Medical Image Analysis (2024)

Contents lists available at ScienceDirect

Medical Image Analysis

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

SimCol3D - 3D Reconstruction during Colonoscopy Challenge

Anita Rau[a,b,*], Sophia Bano[a,*], Yueming Jin[a,c,*], Pablo Azagra[d], Javier Morlana[d], Rawen Kader[a], Edward Sanderson[e], Bogdan J. Matuszewski[e], Jae Young Lee[f], Dong-Jae Lee[f], Erez Posner[g], Netanel Frank[g], Varshini Elangovan[h], Sista Raviteja[i], Zhengwen Li[j], Jiquan Liu[j], Seenivasan Lalithkumar[c,l], Mobarakol Islam[k], Hongliang Ren[c,l], Laurence B. Lovat[a], José M.M. Montiel[d], Danail Stoyanov[a]

[a] *Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) and Department of Computer Science, University College London, London, UK*
[b] *Stanford University, Stanford, California, USA*
[c] *National University of Singapore, Singapore*
[d] *University of Zaragoza, Zaragoza, Spain*
[e] *Computer Vision and Machine Learning (CVML) Group, University of Central Lancashire, Preston, UK*
[f] *Korea Advanced Institute of Science and Technology, Daejeon, Korea*
[g] *Intuitive Surgical*
[h] *College of Engineering, Guindy, India*
[i] *Indian Institute of Technology Kharagpur, Kharagpur, India*
[j] *Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering & Instrument Science, Zhejiang University, China*
[k] *Imperial College London, London, UK*
[l] *The Chinese University of Hong Kong, HK, China*

ARTICLE INFO

Article history:
Received -
Received in final form -
Accepted -
Available online -

2000 MSC: 41A05, 41A10, 65D05, 65D17

Keywords: Computer-assisted interventions, Surgical data science, 3D reconstruction, Depth prediction, Camera pose estimation, Navigation, Colonoscopy

ABSTRACT

Colorectal cancer is one of the most common cancers in the world. While colonoscopy is an effective screening technique, navigating an endoscope through the colon to detect polyps is challenging. A 3D map of the observed surfaces could enhance the identification of unscreened colon tissue and serve as a training platform. However, reconstructing the colon from video footage remains difficult. Learning-based approaches hold promise as robust alternatives, but necessitate extensive datasets. Establishing a benchmark dataset, the 2022 EndoVis sub-challenge SimCol3D aimed to facilitate data-driven depth and pose prediction during colonoscopy. The challenge was hosted as part of MICCAI 2022 in Singapore. Six teams from around the world and representatives from academia and industry participated in the three sub-challenges: synthetic depth prediction, synthetic pose prediction, and real pose prediction. This paper describes the challenge, the submitted methods, and their results. We show that depth prediction from synthetic colonoscopy images is robustly solvable, while pose estimation remains an open research question.

1. Introduction

The Endoscopic Vision (EndoVis) challenges at MICCAI have been an accelerator for surgical data science for several years (Maier-Hein et al., 2017, 2020, 2022). Past challenges have evaluated a range of tasks such as segmentation, image generation, or action triplet detection[1]. Although the applications are widely different, all challenges share a profound contribution to their respective research fields by improving data

*Corresponding authors:
e-mail: arau@stanford.edu (Anita Rau), sophia.bano@ucl.ac.uk (Sophia Bano), ymjin@nus.edu.sg (Yueming Jin)

[1] https://endovis.grand-challenge.org/

availability and bringing attention to research gaps. In the spirit of this tradition, the *SimCol3D - 3D Reconstruction during Colonoscopy* challenge was born. SimCol3D is the first challenge to contribute both synthetic and real colonoscopy procedure sequences to address depth estimation and 6D pose estimation from monocular colonoscopy.

Colorectal cancer (CRC) is a leading cause of death (Araghi et al., 2019), third only to lung and breast (for female) and prostate (for male) cancer. Despite its prevalence, survival rates are high among individuals undergoing screening (Kaminski et al., 2010). The slow progression of CRC allows for an extended window for detecting and treating pre-cancerous growths. But to be treated, such growths first need to be accurately detected—an exceedingly difficult task. Fortunately, a cohort of AI-based platforms has declared missed polyps a relic of the past (Puyal et al., 2022; Ji et al., 2021; Zhao et al., 2022; Chadebecq et al., 2023), making it possible to assist clinicians in identifying polyps on the colon mucosa during colonoscopy in real-time. Yet, challenges persist, particularly in detecting polyps hidden behind folds, which constitute up to three-quarters of all missed polyps (Pickhardt et al., 2004). Additionally, other lesions, such as dysplasia in Inflammatory Bowel Disease (IBD) patients, pose an exceptional challenge, necessitating meticulous screening of the entire colon mucosa. The quality of the screening is often quantified as the time taken to withdraw the colonoscope, a critical aspect of the procedure for lesion detection. Withdrawal time is a key surrogate marker for adenoma detection rate (Butterly et al., 2014), which, in turn, is associated with post-colonoscopy CRC rate (Corley et al., 2014). But withdrawal time as a measure of performance has significant limitations. It measures overall time and fails to ensure sufficient attention to each colon segment. A 3D map could help provide more useful quality indicators such as withdrawal time per segment, or ratio of screened colon mucosa.

Researchers have thus proposed to generate an on-the-fly 3D map of the colon during a colonoscopy that can flag areas of the colon that need to be re-screened for colorectal polyps. But providing such a map is difficult. The poor quality of real colonoscopy videos, caused by artifacts such as specularities, air bubbles, blur, saturated pixels, and lack of contrast Ali et al. (2021), presents a significant hurdle to feature-based methods. Repetitive textures and geometries, extreme deformation, and challenging and view-dependent lighting additionally challenge feature matching between images.

Data-driven approaches circumvent the need for robust features and divide the task into depth prediction and pose estimation. But despite the significant progress made by deep learning in reconstructing 3D scenes [cite], the translation of such approaches to colonoscopy is limited by data availability. While cities or rooms can be scanned using lidar or infrared sensors, such scanners are not deployable within a spatially constrained colonoscope. To date, there exists no dataset containing RGB images, camera poses, and depth maps from a real colonoscopy. Attempts to work around this limitation, such as registering previously acquired computer tomography (CT) scans of the colon with images from the procedure, fail due to the immense deformation of the colon during its inspection. Similarly, the calibration of non-medical-grade structured-light sensors, electric-magnetic tracking sensors, and standard colonoscopes is exceptionally difficult, often inaccurate, and only applicable to phantoms that deviate significantly in visual and haptic characteristics from real colons. Synthetic data, though visually distinct, offers precise and abundant annotations.

Previous work leveraging synthetic depth data mostly focused on bridging the domain gap between real and synthetic data (Mahmood and Durr, 2018; Rau et al., 2019; Mathew et al., 2020; Cheng et al., 2021; Itoh et al., 2021; Rodriguez-Puigvert et al., 2022) and employed existing depth networks. In contrast, the challenge organizers were curious to explore depth prediction without accounting for the domain shift between real and synthetic data and chose to evaluate methods directly on synthetic depth. Data-driven pose prediction had yet to be explored widely before the SimCol3D challenge, mostly due to the lack of camera pose ground truth (Rau et al., 2022). We therefore provided synthetic and real pose labels to differentiate between the scenario in which pose networks can be learned in a supervised manner and a scenario where no ground truth is available.

We believe a mapping technology for colonoscopy to be within reach and created the SimCol3D challenge to bringing us one step closer to reliable 3D reconstruction of the colon.

In this paper, we

- introduce the SimCol3D challenge: the first of its kind for depth and pose prediction in colonoscopy;
- analyze each participating group's results, identifying trends and best practices across three subtasks: synthetic depth prediction, synthetic pose estimation, and real pose estimation;
- establish a benchmark for future comparisons of depth and pose estimation methods in colonoscopy;
- introduce synthetic data based on two additional human CT scans, augmenting our existing dataset;
- provide COLMAP labels for real colonoscopy sequences;
- highlight avenues for future investigation.

2. Related work

To date, a profound gap exists between research efforts in depth prediction and pose estimation, both intrinsic subtasks of 3D reconstruction. Depth prediction solves the task of regressing or classifying each pixel in an image, and such tasks are more easily learnable for neural networks if sufficient training data exists. However, understanding camera movement and its geometric implications through regression alone is a much more challenging task and remains underexplored (Rau et al., 2022). Some works have thus focused on leveraging the depth and pose networks in a mutual framework. This section briefly reviews essential works in the field to give context to the participants' contributions.

Most works on depth prediction during colonoscopy have two things in common: they are borrowed from general computer vision approaches, and they incorporate synthetic data in

some way. Some notable virtually generated or phantom-based public datasets were proposed by Rau et al. (2019), Zhang et al. (2020), Ozyoruk et al. (2021), Bobrow et al. (2022), and (Rau et al., 2022). While they were an important addition to the research community, they all consist of one anatomy only, and cannot be used to evaluate accuracy on an unseen patient. Mahmood and Durr (2018) proposed one of the first depth networks for colonoscopy and is based mainly on convolutional neural fields proposed in (Liu et al., 2015). The authors trained one network for depth prediction on synthetic data and used a second, independently optimized network to translate between the appearance of real and synthetic images. Rau et al. (2019) use the well-known pix2pix network (Isola et al., 2017) to integrate the depth and domain translations networks into a single framework trained on both synthetic and real data. Cheng et al. (2021) propose to train a well-known GAN (Wang et al., 2018) on synthetic data with supervision and, in a second, independent step, train the initialized network on real images with self-supervision. Mathew et al. (2020) base their method on the well-known CycleGAN network that maps virtual images to real images and vice versa. Itoh et al. (2021) also borrow the cycle-consistency losses from CycleGAN and decompose images based on a Lambertian-reflection model to train their network on synthetic and real data. Rodriguez-Puigvert et al. (2022) based their method on MonoDepth2 (Godard et al., 2019) and trained an ensemble method with a teacher trained on synthetic data. Though these methods help progress the field, all of these method primarily focus on bridging the domain gap between synthetic and real images, not on improving the architectures of the respecitve depth networks. Accordingly, the evaluation protocols focused on real colonoscopy frames that are oftentime borrowed from in-house datasets. A common benchmark allowing a systematic comparison of these methods is missing.

