validates our Col3D-MTL network. We have included a self-supervised masked image modelling-based approach for improving the generalisability of our proposed model on real patient colonoscopy datasets. Our framework is compared against other state-of-the-art monocular depth estimation methods on the C3VD dataset, which is entirely recorded with a high-definition clinical colonoscope on a silicone phantom model that mimics the vascular patterns and the specular appearance of the colon mucosa. Our quantitative results show that our proposed networks outperform current state-of-the-art methods. The most remarkable improvement of our method is achieved on a colon segment that was not given during the training stage. Our qualitative results support adding each proposed module towards a refined feature representation of the colon scene. A qualitative comparison among the 3D projections generated from the depth predictions of each method shows the ability of our proposed framework to generate sharper boundaries at transition zones, recover the shape of small protuberances, and decrease the generation of visual artefacts.

Limitations and future work

The limitations of the proposed frameworks include inaccurate depth and surface normal predictions in regions with low lighting conditions, usually located at the farthest region of the colonoscopy scene. Other cases in which our surface normal decoder does not recover the surface orientation include small regions with high orientation variability, usually encountered as the region gets farther from the colonoscope.

CRediT authorship contribution statement

Pedro Esteban Chavarrias Solano: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Andrew Bulpitt:** Writing – original draft, Supervision, Project administration. **Venkataraman Subramanian:** Writing – original draft, Supervision, Resources, Data curation, Conceptualization. **Sharib Ali:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Code availability

Code is publicly available to support open source and reproducibility at the link: https://github.com/aimsgroup-Leeds/Col3D-MTL.git.

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.

Acknowledgements

This work was supported by Crohn's and Colitis UK (ref: M2023-5). The work was undertaken partly on ARC4, part of the High Performance Computing facilities at the University of Leeds, UK. Part of the work also made use of the facilities of the N8 Centre of Excellence in Computationally Intensive Research (N8 CIR) provided and funded by the N8 research partnership and EPSRC (Grant No. EP/T022167/1). The Centre is co-ordinated by the Universities of Durham, Manchester and York.

Data availability

Publicly available data has been used in this work.

References

Abdelrahim, M., Saiga, H., Maeda, N., Hossain, E., Ikeda, H., Bhandari, P., 2022. Automated sizing of colorectal polyps using computer vision. Gut 71 (1), 7–9. http://dx.doi.org/10.1136/gutjnl-2021-324510.

Ali, S., Bailey, A., Ash, S., Haghighat, M., Investigators, T., Leedham, S.J., Lu, X., East, J.E., Rittscher, J., Braden, B., 2021. A pilot study on automatic three-dimensional quantification of Barrett's esophagus for risk stratification and therapy monitoring. Gastroenterology 161 (3), 865–878.e8. http://dx.doi.org/10.1053/j.gastro.2021.05.059.

Ali, S., Jha, D., Ghatwary, N., Realdon, S., Cannizzaro, R., Salem, O.E., Lamarque, D., Daul, C., Riegler, M.A., Anonsen, K.V., Petlund, A., Halvorsen, P., Rittscher, J., De Lange, T., East, J.E., 2023. A multi-centre polyp detection and segmentation dataset for generalisability assessment. Sci. Data 10 (1), 75. http://dx.doi.org/10.1038/s41597-023-01981-y.

Alistair, W., Joao, C., Chi, X., Joseph, D., Stamatia, G., 2023. Regularising disparity estimation via multi task learning with structured light reconstruction. Comput. Methods Biomech. Biomed. Eng. 11 (4), 1206–1214. http://dx.doi.org/10.1080/21681163.2022.2156391.

Armin, M.A., Chetty, G., De Visser, H., Dumas, C., Grimpen, F., Salvado, O., 2016. Automated visibility map of the internal colon surface from colonoscopy video. Int. J. Comput. Assist. Radiol. Surg. 11 (9), 1599–1610. http://dx.doi.org/10.1007/s11548-016-1462-8.

Bae, G., Budvytis, I., Cipolla, R., 2022. Multi-view depth estimation by fusing single-view depth probability with multi-view geometry. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2842–2851. http://dx.doi.org/10.1109/CVPR52688.2022.00286.

Bao, H., Dong, L., Piao, S., Wei, F., 2022. BEiT: BERT pre-training of image transformers. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=p-BhZSz59o4.

Bernal, J., Sánchez, J., Vilariño, F., 2012. Towards automatic polyp detection with a polyp appearance model. Pattern Recognit. 45 (9), 3166–3182. http://dx.doi.org/10.1016/j.patcog.2012.03.002.

Bobrow, T.L., Golhar, M., Vijayan, R., Akshintala, V.S., Garcia, J.R., Durr, N.J., 2023. Colonoscopy 3D video dataset with paired depth from 2D-3D registration. Med. Image Anal. 90, 102956. http://dx.doi.org/10.1016/j.media.2023.102956.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020b. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. pp. 1597–1607.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848. http://dx.doi.org/10.1109/TPAMI.2017.2699184.

Chen, K., Salz, D., Chang, H., Sohn, K., Krishnan, D., Seyedhosseini, M., 2023. Improve supervised representation learning with masked image modeling. http://dx.doi.org/10.48550/ARXIV.2312.00950.

Chen, J., Yang, X., Jia, Q., Liao, C., 2020a. DENAO: Monocular depth estimation network with auxiliary optical flow. IEEE Trans. Pattern Anal. Mach. Intell. http://dx.doi.org/10.1109/TPAMI.2020.2977021, 1–1.

Cheng, K., Ma, Y., Sun, B., Li, Y., Chen, X., 2021. Depth estimation for colonoscopy images with self-supervised learning from videos. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Springer International Publishing, Cham, pp. 119–128. http://dx.doi.org/10.1007/978-3-030-87231-1_12.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Vol. 2, MIT Press, pp. 2366–2374.

Farooq Bhat, S., Alhashim, I., Wonka, P., 2021. AdaBins: Depth estimation using adaptive bins. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, pp. 4008–4017. http://dx.doi.org/10.1109/CVPR46437.2021.00400.

Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Mac Kain, A., Saillard, C., Schiratti, J.-B., 2023. Scaling self-supervised learning for histopathology with masked image modeling. http://dx.doi.org/10.1101/2023.07.21.23292757, medRxiv.

Gan, T., Jin, Z., Yu, L., Liang, X., Zhang, H., Ye, X., 2023. Self-supervised representation learning using feature pyramid siamese networks for colorectal polyp detection. Sci. Rep. 13 (1), 21655. http://dx.doi.org/10.1038/s41598-023-49057-6.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3354–3361. http://dx.doi.org/10.1109/CVPR.2012.6248074.

Goncharov, M., Pisov, M., Shevtsov, A., Shirokikh, B., Kurmukov, A., Blokhin, I., Chernina, V., Solovev, A., Gombolevskiy, V., Morozov, S., Belyaev, M., 2021. CT-based COVID-19 triage: Deep multitask learning improves joint identification and severity quantification. Med. Image Anal. 71, 102054. http://dx.doi.org/10.1016/j.media.2021.102054.

Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., Tao, D., 2024. A survey on self-supervised learning: Algorithms, applications, and future trends. IEEE Trans. Pattern Anal. Mach. Intell. 1–20. http://dx.doi.org/10.1109/TPAMI.2024.3415112.

Hansen, N., Müller, S.D., Koumoutsakos, P., 2003. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). Evol. Comput. 11 (1), 1–18. http://dx.doi.org/10.1162/106365603321828970, URL: https://direct.mit.edu/evco/article/11/1/1-18/1139.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Islam, M., Vibashan, V.S., Lim, C.M., Ren, H., 2021. ST-MTL: Spatio-Temporal multitask learning model to predict scanpath while tracking instruments in robotic surgery. Med. Image Anal. 67, 101837. http://dx.doi.org/10.1016/j.media.2020.101837.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5967–5976. http://dx.doi.org/10.1109/CVPR.2017.632.

Jeong, B.H., Kim, H.K., Son, Y.D., 2024. Depth estimation from monocular endoscopy using simulation and image transfer approach. Comput. Biol. Med. 181, 109038. http://dx.doi.org/10.1016/j.compbiomed.2024.109038.