While methods that predict depth only largely rely on synthetic data, approaches combining depth and pose networks can directly learn from real data. Bae et al. (2020) use sparse SfM pseudo ground truth to supervise their colon reconstruction pipeline. They reconstruct small colon sections from eight consecutive frames using the derived poses and sparse depth supervision to guide the initial U-Net based (Ronneberger et al., 2015) depth estimation. Ma et al. (2019) propose a SLAM pipeline that integrates a well-known recurrent neural net for depth and pose estimation Wang et al. (2019). Freedman et al. (2020) and Ozyoruk et al. (2021) propose self-supervised networks based on the popular depth and pose networks (Gordon et al., 2019) and (Bian et al., 2019), respectively. All these approaches do not require synthetic data; however, they can only be as accurate as the underlying feature-based SfM reconstruction. Additionally, these works use in-house datasets and do not provide a sufficient comparison between each other.

3. Tasks and datasets

3.1. Challenge Tasks

The SimCol3D challenge aims to facilitate depth and camera pose prediction during colonoscopy by providing a new public dataset with ground truth depths and poses for training and testing. The challenge comprises three tasks: Task 1 invited participants to train networks to predict depth from simulated colonoscopy images. Task 2 evaluates predicted camera poses from simulated colonoscopy. Task 3 extends the challenge into the realm of real-world clinical practice, tasking participants with predicting poses from real colonoscopy procedures.

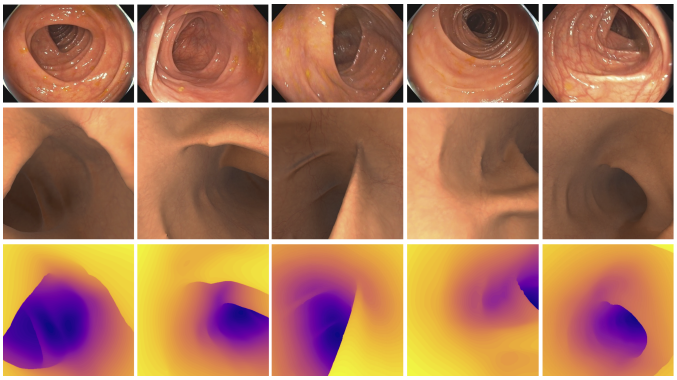

Fig. 1: Overview of the real images (top), synthetic images (center), and synthetic depth maps (bottom) used in the challenge.

Sub-dataset	# Train traj.	# Test traj.	# Images
Synthetic colon I (Public mesh)	12	3	18k
Synthetic colon II (Patient A)	12	3	18k
Synthetic colon III (Patient B)	0	3	1.8k
Real Sequences	59	7	-

Table 1: Overview of the datasets in the SimCol3D challenge, indicating the number of trajectories (traj.) and images per scene. The real sequences provide videos only.

3.2. Data

The SimCol3D challenge encompasses both synthetic and real colonoscopy sequences. Table 1 provides an overview of the data used in the challenge, and Figure 1 shows illustrates qualitative examples.

3.2.1. Simulated colonoscopy data for Tasks 1 and 2

The synthetic data for Tasks 1 and 2 builds upon the dataset introduced in (Rau et al., 2022), but expands its scope from one anatomy (Synthetic Colon I) to encompass three distinct human colons (Synthetic Colons I, II, and III). Two of the three subsets (namely I and II) contain 15 trajectories of which 12 were randomly assigned for training and three for testing. Synthetic Colon III only contains 3 trajectories for testing and no training data. This setup allows to evaluate generalizability to new anatomies. Each training trajectory contains 1201 images, ground truth depth maps, and ground truth camera poses. Each test trajectory contains either 1201 or 601 frames and their labels. The simulated colon meshes were extracted from computer tomography scans of human colons, and the images were rendered using a Unity simulation environment (Rau et al., 2019). The CT scan for Synthetic Colon I is publicly available (Ozyoruk et al., 2021), while the CT scans for Colons II and III were acquired at University College London Hospital. In the simulation environment, a virtual colonoscope followed a path

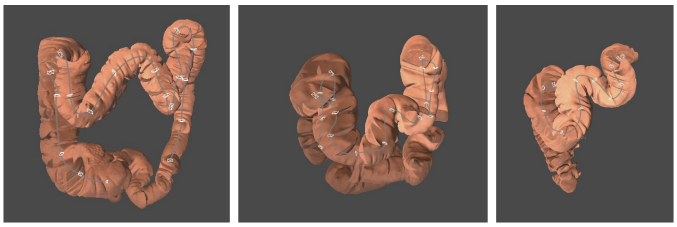

Fig. 2: Synthetic Colons I, II, and III in Unity environment with camera paths along the center of the mesh. Synthetic Colons I and II include training and test trajectories. Synthetic Colon III provides test trajectories only.

through the center of the meshes recording rendered RGB images, depth maps, and camera poses. This path was randomly manipulated each time a new trajectory was recorded resulting in different, random trajectories within the same anatomy. In total, the training data for Task 1 and Task 2 contains 14,412 frames, and the test set contains 9,009. For each frame the challenge organizers provided the corresponding:

- 3×3 camera intrinsics matrix saved as *txt* file.
- Depth map in *png* format, including the depth value for each pixel in the corresponding RGB image.
- Absolute camera pose in the Unity coordinate frame provide in a .txt file. We represent camera pose as 7D vector $[t_x, t_y, t_z, q_x, q_y, q_z, q_w]$, where t denotes the translation along the x-, y-, and z-axes, and q denotes the rotation in quaternion representation, where w denotes the scalar part, and x, y, and z describe the imaginary parts.

More details about the data generation process and the coordinate systems used to represent the data can be found in the original publication (Rau et al., 2022). The full synthetic dataset is publicly available here: `https://www.ucl.ac.uk/interventional-surgical-sciences/simcol3d-data`. Participants were allowed to use additional datasets as long as they were publicly available.

3.2.2. Real patient data for Task 3

For Task 3, the testing data contained three patients' anatomies with 1–3 trajectories from each and 7 in total. The real data comes from the EndoMapper dataset (Azagra et al., 2022), which is a collection of complete endoscopy sequences obtained during regular medical procedures[2]. It includes 59 sequences with over 15 hours of video and is the first endoscopic dataset to include geometric and photometric endoscope calibration. The dataset also includes meta-data and annotations. The participants were encouraged to train their models on the EndoMapper sequences that were not in the test set. For this task, we generated COLMAP pseudo ground truth of the 7 testing sequences for method evaluation. As COLMAP is not reliable in colonoscopy (therefore the need for this challenge), two challenge organizers and a gastroenterologist , qualitatively verified each of the generated COLMAP trajectories and sparse point clouds and chose those that were visually coherent with respect to the direction of the movement of the endoscope observed in the corresponding video.

[2] `https://www.synapse.org/Synapse:syn26707219/wiki/615178`

3.3. Evaluation metrics

3.3.1. Task 1: Depth estimation

We utilize three standard evaluation metrics to assess the performance of the depth prediction methods. We define the per image errors as

$$L_1 = \frac{1}{D} \sum_{d=1,\dots,D} \|Y(d) - s \cdot Y'(d)\|_1 \tag{1}$$

$$L_{rel} = \mu_d \left(\left\| \frac{Y(d) - s \cdot Y'(d)}{Y(d)} \right\|_1 \right) \tag{2}$$

$$L_{RMSE} = \sqrt{\frac{1}{D} \sum_{d=1,\dots,D} (Y(d) - s \cdot Y'(d))^2} \tag{3}$$

where Y denotes the ground truth depth map, Y' denotes the predicted depth map, D is the number of pixels in Y, and μ_d represents the median calculated for all valid arguments d. Let

$$\bar{Y}_i = \frac{1}{D} \sum_{d=1,\dots,D} Y_i(d), \tag{4}$$

denote the mean depth over all pixels in a depth map i, then the scale s is calculated per trajectory as

$$s = \frac{\sum_{i \in I} \bar{Y}_i \cdot \bar{Y}'_i}{\sum_{i \in I} \bar{Y}'_i \cdot \bar{Y}'_i}, \tag{5}$$

where I denotes the number of images in a trajectory. We chose to evaluate the scaled depths, as the task of monocular depth estimation is ill-posed and networks are expected to predict depth up to scale. We compute the L_1 loss as the mean of the absolute differences between the ground truth depth $Y(d)$ and the predicted depth $Y'(d)$ over all pixels in a depth map. As the relative loss, L_{rel}, is sensitive to outliers, we use the median instead of the mean over the per-pixel relative L_1 errors. Lastly, we measure the L_{RMSE} as it weights outliers more heavily than the L_1 loss. The per-depth map errors are then averaged over all depth maps in a scene.

As we found all three metrics to be equally descriptive of performance, but due to their different scales not comparable, we use a point system for Task 1. We report the final score, $\sum_1$, as the sum of ranks per scene. For each of the three scenes and each of the three metrics, the winner received six points, the runner-up five points, etc. The task winners were the groups with the most points.