Kim, D., Lee, S., Lee, J., Kim, J., 2020. Leveraging contextual information for monocular depth estimation. IEEE Access 8, 147808–147817. http://dx.doi.org/10.1109/ACCESS.2020.3016008.

Koutilya, P.N.V.R., Zhou, H., Jacobs, D., 2020. SharinGAN: Combining synthetic and real data for unsupervised geometry estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 13971–13980. http://dx.doi.org/10.1109/CVPR42600.2020.01399.

Lee, J.H., Han, M.-K., Ko, D.W., Suh, I.H., 2021. From Big to Small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326.

Li, L., Qin, J., Lv, L., Cheng, M., Wang, B., Xia, D., Wang, S., 2023a. ICUnet++: An inception-CBAM network based on Unet++ for MR spine image segmentation. Int. J. Mach. Learn. Cybern. 14 (10), 3671–3683. http://dx.doi.org/10.1007/s13042-023-01857-y.

Li, S., Wu, D., Wu, F., Zang, Z., Li, S.Z., 2023b. Architecture-agnostic masked image modeling – from ViT back to CNN. In: Proceedings of the 40th International Conference on Machine Learning. pp. 20149 – 20167.

Liu, S., Fan, J., Song, D., Fu, T., Lin, Y., Xiao, D., Song, H., Wang, Y., Yang, J., 2022. Joint estimation of depth and motion from a monocular endoscopy image sequence using a multi-loss rebalancing network. Biomed. Opt. Express 13 (5), 2707. http://dx.doi.org/10.1364/BOE.457475.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J., 2023. Self-supervised learning: Generative or contrastive. IEEE Trans. Knowl. Data Eng. 35 (1), 857–876. http://dx.doi.org/10.1109/TKDE.2021.3090866.

Long, X., Lin, C., Liu, L., Li, W., Theobalt, C., Yang, R., Wang, W., 2021. Adaptive surface normal constraint for depth estimation. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, IEEE Computer Society, pp. 12829–12838. http://dx.doi.org/10.1109/ICCV48922.2021.01261.

Ma, R., Wang, R., Zhang, Y., Pizer, S., McGill, S.K., Rosenman, J., Frahm, J.-M., 2021. RNNSLAM: Reconstructing the 3D colon to visualize missing regions during a colonoscopy. Med. Image Anal. 72, 102100. http://dx.doi.org/10.1016/j.media.2021.102100.

Mahmood, F., Durr, N.J., 2018. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. Med. Image Anal. 48, 230–243. http://dx.doi.org/10.1016/j.media.2018.06.005.

Masahiro, O., Hayato, I., Kiyohito, T., Hirotsugu, T., Masaki, M., Hiroshi, N., Kensaku, M., 2022. Depth estimation from single-shot monocular endoscope image using image domain adaptation and edge-aware depth estimation. Comput. Methods Biomech. Biomed. Eng. 10 (3), 266–273. http://dx.doi.org/10.1080/21681163.2021.2012835.

McGill, S.K., Rosenman, J., Zhao, Q., Wang, R., Ma, R., Fan, M., Niethammer, M., Alterovitz, R., Frahm, J.-M., Tepper, J., Pizer, S., 2018. Sa1930 missed colonic surface area at colonoscopy can be calculated with computerized 3D reconstruction. Gastrointest Endosc. 87 (6), AB254. http://dx.doi.org/10.1016/j.gie.2018.04.452.

Ming, Y., Meng, X., Fan, C., Yu, H., 2021. Deep learning for monocular depth estimation: A review. Neurocomputing 438, 14–33. http://dx.doi.org/10.1016/j.neucom.2020.12.089.

Minglan, Z., Weiqi, C., Yisheng, Z., Chun, Z., Linfu, S., Min, H., 2023. A multi-scale deep image completion model fused capsule network. In: 2023 18th International Conference on Intelligent Systems and Knowledge Engineering. ISKE, pp. 288–293. http://dx.doi.org/10.1109/ISKE60036.2023.10481245.