3.3.2. Task 2: Camera pose estimation on simulated data

To evaluate the predicted camera poses, we first composite the relative poses Ω_i to produce the complete trajectory of absolute poses P_i. The absolute pose of a camera τ in the world space is $P_1\Omega_1 \cdots \Omega_{\tau-1}$, where each Ω_i sequentially projects the initial pose P_1 to the next one. As monocular video can only

be interpreted up to scale, the predicted trajectory needs to be scaled using:

$$s_{rel} = \frac{\Sigma_\tau trans(\Omega_\tau)^T \cdot trans(\Omega'_\tau)}{\Sigma_\tau trans(\Omega'_\tau)^T \cdot trans(\Omega'_\tau)}, \tag{6}$$

where *trans* denotes the translation of a projection matrix. We then assess the scaled predicted trajectory's accuracy with the Absolute Translation Error (ATE), Relative Translation Error (RTE), and Rotation Error (ROT).

$$\begin{aligned} RTE &= \mu_\tau(\|trans(\Omega_\tau^{-1}\Omega_\tau^{'})\|) \\ ATE &= \mu_\tau(\|trans(P_\tau) - trans(P_\tau^{'})\|) \\ ROT &= \mu_\tau(\frac{trace(Rot(\Omega_\tau^{-1}\Omega_\tau^{'})) - 1}{2} \cdot \frac{180}{\pi}) \end{aligned} \tag{7}$$

where Rot denotes the projection rotation, $\Omega^{'}, P^{'}$ are the scaled predicted relative and absolute poses, and $\|\cdot\|$ is the two-norm. The ATE measures drift and the overall consistency of a predicted trajectory. The ROT measures the magnitude of the rotation errors locally. The RTE reflects both translation and rotation errors locally. To achieve a small RTE, the predicted relative pose Ω' must be close to the ground truth Ω, so that $\Omega_\tau^{-1}\Omega_\tau^{'}$ is close to an identity matrix. This is achieved, when both $trans(\Omega')$ and $Rot(\Omega')$ are accurate. We consider the forward direction only. Evaluating these three evaluation metrics, we obtain a comprehensive assessment of the performance of the pose prediction models. To determine the winner of Task 2, we define the task loss $\sum_2$ as the weighted average of RTEs on the three scenes, where we weight SynCol III twice to account for the increased difficulty of pose prediction on an unseen scene.

3.3.3. Task 3: Camera pose estimation on real-world data

We use the same evaluation metric for Task 3 as for Task 2. In particular, we determine ATE, RTE, and ROT as defined in Equation 7. However, we scale the entire trajectory based on the absolute poses to reflect that we are more interested in the global consistency in Task 3, than in local accuracy. The scaling factor in Task 3 is defined as:

$$s_{abs} = \frac{\Sigma_\tau trans(P_\tau)^T \cdot trans(P'_\tau)}{\Sigma_\tau trans(P'_\tau)^T \cdot trans(P'_\tau)}, \tag{8}$$

The task score $\sum_3$ for Task 3 has three components: ATE, RTE, and ROT averaged over all seven scenes.

3.4. Challenge organization

The challenge was a one-time event with fixed submission deadline of September 2022. In order to access the train and test data, participants had to register participation in the challenge on the challenge website[3]. The teams provided their predictions for the test sets via the challenge website based on detailed submission guidelines including docker templates and evaluation scripts that participants could use for validation [4]. The ground truth for the test data was published after the challenge had ended. The participants were not required to publish their code, but links to the code bases of the teams that chose to are provided in Section 4. Ethics approval was not necessary for this challenge. In total, we received and approved 51 challenge registration requests and 13 team registration requests.

4. Methods for Task 1: Depth prediction from synthetic images

For Task 1, final submissions were received from six teams. Table 2 summarizes the key features of the teams' methodology for Task 1. Team details and the methodology proposed by each participating team are presented below.

4.1. FCBFormer adaptation by Team CVML

Team CVML are Edward Sanderson and Bogdan J. Matuszewski from the University of Central Lancashire (UK). Team CVML proposed the FCBFormer-D (as shown in Fig. 3 (I)), which is an adaptation of the FCBFormer (Sanderson and Matuszewski, 2022).

The overall architecture of FCBFormer-D is shown in Fig. 3(I-a). The method consists of two branches: a transformer-based branch (TB) (Fig. 3(I-b)) extracting global features, and a convolutional branch (CB) (Fig. 3(I-c)) extracting local features that the TB could potentially neglect. For the Transformer branch, the Pyramid Vision Transformer v2 (PVTv2) (Wang et al., 2022) (B3 variant pre-trained on ImageNet), which serves as image encoder and provides robust multiscale features for dense prediction, is employed. PVTv2 then feeds into a lightweight decoder. The convolutional branch is based on a UNet-style architecture inspired by (Nichol and Dhariwal, 2021) and includes multi-head self-attention at the lower levels to provide the model with global context for this feature extraction. The feature maps from both branches are then concatenated and fused using a UNet-style architecture also inspired by (Nichol and Dhariwal, 2021) in the fusion module (FM) Fig. 3(I-d).

Finally, the output of the fusion module is passed through the prediction head (PH) Fig. 3(I-e). The prediction head is a 1x1 convolutional layer with sigmoid activation that outputs dense depth map. The depths are then upsampled to the original size of 475×475 using bilinear interpolation.

The implemented network takes a 352×352 RGB image with pixel intensities in the range $[-1, 1]$ as inputs. This involves resizing the 475×475 8-bit RGB images using bilinear interpolation with anti-aliasing prior to normalization. The output of the proposed network then provides a 475×475 depth map with relative depth values in the range $[0, 1]$. During training and validation, the ground truth depth values were scaled to a range of $[0, 1]$, corresponding to [0cm, 20cm], and the model was optimized to minimize the mean squared error (MSE) loss. Team *CVML* used AdamW optimizer with a learning rate of $1e-4$, which was scheduled to halve when the MSE on the validation data did not decrease over 10 epochs. The inputs were randomly horizontally and vertically flipped with a probability of 0.5. The model was trained for 300 epochs with a batch size

[3] https://www.synapse.org/Synapse:syn28548633/wiki/
[4] https://github.com/anitarau/simcol

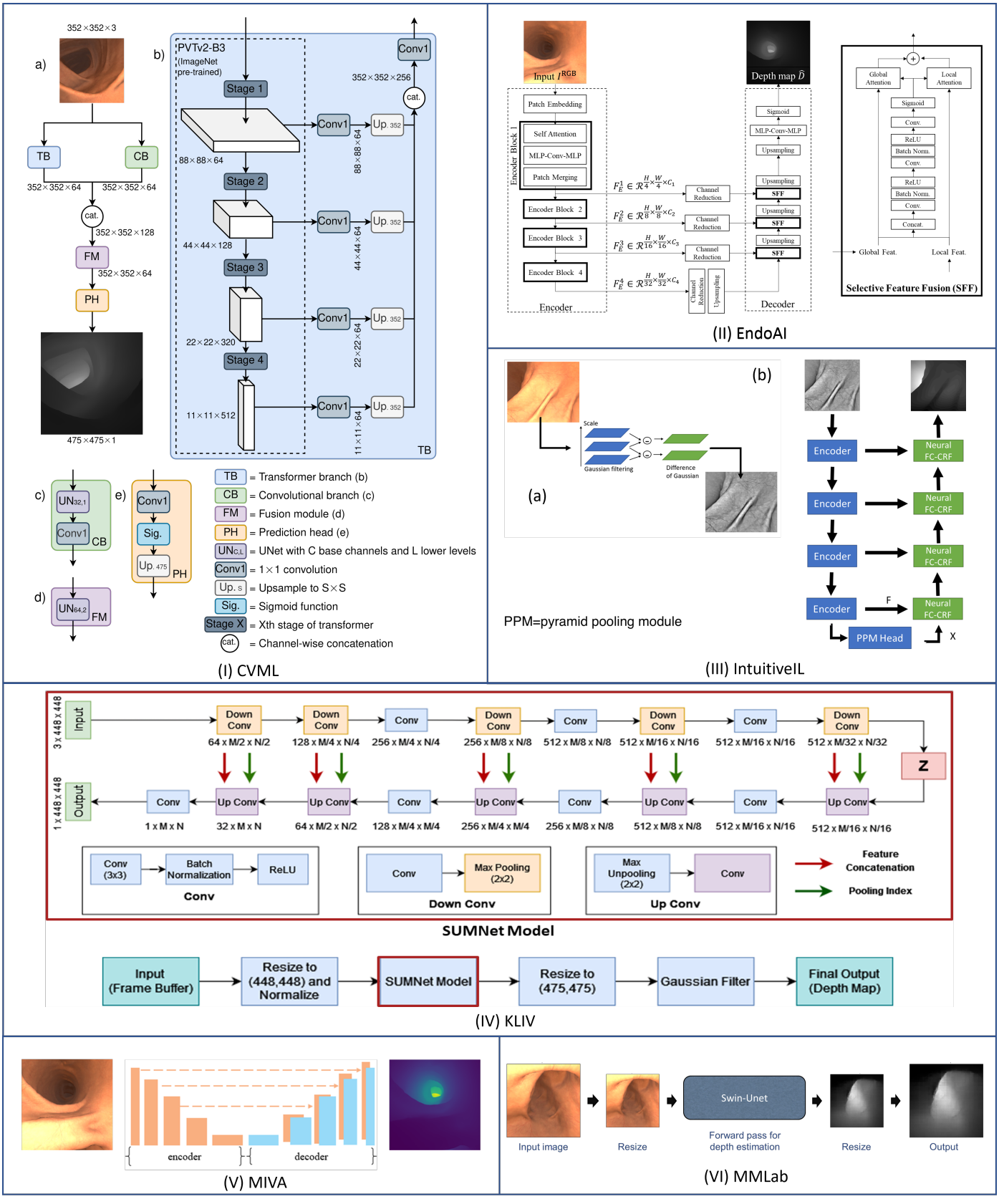

Fig. 3: Architecture overview for Task 1 (depth prediction) of the 6 participating teams. (I) Team CVML adapted FCBFormer (Sanderson and Matuszewski, 2022), (II) Team EndoAI utilized GLPDepth (Kim et al., 2022) with Segformer encoder (Xie et al., 2021), (III) Team IntuitiveIL applied multiple DoG filters with varying scales as preprocessing and used a NeW CRF network for depth prediction, (IV) Team KLIV utilized SUMNet (Nandamuri et al., 2019), (V) Team MIVA utilized DenseDepth (Alhashim and Wonka, 2018) as an encoder-decoder network with skip connections, (VI) Team MMLab utilized Swin-UNet (Cao et al., 2023).

Table 2: Summary of the participating teams of the SimCol Challenge – Task 1

Team name	Algorithm	T/C	Loss function	Preprocessing	Data augmentation	Post-processing
CVML	FCBFormer adaptation (Sanderson and Matuszewski, 2022)	T&C	MSE	Alpha channel removed, pixel intensity normalization, resize, depth scaling	Horizontal & vertical flips	None
EndoAI	GLPDepth (Kim et al., 2022)	T&C	SILog	Normalization, Horizontal flip	Vertical cut depth	Conv2D-ReLU-Conv2D block to adjust the resolution of output images
IntuitiveIL	NeWCRFs (Yuan et al., 2022)	T	SILog	GC, DoG, normalization, HSV	AS, FA	None
KLIV	SUMNet (Nandamuri et al., 2019)	C	MAE + MSE + SIL + BL	Resizing, normalization		Low-pass GB
MIVA	DenseDepth (Alhashim and Wonka, 2018)	C	MAE + SSIM	normalization	Horizontal flip	None
MMLab	Swin-UNet (Cao et al., 2023)	T	L1	Downsampling	None	Upsampling

T: Transformer backbone; C: Convolutional backbone; MAE: Mean Absolute Error; MSE: Mean Squared Error; BL: Berhu Loss; GC: Gamma Correction; DoG: Difference of Gaussian filter; AS: Average Shape; FA: Feature Augmentation; GB: Gaussian Blur; SSIM: Structural Similarity loss; SILog: Scale-Invariant Logarithmic loss; HSV: Hue Saturation Value.

of 24. The network weights with the smallest MSE on the validation set were saved. Training was performed on an ASUS ESC8000-G4 GPU server with six NVIDIA RTX A6000 48GB GPUs.

The groups' method is inspired by their observation that a standard UNet performs relatively weak at inferring the edges of the geometry, as well as the depth of far away surfaces. Their network thus aims to capture both, global features that help understand depth at all distances, and local features that can infer steps in depth.

4.2. GLPDepth adaptation by Team EndoAI

Team EndoAI are Jiwoon Jeon from EndoAI (Korea) and Jae Young Lee, Dong Jae Lee and Woonghyun Ka from Korea Advanced Institute of Science and Technology (Korea), who participated in all three Tasks. Team EndoAI proposed to use GLPDepth (Kim et al., 2022), a Transformer-based network for depth prediction (as shown in Fig. 3(b), for the depth prediction task because this method has shown higher generalization ability and robustness compared to previously developed networks. To obtain the depth map prediction D_{pred} from the input I^{RGB}, the local and global features are fused by Selective Feature Fusion (SFF) in the decoder. For the encoder, Segformer (Xie et al., 2021) is utilized.

The last layer of the original GLPDepth network decoder is modified to include a Conv2D-ReLU-Conv2D block to adjust the resolution of the resulting depth map. Further, to avoid scale adjusting, the model is directly trained to predict depth maps in the range of $[0, 1]$ (corresponding to [0cm, 20cm]) instead of using median scaling. GLPDepth uses the Scale-Invariant Logarithmic (SILog) loss (Eigen et al., 2014) given by:

$$L(D_{pred}, D_{GT}) = \sqrt{\frac{1}{T}\sum_i d_i^2 - \left(\frac{1}{T}\sum_i d_i\right)^2}, \tag{9}$$

where d_i is the pixel-wise log loss

$$d_i = \log(D_{pred}(i)) - \log(D_{GT}(i)) \tag{10}$$

and T denotes the number of pixels in the depth map. For training the GLPDepth model, the original hyperparameters from (Kim et al., 2022) are used. The model is fine-tuned for 20 epochs using the *CosineAnnealingWarmRestarts* learning rate scheduler (Loshchilov and Hutter, 2016) on the challenge metrics: L1 depth error, RMSE, and relative depth error. The final model is chosen based on the performance of all metrics on the validation set.

4.3. NewCRFs adaptation by Team IntuitiveIL

Team IntuitiveIL are Erez Posner, Netanel Frank, and Moshe Bouhnik from the Intuitive Surgical, who proposed to adapt Neural Window Fully-connected Conditional Random Fields (NeW CRFs) (Yuan et al., 2022) to accomplish colonoscopy monocular depth estimation leveraging the advantages of fully-connected (FC) CRFs (He et al., 2004). In addition, they employed data augmentation techniques to address the issue of illumination changes, which involved creating partially illumination-invariant images.

For depth estimation, NeW CRFs are selected because they overcome the limitations of traditional depth estimation methods that rely on Markov Random Fields (MRFs) or CRFs (Saxena et al., 2008, 2005). NeW CRFs embed a vision transformer to capture pairwise interactions with multi-head attention as the encoder and the neural CRFs module in a network as the decoder. NeW CRFs can capture the relationship between any node in a graph, making them much stronger than neighbor CRFs. By splitting the input into windows and performing FC-CRFs optimization within each window, NeW CRFs reduce computation complexity while maintaining the advantages of FC-CRFs. Additionally, the use of multi-head attention within a neural CRFs module further improves depth estimation performance. As shown in Fig. 3III(b), the encoder initially extracts features across four levels. A Pyramid Pooling Module (PPM) combines both global and local data, generating the preliminary prediction X using the uppermost image feature F. Subsequently, within each level, the neural window fully-connected CRF component constructs multi-head energy from X and F, refining it to an improved prediction X'.

In colon augmentation, the method originally proposed in (Ye et al., 2014) for face recognition is utilized, which contains the following steps to create the grayscale illumination-

invariant image:

$$\begin{aligned} I_{gamma} &= GammaCorrection(\text{Image}) \\ I_{DoG} &= DoG(I_{gamma}) \\ I_{norm} &= \frac{I_{DoG}}{mean(|I_{DoG}|^a)^{\frac{1}{a}}} \\ I_{norm} &= \frac{I_{norm}}{mean(min((\tau), |I_{norm}|^a))^{\frac{1}{a}}} \\ I_{norm} &= \frac{\tau * tanh(I_{norm})}{\tau}, \end{aligned} \tag{11}$$

where *GammaCorrection* involves gamma correcting all images to the same value, and *DoG* represents the difference of Gaussians filter. In the augmentation process, the original DoG image is replaced with an average of several DoG filters with varied scales (as illustrated in Fig. 3III(a)). This augmentation aims to improve local texture and accommodates features of various sizes. Additionally, the input image is changed from an RGB to an HSV representation, the value channel is swapped for the algorithm's output in grayscale, and the resulting image is then converted back to an RGB representation. This allowed stronger features even in the colon's distant areas. Scale-Invariant Logarithmic (SILog) loss is utilized as the loss function. SILog supervises the training by first calculating the logarithm difference between the predicted and the ground-truth depth map. For K pixels with valid depth values in an image, the scale-invariant loss is computed to measure the performance of the depth estimation (Yuan et al., 2022).

4.4. SUMNet adaptation by Team KLIV

Team KLIV are Varshini Elangovan from College of Engineering, Guindy (India), and Sista Raviteja, Rachana Sathish, Debdoot Sheet from the Indian Institute of Technology Kharagpur (India). KLIV proposed to apply a fully convolutional neural network SUMNet (Nandamuri et al., 2019) to effectively generate colon depth maps from frame buffers while preserving conformity around small structures and preventing the loss of critical information.

Concretely, SUMNet (Nandamuri et al., 2019) consists of an encoder network with VGG11 architecture, activation concatenation, and pooling index transfer. Several loss functions are taken into account during the training process, including Mean Absolute Error (MAE), Mean Squared Error (MSE), scale-invariant loss (Eigen et al., 2014), and Berhu loss (Carvalho et al., 2018). In order to reduce the aliasing effect in the predicted depth maps, a post-processing step is used to apply a Gaussian Blur low-pass filter with a kernel size of 7×7.

From the provided training data 10,309 frames are used for training, and 3,603 frames are used for validation. To give a more thorough summary, the frame buffers in the simulated dataset were initially in RGBA format, but for network compatibility, they are converted to RGB images and resized to 448 $\times$ 448. Additionally, the images are normalized using the training dataset's mean and standard deviation. The depth maps are scaled to 448×448 and translated into grayscale images. These preprocessed images and depth maps are then used for training the SUMNet model for depth estimation of synthetic colonoscopic images. The network is implemented in PyTorch and trained for 50 epochs on an Nvidia GeForce GTX TITAN X GPU with a batch size of 16 using the ADAM optimizer with an initial learning rate of 0.001 and an exponential learning rate scheduler with a decay factor of 0.98. The complete training took 24 hours.

The effectiveness of the model and the reliability of its predictions are assessed using the L1 error, relative error, and root-mean-square error. The model trained on MSE loss predicted results that are more reliable and accurate, in comparison to the models trained on the other loss functions. KLIV's code is available[5].