Nakagawa, Y., Uchiyama, H., Nagahara, H., Taniguchi, R.-I., 2015. Estimating surface normals with depth image gradients for fast and accurate registration. In: 2015 International Conference on 3D Vision. pp. 640–647. http://dx.doi.org/10.1109/3DV.2015.80.

Nathan Silberman, P.K., Fergus, R., 2012. Indoor segmentation and support inference from RGBD images. In: European Conference on Computer Vision. ECCV, pp. 746–760. http://dx.doi.org/10.1007/978-3-642-33715-4_54.

Patil, V., Sakaridis, C., Liniger, A., Van Gool, L., 2022. P3Depth: Monocular depth estimation with a piecewise planarity prior. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1600–1611. http://dx.doi.org/10.1109/CVPR52688.2022.00166.

Piccinelli, L., Sakaridis, C., Yu, F., 2023. iDisc: Internal discretization for monocular depth estimation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 21477–21487. http://dx.doi.org/10.1109/CVPR52729.2023.02057.

Qi, X., Liao, R., Liu, Z., Urtasun, R., Jia, J., 2018. GeoNet: Geometric neural network for joint depth and surface normal estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 283–291. http://dx.doi.org/10.1109/CVPR.2018.00037.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V., 2022. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans. Pattern Anal. Mach. Intell. 44 (03), 1623–1637. http://dx.doi.org/10.1109/TPAMI.2020.3019967.

Rau, A., Bhattarai, B., Agapito, L., Stoyanov, D., 2023. Task-guided domain gap reduction for monocular depth prediction in endoscopy. In: Data Engineering in Medical Imaging. Springer Nature Switzerland, pp. 111–122. http://dx.doi.org/10.1007/978-3-031-44992-5_11.

Rau, A., Edwards, P.J.E., Ahmad, O.F., Riordan, P., Janatka, M., Lovat, L.B., Stoyanov, D., 2019. Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. Int. J. Comput. Assist. Radiol. Surg. 14 (7), 1167–1176. http://dx.doi.org/10.1007/s11548-019-01962-w.

Recasens, D., Lamarca, J., Fácil, J.M., Montiel, J.M.M., Civera, J., 2021. Endo-Depth-and-Motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints. IEEE Robot. Autom. Lett. 6 (4), 7225–7232. http://dx.doi.org/10.1109/LRA.2021.3095528.

Rex, D.K., Schoenfeld, P.S., Cohen, J., Pike, I.M., Adler, D.G., Fennerty, M.B., Lieb, J.G., Park, W.G., Rizk, M.K., Sawhney, M.S., Shaheen, N.J., Wani, S., Weinberg, D.S., 2015. Quality indicators for colonoscopy. Gastrointest Endosc. 81 (1), 31–53. http://dx.doi.org/10.1016/j.gie.2014.07.058.

Shao, S., Pei, Z., Chen, W., Wu, X., Li, Z., 2023. NDDepth: Normal-distance assisted monocular depth estimation. In: 2023 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 7897–7906. http://dx.doi.org/10.1109/ICCV51070.2023.00729.

Shao, S., Pei, Z., Chen, W., Zhu, W., Wu, X., Sun, D., Zhang, B., 2022. Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. Med. Image Anal. 77, 102338. http://dx.doi.org/10.1016/j.media.2021.102338.

Siegel, R.L., Wagle, N.S., Cercek, A., Smith, R.A., Jemal, A., 2023. Colorectal cancer statistics, 2023. CA: Cancer J. Clin. 73 (3), 233–254. http://dx.doi.org/10.3322/caac.21772, URL: https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21772.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations.

Tukra, S., Giannarou, S., 2022. Randomly connected neural networks for self-supervised monocular depth estimation. Comput. Methods Biomech. Biomed. Eng. 10 (4), 390–399. http://dx.doi.org/10.1080/21681163.2021.1997648.

Van Rijn, J.C., Reitsma, J.B., Stoker, J., Bossuyt, P.M., Van Deventer, S.J., Dekker, E., 2006. Polyp miss rate determined by tandem colonoscopy: A systematic review. Am. J. Gastroenterol. 101 (2), 343–350. http://dx.doi.org/10.1111/j.1572-0241.2006.00390.x.