4.5. DenseDepth adaptation by Team MIVA

Team MIVA are Zhengwen Li and Yichen Zhu from Zhe-Jiang University (China), who participated in all three Tasks. MIVA used DenseDepth (Alhashim and Wonka, 2018) which is a fully convolutional encoder-decoder architecture with skip connections (as shown in Fig. 3(V)). The encoder is a DenseNet-169 (Huang et al., 2017) pre-trained on ImageNet (Deng et al., 2009) as proposed by the original DenseDepth. The authors also experimented with a DenseNet-201, which performed worse in their experiments. To train the network, MIVA used the loss L as the weighted sum between the depth and SSIM loss:

$$L(Y, Y') = 0.1 \cdot L_{depth}(Y, Y') + L_{SSIM}(Y, Y'). \tag{12}$$

The Loss term L_{depth} is the point-wise L1 loss defined on the depth values and L_{SSIM} uses the Structural Similarity (SSIM). The authors replace the original augmentation strategy with a 50% random horizontal flipping and image normalization only. The synthetic data provided is split into training set(Rau et al., 2022) and validation set in the way recommended by the SimCol3D challenge organizers, and the mean and standard deviation in normalization are calculated from all images in the training set. The participants trained their method on an NVIDIA GeForce RTX 3090 GPU using a batch size of 16 and a learning rate of 10^{-4} with Adam optimizer for 40 epochs.

4.6. Swin-UNet adaptation by MMLAB

Team MMLAB are Seenivasan Lalithkumar, Islam Mobarakol and RenHongliang are from National University of Singapore (Singapore), Imperial College London (UK) and Chinese University of Hong King (China), who participated in Task 1 and 2.

For the depth estimation task, a Unet-like Swin-Transformer (Swin-UNet) (Cao et al., 2023) (Fig. 3(VI)), a medical image segmentation model, is used. Swin-UNet forms a hierarchical Swin Transformer with shifted windows in the encoder, a decoder with patch expanding layer to perform upsampling on the feature maps and skip connections for local-global semantic feature learning. Overall, there are three blocks of the encoder and corresponding decoder in Swin-UNet. The model

[5] `https://github.com/SistaRaviteja/Colonoscopy-Depth-Estimation`

Table 3: Summary of the participating teams of the SimCol3D Challenge – Task 2 and Task 3

Team	Task	Algorithm	Loss function	Data augmentation
EndoAI	2 (Pose Syn.)	MonoDepth2 (Godard et al., 2019)	MSE	None
	3 (Pose Real)	CycleGAN + MonoDepth2	Same as Task 2	CycleGan real-to-syn conversion
MIVA	2 (Pose Syn.)	SC-sfMLearner (Bian et al., 2021)	SC-sfMLearner + Densedepth	Image normalization
	3 (Pose Real)	CycleGAN + SC-sfMLearner	Same as Task 2	Crop, resize, CycleGAN real-to-syn conversion
MMLab	2 (Pose Syn.)	Curriculum learning, linear regression	MSE	None
	3 (Pose Real)	N/A	N/A	N/A

was trained using L1 loss and SGD optimizer with a learning rate of 0.01, a decay factor of 1e-4, and a momentum of 0.9. The input images are resized to 224×224 during training and upsampled to the original size at test time after the prediction. The participants experimented with different loss functions such as L1, mean square error (MSE), structural similarity index (SSIM), and binary cross entropy (BCE). Ultimately, the L1 loss outperformed other loss functions with an MSE of 0.000115 and an SSIM of 0.984670. The team's code is publicly available[6].

5. Methods for Task 2 and 3: Pose prediction from synthetic and real images

In total, 3 teams (EndoAI, MIVA and MMLab) participated in Task 2 (pose prediction from synthetic), two of which (EndoAI and MIVA) also participated in Task 3 (pose prediction from real images). Table 3 provides an overview of the key features of the teams' methodology. The remainder of this section describes the participants' methods in detail.

5.1. SC-SfMLearner adaptation by MIVA

For the pose estimation task, MIVA used a method based on SC-SfMLearner as shown in Fig. 4(II), which includes two parts: a depth estimation module and a pose estimation module. In addition, they replaced the DispResNet depth estimation module in the original SC-SfMLearner with a DenseDepth network. As ground truth depth for synthetic data was known, MIVA made use of this information while training the formerly self-supervised SC-SfMLearner. To supervise the depth module, the loss of Densedepth was added to the original loss of SC-SfMLearner. The modified loss function is

$$L = L_{SC-sfmformer} + \omega \cdot L_{densedepth}, \tag{13}$$

where the weight ω was set to 1.

The team divided the dataset according to their split for Task 1 and also normalized the input images. MIVA's model was trained on an NVIDIA GeForce RTX 3090 GPU with a batch size of 8, learning rate of 10^{-4} and Adam optimizer. The network was trained for 40 epochs.

For the Task 3, MIVA used CycleGAN, which consists of two generators and two discriminators as shown in Fig. 4(IV), where A represents the real colonoscopy image domain, and B represents the virtual colonoscopy image domain. The input image A generates Fake_B through Generator G, and Fake_B generates Rec_A through Generator F. After two transformations, Rec_A is mapped back to the A domain. The model is optimized by comparing the similarity between Input_A and Rec_A. Input_B is processed in the same way. The generator in this paper adopts a ResNet backbone (He et al., 2016), and the discriminator uses a PatchGAN structure. The EndoMapper (Azagra et al., 2022) dataset and the synthetic dataset (Rau et al., 2022) provided by the SimCol3D Challenge were used for training the CycleGAN. Since there are black areas in the four corners of the EndoMapper dataset, MIVA cropped the areas from (155, 0) to (1162, 1007) and reduced them to a 480×480 square. During training, MIVA applied random horizontal flipping and normalization to the data. Preliminary validation results show that CycleGAN's generator can map multiple inputs to the same output. For example, a real colonoscopy image is converted to generate a completely different virtual image. For this reason, MIVA experimented with identity loss, self-regularization loss and SSIM loss to guide the generator. Ultimately, the team used an identity loss in the final submitted model. The experiments for Task 3 was carried out on an NVIDIA GeForce RTX 2080ti.

5.2. MonoDepth2 adaptation by EndoAI

EndoAI's camera pose estimation framework is based on the self-supervised monocular depth estimation method called MonoDepth2 (Godard et al., 2019) (as shown in Fig. 4(I-a)) but is trained using supervision with the ground truth translations and rotations. In addition to the original self-supervised loss, the team added the supervised loss

$$L(P_{pred}, P_{GT}) = \sum_{i,j} \|P_{pred} - P_{GT}\|_1, \tag{14}$$

where P_{pred} and P_{GT} are the prediction and ground truth 4×4 matrices, and (i, j) represents row and column indices of the matrices, respectively, such that $(1 \leq i \leq 3, 1 \leq j \leq 4)$. The last row is not used.

Further, the depth network in Monodepth2 is replaced with GLPDepth (Kim et al., 2022) (Fig. 4(I-b)) and the pose network employs a ResNet18 encoder and a decoder (Fig. 4(I-c)). For Task 2, the depth network is trained from scratch (weights from Task 1 are not used). The team used the hyperparameters proposed in the original MonoDepth2. At training time, as used in Monodepth2, both forward and backward path trajectories are trained, simultaneously. The model is trained for 20 epochs, but the epoch with the best performance on the validation set is submitted. The same model is used for Task 3, where additionally, a CycleGAN was used to translate real images to synthetic images before feeding into the Monodepth2 network.

5.3. Curriculum learning with linear regression by MMLab

For Task 2, MMlab employed ResNet18 (He et al., 2016) and a series of linear layers as shown in Fig. 4(III). Furthermore, they employed Laplacian of Gaussian (LoG) kernel-based filters to enforce attention to contours and perform curriculum

[6] https://github.com/lalithjets/SimCol3D_challenge_2022

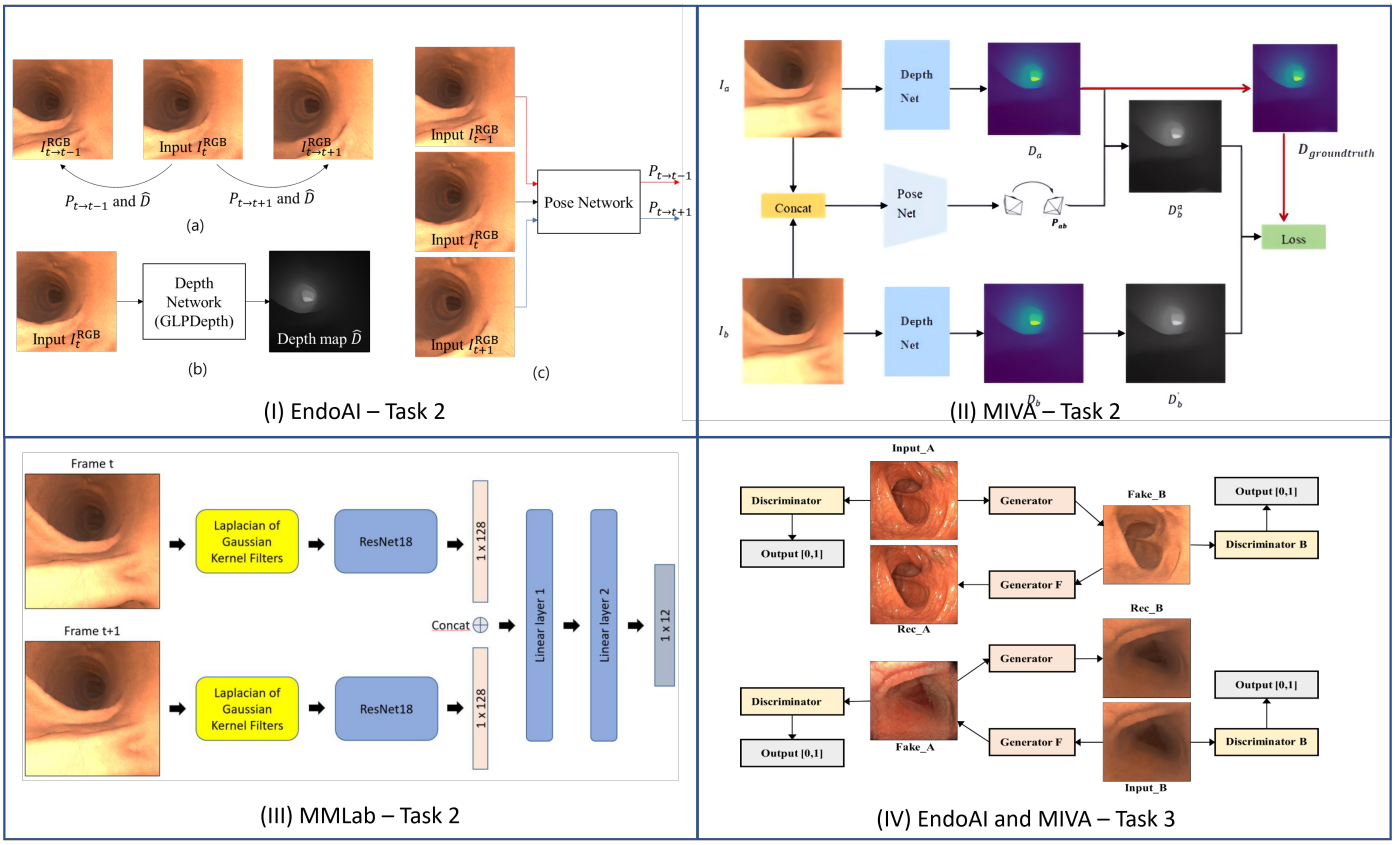

Fig. 4: Architecture overview for Task 2 (pose prediction from synthetic) and Task 3 (pose prediction from real) images of the participating teams. For Task 2, (I) Team EndoAI utilized MonoDepthv2 (Godard et al., 2019), (II) Team MIVA utilized SC-SfMLearner (Bian et al., 2021), and (III) Team MMLab implemented curriculum learning with linear regression. For Task 3, (IV) Team EndoAI and Team MIVA utilized the CycleGAN model for Sim2Real image generation.

learning. Initially, the ResNet module is loaded with the PyTorch ImageNet pre-trained weights. Then the whole model is trained based on mean-square-error (MSE) loss using Adam optimizer with a learning rate of 7.5×10^{-6} for 45 epochs. During training, the values of the LOG kernel (with kernel size = 3) are updated with a factor of 0.9 to allow more features to pass through the model as the learning progresses and to enforce attention to contours. While the relative ground truth pose has 16 values, the module regresses 12 values as the last four values are constant [0.0, 0.0, 0.0, 1.0].

6. Results and discussions

This section summarizes and discusses the submitted results of all participating teams on the three tasks.

6.1. Task 1: Depth estimation

All teams that participated in Task 1 delivered impressive results on the test scenes as presented in Table 4. The L_1 error ranged between 0.03 cm and 0.201 cm across teams and scenes.

Among the three best-performing methods, one method was fully convolutional (MIVA), and the other two were a combination of a convolutional model and transformer (EndoAI, CVML). Achieving sub-millimeter errors on all scenes, team CVML demonstrates that depth prediction from synthetic data can be considered a robustly solvable task. CVML outperformed all other teams on all metrics and all scenes. Even on SynCol III, a scene that has not been seen during training, the average L_1 error of CVML is below one millimeter. The winning team used a model that combines both a transformer-based and a CNN-based branch in a single network. To develop their method, the team performed detailed validation of their backbone model, which inspired their modifications to FCBFormer. The team reported that the addition of the 1×1 convolutional layers to the convolutional branch and the inclusion of the fusion module was instrumental to their method's accuracy. Furthermore, the multi-head self-attention in both the U-Net style architecture in the convolutional branch and the fusion module boosted performance but necessitated replacing the Transformer branch decoder and the prediction head with lightweight alternatives to reduce computational complexity.

Maybe surprisingly, the runner-up method proposed by MIVA is based on a convolutional neural network from 2018 and was applied out of the box, without further adaptations to the method, or complex augmentations or post-processing. The only changes the authors made was replacing the original augmentations with less pronounced endoscopy-suitable augmentations, namely flipping and normalizing only, which might be a good strategy for synthetic colonoscopy frames as their appearance does not vary.

EndoAI's method ranked third and was also a direct adapta-

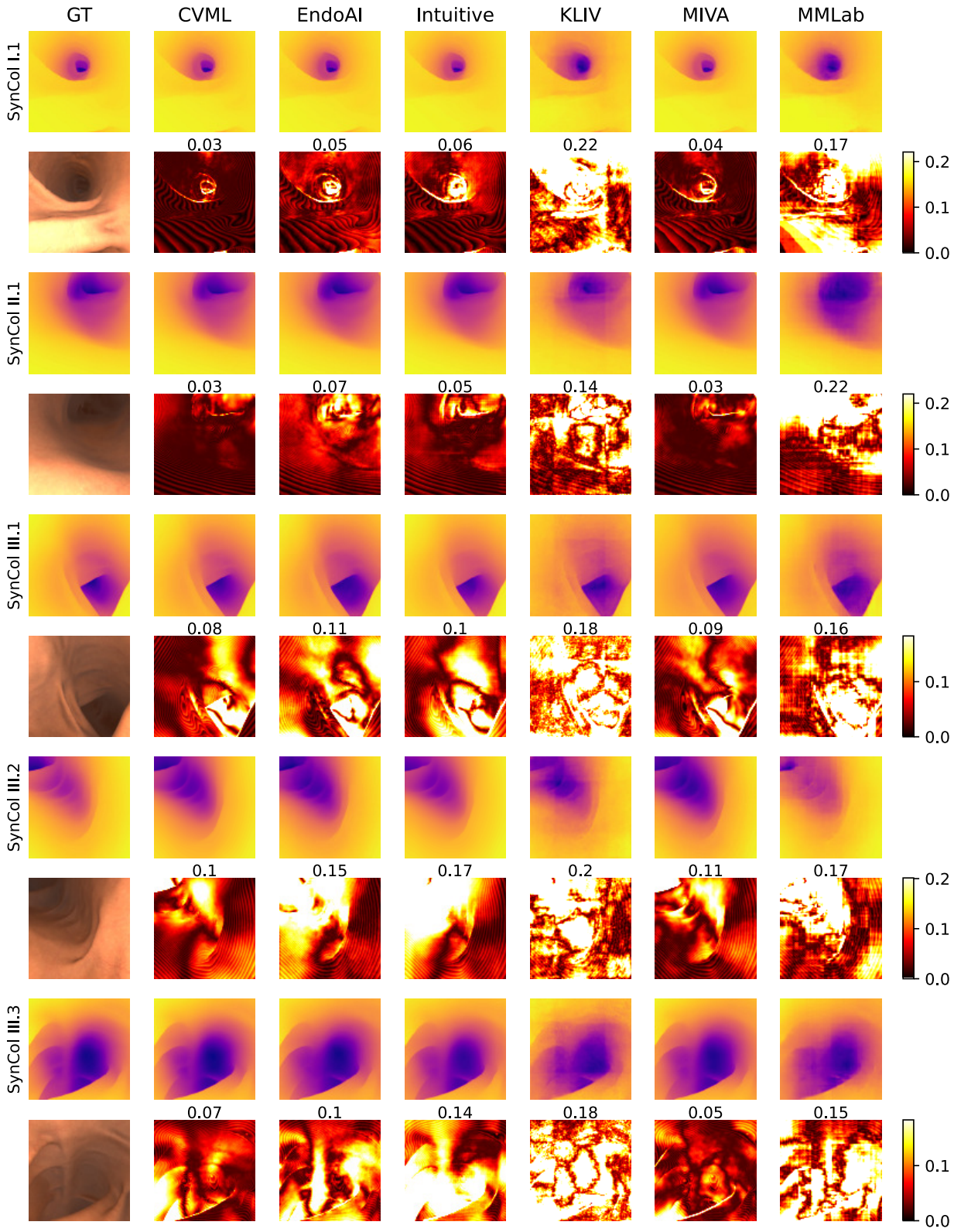

Fig. 5: Comparison of depth predictions generated by the participant teams. For Synthetic Colons I and II we show one example from one test trajectory each. For Synthetic Colon III, we show an example for all three test trajectories. We show the average L1 error above each error map. The colorbar's scale is in cm. Visually, the results of CVML, EndoAI, IntuitiveIL, and MIVA are barely distinguishable from the ground truth. Though when observing the L1 error, CVML is found to be the best performing one, closely followed by MIVA.

Table 4: Task 1 results on the three test scenes. We report the mean over three sequences per test scene. Winners are indicated in bold, the runner-up is underlined, and third-placed teams are shown in italics. Asterisks (*) indicate scenes that provided trajectories with groundtruth for training. All results are reported in cm.

	SynCol I*			SynCol II*			SynCol III			$\sum_1 \uparrow$
	L1 ↓	Rel ↓	RMSE ↓	L1 ↓	Rel ↓	RMSE ↓	L1 ↓	Rel ↓	RMSE ↓	
CVML	**0.030**	**0.012**	**0.045**	**0.030**	**0.009**	**0.044**	**0.099**	**0.025**	**0.141**	**54**
EndoAI	*0.040*	*0.015*	*0.067*	*0.039*	*0.011*	0.063	*0.111*	*0.028*	*0.168*	*37*
IntuitiveIL	0.050	0.017	0.091	0.059	0.016	0.103	0.167	0.047	0.233	26
KLIV	0.155	0.055	0.228	0.166	0.045	0.236	0.187	0.048	0.277	12
MIVA	0.038	0.014	0.065	0.038	0.010	*0.065*	0.107	0.025	0.163	44
MMLAB	0.109	0.037	0.185	0.201	0.047	0.330	0.171	0.040	0.277	16

Table 5: Task 2 results on the three test scenes. We report the mean over three sequences per test scene. Winners are indicated in bold, and the runner-up is underlined. ATE is measured in dm, RTE in cm, and ROT in degrees. Asterisks (*) indicate scenes that provided trajectories with ground truth for training.