Wang, Y., Ni, H., Zhou, J., Liu, L., Lin, J., Yin, M., Gao, J., Zhu, S., Yin, Q., Zhu, J., Li, R., 2024. A semi-supervised learning framework for classifying colorectal neoplasia based on the NICE classification. J. Imaging Inform. Med. http://dx.doi.org/10.1007/s10278-024-01123-9, URL: https://link.springer.com/10.1007/s10278-024-01123-9.

Wang, R., Pizer, S.M., Frahm, J.-M., 2019. Recurrent neural network for (un-)supervised learning of monocular video visual odometry and depth. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5550–5559. http://dx.doi.org/10.1109/CVPR.2019.00570.

Wang, J., Zheng, Y., Ma, J., Li, X., Wang, C., Gee, J., Wang, H., Huang, W., 2023. Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation. Med. Image Anal. 83, 102687. http://dx.doi.org/10.1016/j.media.2022.102687.

Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M., 2021. The temporal opportunist: Self-supervised multi-frame monocular depth. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1164–1174. http://dx.doi.org/10.1109/CVPR46437.2021.00122.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. CBAM: Convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018. Vol. 11211, Springer International Publishing, pp. 3–19. http://dx.doi.org/10.1007/978-3-030-01234-2_1.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H., 2022. SimMIM: A simple framework for masked image modeling. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 9643–9653. http://dx.doi.org/10.1109/CVPR52688.2022.00943.

Xue, W., Brahm, G., Pandey, S., Leung, S., Li, S., 2018. Full left ventricle quantification via deep multitask relationships learning. Med. Image Anal. 43, 54–65. http://dx.doi.org/10.1016/j.media.2017.09.005.

Yang, Y., Shao, S., Yang, T., Wang, P., Yang, Z., Wu, C., Liu, H., 2023. A geometry-aware deep network for depth estimation in monocular endoscopy. Eng. Appl. Artif. Intell. 122, 105989. http://dx.doi.org/10.1016/j.engappai.2023.105989, URL: https://www.sciencedirect.com/science/article/pii/S0952197623001732.

Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P., 2022. Neural window fully-connected CRFs for monocular depth estimation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3906–3915. http://dx.doi.org/10.1109/CVPR52688.2022.00389.

Zamir, A.R., Sax, A., Cheerla, N., Suri, R., Cao, Z., Malik, J., Guibas, L.J., 2020. Robust learning through cross-task consistency. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 11194–11203. http://dx.doi.org/10.1109/CVPR42600.2020.01121.

Zhang, Y., Frahm, J.-M., Ehrenstein, S., McGill, S.K., Rosenman, J.G., Wang, S., Pizer, S.M., 2021b. ColDE: A depth estimation framework for colonoscopy reconstruction. arXiv preprint arXiv:2111.10371.

Zhang, J., Peng, H., Wu, K., Liu, M., Xiao, B., Fu, J., Yuan, L., 2022. MiniViT: Compressing vision transformers with weight multiplexing. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 12135–12144. http://dx.doi.org/10.1109/CVPR52688.2022.01183.

Zhang, S., Zhao, L., Huang, S., Ye, M., Hao, Q., 2021a. A template-based 3D reconstruction of colon structures and textures from stereo colonoscopic images. IEEE Trans. Med. Robot. Bionics 3 (1), 85–95. http://dx.doi.org/10.1109/TMRB.2020.3044108.

Zhou, Q.-Y., Park, J., Koltun, V., 2018. Open3D: A modern library for 3D data processing. arXiv:1801.09847.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T., 2022. iBOT: Image BERT pre-training with online tokenizer. In: International Conference on Learning Representations. ICLR.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on Computer Vision. ICCV, pp. 2242–2251. http://dx.doi.org/10.1109/ICCV.2017.244.

Zou, Y., Luo, Z., Huang, J.-B., 2018. DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency. In: Computer Vision – ECCV 2018. Springer International Publishing, pp. 38–55. http://dx.doi.org/10.1007/978-3-030-01228-1_3.