	SynCol I*			SynCol II*			SynCol III			$\sum_2$
	ATE ↓	RTE ↓	ROT ↓	ATE ↓	RTE ↓	ROT ↓	ATE ↓	RTE ↓	ROT ↓	
EndoAI	**0.574**	**0.081**	0.144	0.336	**0.084**	**0.148**	**0.325**	0.247	0.367	**0.165**
MIVA	0.860	0.124	**0.141**	**0.325**	0.158	0.180	0.422	**0.226**	**0.275**	0.183
MMLAB	0.819	0.082	2.818	1.206	0.139	1.880	0.572	0.458	1.833	0.284

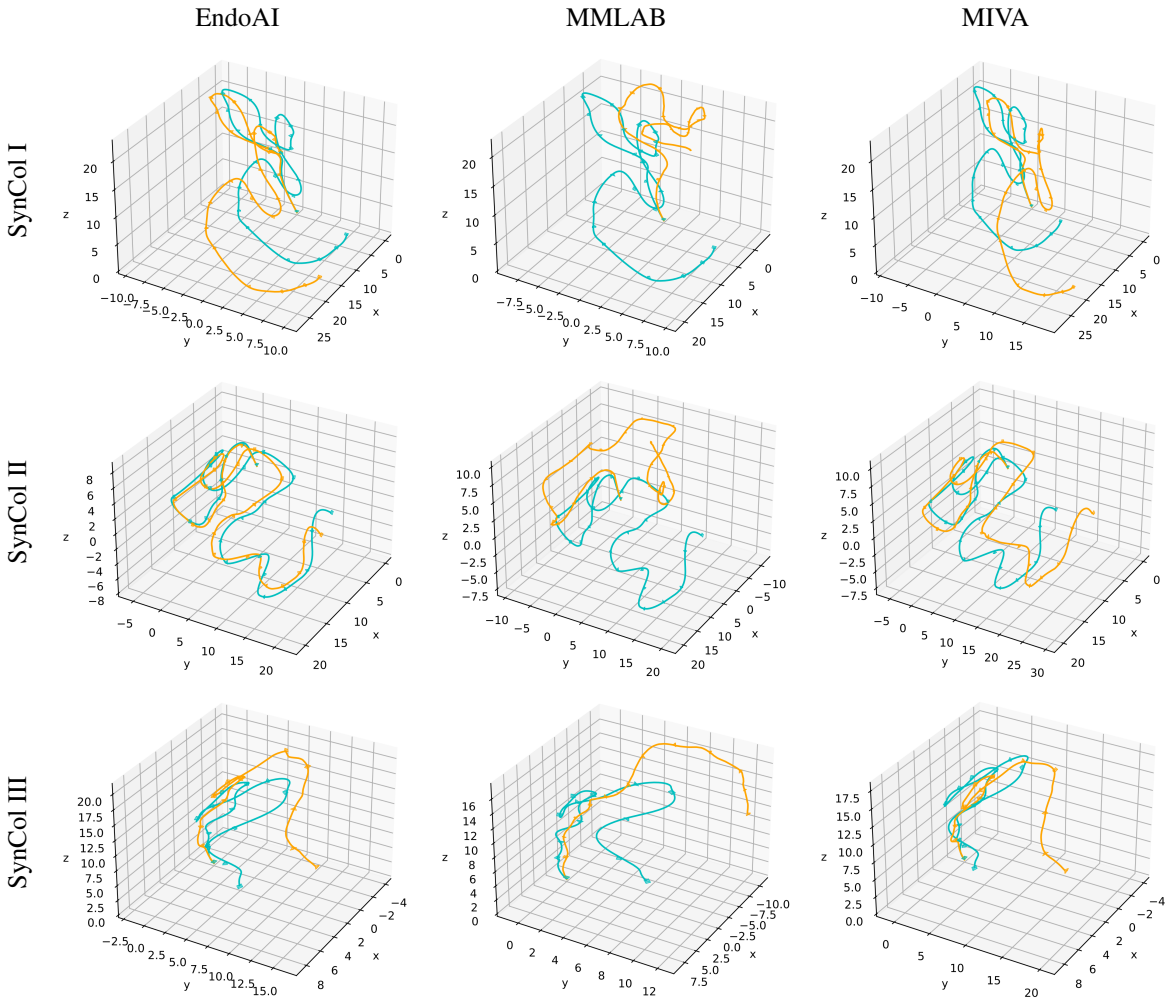

Fig. 6: Overview of Task 2 results showing predicted and ground truth trajectories. For each test scene, we show the first of three trajectories. Qualitatively, it can be observed that Team EndoAI performed the best.

tion of an existing method, but, like the winner, the model is based on a combination of a convolutional model and a Transformer.

Comparing all six methods, we find that all teams used distinctly different baseline methods. Perhaps surprisingly, there was no consensus on the best method for depth prediction during the time of the challenge. Most teams used recent works from 2022 to build upon except for two teams, one of which ranked second. All recent methods were transformer-based, while the older methods are CNNs. Team IntuitiveIL was the only team to develop new augmentation strategies tailored towards colonoscopy applications. The team introducing most changes to a baseline method is CVML which won the first task of the challenge.

Interestingly, the winning method and the last and second-to-last methods employed networks that were initially developed for medical image segmentation. All other teams used networks that were developed for depth prediction. Given the discrepancy between the results, it appears that in this challenge, segmentation models are neither better nor worse than depth prediction networks.

Similarly, fully convolutional networks ranked both second, and last, so that a method's performance cannot be attributed to this design choice alone. Transformer-only networks ranked fourth and fifth, suggesting, that perhaps, the transformers used in this challenge were not equipped to capture the detailed geometry of the endoscopic scenes. Although Vision Transformers have greatly impacted the broad field of computer vision, further investigations into their ability to predict depth from endoscopic images are required. One design, that performed well throughout, is a combination of transformer and convolutional layers. As described by Sanderson et al. (Sanderson and Matuszewski, 2022), who also participated in this challenge as team CVML, the combination of transformer and convolutional layers helps leverage both global and local features in endoscopic images.

Comparing quantitative results in Table 4, we can observe that three best performing methods all lead to similar errors. For instance, on SynCol I, CVML, EndoAI, and MIVA achieve L1 errors of $0.03-0.04$ cm. The other methods perform considerably worse ($0.05-0.16$). Further, all methods perform significantly worse on SynCol III, for which, as opposed to SynCols I and II, there were no training sequences released. Nonetheless, all methods achieve an L1 loss of less than 2mm. This speaks to the ability of these methods to accurately generalize to unseen geometries.

A qualitative comparison of all methods on a few representative images from all three scenes is provided in Figure 5 along with the L1 error of individual predicted masks. We randomly sampled one frame for visualization per trajectory. The results of CVML, EndoAI, IntuitiveIL, and MIVA are barely distinguishable from the ground truth. Only when assessing the individual L1 errors, we can observe that CVML performs slightly better than MIVA, followed by EndoAI and IntuitiveIL. KLIV and MMLab show visible checkerboard artefacts, which is consistent with the quantitative results in Table 4, where KLIV and MMLab rank fifth and sixth.

Table 6: Task 3 results. Winners are indicated in bold. The ROT error is reported in degrees. The absolute scale of the ATE and RTE is unknown.

Sequence	1	2	3	4	5	6	7	$\sum_3$
#Frames/seq	76	144	119	69	127	86	56	
	ATE ↓							ATE ↓
EndoAI	3.34	10.19	7.70	**1.17**	**1.47**	11.58	14.69	7.16
MIVA	**0.97**	**4.50**	**3.12**	2.38	3.55	**4.10**	**6.48**	**3.59**
	RTE ↓							RTE ↓
EndoAI	0.104	0.200	0.142	0.307	0.144	**0.191**	**0.493**	0.23
MIVA	**0.065**	**0.104**	**0.130**	**0.174**	**0.116**	0.300	0.625	**0.22**
	ROT ↓							ROT ↓
EndoAI	0.709	0.960	0.836	0.643	0.776	0.823	1.310	0.87
MIVA	**0.264**	**0.634**	**0.551**	**0.453**	**0.622**	**0.478**	**0.804**	**0.54**

6.2. Task 2: Camera pose estimation on simulated data

Three of the six teams participated in Task 2. The results of these teams are summarized in Table 5. The challenge organizers were particularly interested in the teams' results on the third test scene, as no training trajectories of scene III were provided to the teams. We thus weighted errors on SynCol III twice, while errors on SynCol I and II were weighted once, to reflect the importance of generalizability to unseen scenes. Based on the mean ATE, Team EndoAI performs best and by a large margin on two out of three trajectories and takes first place. EndoAI also performs best on SynCol III, which is the only unseen scene. Based on the RTE, EndoAI performs best on SynCol I and II, but even when weighting results on scene III twice, EndoAI outperforms the other methods. MIVA performs best on all measures in at least one scene, but ranks second overall, followed by MMLAB.

Qualitative results are shown in Figure 6. We chose to show one trajectory per scene only, as the differences between trajectories on one scene are small. It can be observed that EndoAI's predictions most closely follow the ground truth trajectories. All models show clear drift in almost all scenes, which is consistent with the frame-wise approaches all teams chose to follow. Especially scene SynCol III, which was not seen during training, suffers from drift. Notably, the two more accurate approaches are both based on warping-based depth and pose networks (MonoDepth2 with updated backbones and SC-SfMLearner), while the third-placed method regresses pose from images directly. Although the warping-based approaches are optimized for the auxiliary task novel-view synthesis, the networks outperform the approach that minimized the pose loss only. Moreover, the two teams employing warping-based networks added supervised losses based on the provided labels in the training to the respective self-supervision methods. Interestingly, MIVA employed a supervised depth loss, while EndoAI used a supervised pose loss, and neither team used both depth and pose labels. As all teams use different backbones, a concluding comparison study remains to be conducted. We can only speculate that EndoAI's performance might result from their more complex back-bone (Transformer-based depth network) in comparison to MIVA who use a UNet-type depth net. It could also result from their supervision with ground truth poses in addition to the self-supervised losses of Monodepth2.

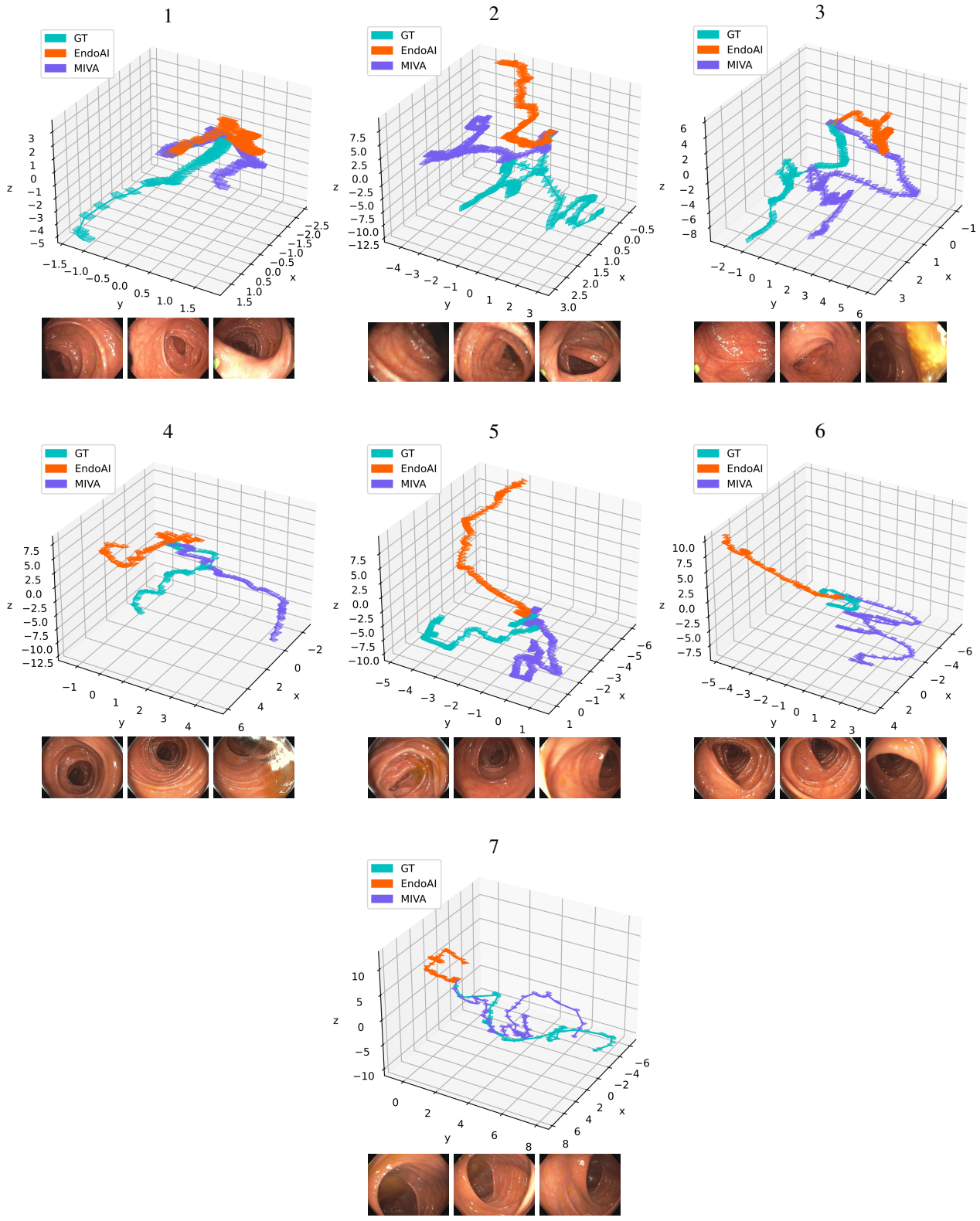

Fig. 7: Overview of Task 3 results showing predicted and ground truth (COLMAP) trajectories. For each test trajectory, we show three sampled frames in order of the video.

6.3. Task 3: Camera pose estimation on real-world data

Two teams participated in Task 3. Both teams used the same method they also used for Task 2, but also applied a CycleGAN to translate appearance between the real and synthetic domains before predicting the camera pose. Results are summarized in Table 6. As methods are compared to COLMAP, and overall scales are not known, the predicted trajectories are scaled before the evaluation of the error metrics. For the same reason, we are more interested in the ATE which better reflects the global consistency than the RTE. MIVA outperforms EndoAI on five out of seven scenes according to the ATE and is thus declared winner. MIVA also yields the smallest RTE in five of seven scenes.

The methods are compared qualitatively in Figure 7. While both methods demonstrate extreme drift, the overall trajectories follow the COLMAP trajectory in some scenes, such as in scene 2, where both methods predict the sharp sideways movement in the first half of the trajectory. Similarly, both models show the quick sideways slip of the camera in the middle of the trajectory in scene 3. And in scene 5, both models follow the "W" shape of the trajectory. We found that none of the participating groups used the publicly available COLMAP poses in the EndoMapper dataset for training. The power of their methods is based entirely on their pose models pretrained on synthetic data. We thus posit that the synthetic data in this challenge provided the models with some understanding of camera pose movement in real colonoscopy. Interestingly, the ranking of both teams is swapped in comparison to Task 2, although both teams use the same pose networks as before, and employ the same CycleGAN for domain adaptation.

6.4. Data limitations and future directions

While synthetic data provides a useful playground to develop algorithms, its applicability to real procedures remains to be elucidated. While we strongly believe that synthetic datasets played a crucial role in enabling early research in the field (Mahmood and Durr, 2018) and in helping push the boundaries further (Rau et al., 2019; Mathew et al., 2020; Itoh et al., 2021; Rodriguez-Puigvert et al., 2022), drawbacks remain. First, the visual discrepancy between real images and our synthetically generated frames is obvious. But visual differences alone can usually be overcome with domain adaptation. More importantly, our synthetic data also misses some physical properties of real colons. For instance, colonoscopists often use water to clean the colon mucosa, resulting in puddles. Specularities and air bubbles are also common in real colonoscopies but are not reflected in our data. For the camera pose dataset, one important difference is the lack of deformation in the synthetic data. In the synthetic dataset, the movement of the colon wall is always due to a camera movement. But in real colonoscopy, the colon walls constantly move due to the colon's own digestive motions, or inflation with air.

So, while synthetic data is useful, the question of how we can move past having to choose between unrealistic synthetic data *or* unlabeled real data remains unanswered. One obvious approach is improving the fidelity of synthetic data to replicate real colon mucosa more closely (Dowrick et al., 2023). However, a domain gap is unavoidable, especially with respect to the behavior of the camera and the relative movement of the colon wall and haustral folds.

A different approach is method-based and focuses on combining both modalities in a useful way Rau et al. (2023). But to evaluate such methods, a real labeled dataset is indispensable.

COLMAP provided useful ground truth poses for this challenge, but the method has serious limitations. It requires reliably matchable features which are extremely sparse in the colon. The reconstruction thus fails on many subsections of the colon. Even if it works, the resulting depth maps are too sparse to be useful, and depths and poses are biased toward a few visible features while ignoring most of the remaining colon wall. Due to the high failure rate, the reconstructions must be visually verified, further biasing the resulting test set towards sub-scenes that are visually interpretable. However, when COLMAP succeeds, it is accurate. We ran COLMAP on Synthetic Colon I and found that it fails to reconstruct 93% of all frames but achieves an RTE of 0.028 cm on the sections where it does not fail. For comparison, the best submission achieved 0.081 cm on the entire Synthetic Colon I.

An alternative route for labeled real datasets could be new hardware. Magnetically actuated soft capsule endoscopes can provide partial ground truth pose, but not depth (Pittiglio et al., 2019). Some capsule colonoscopes provide stereo vision, paving the door for more accurate, but still sparse, depth prediction (Bianchi et al., 2017). Similarly, full spectrum colonoscopy provides two additional lateral cameras (Kurniawan and Keuchel, 2017). While these advances currently focus on improving the visualization of the colonoscopic scenes for the operator in real-time, we hope that future advances incorporate other sensors, such as for position or depth.

A last alternative to synthetic data is colon phantoms made of synthetic materials, such as silicone. Phantoms are, perhaps, the most flexible approach. They can, in theory, be produced in any size, allowing the integration of mounted depth and pose sensors. One drawback of phantoms is their material. Phantoms are either rigid, preventing a colonoscope from moving through it, especially around corners. Or they are non-rigid, rendering electromagnetic poses invalid as the sensor can move relative to the magnetic field while staying in place relative to the phantom. Further, the rubber-like surface looks unrealistic and prevents the camera from replicating realistic camera movements due to friction. As they are expensive to produce, a collection of many phantoms is unrealistic, such that data availability and diversity are limited. Lastly, hand-eye calibration between the camera and EM tracker and temporal synchronization introduces errors in the ground truth. Nonetheless, the creation of cheap and realistic looking and feeling phantoms could be a promising future direction.

7. Conclusions

This paper discusses the SimCol3D 2022 EndoVis Subchallenge and the methods employed by participating teams. The primary objective of this challenge was to promote research on

3D reconstruction during colonoscopy. Six teams from various parts of the world participated in the challenge and achieved impressive results. Particularly, the task of depth prediction on synthetic data proved to be both interesting and solvable. Achieving sub-millimeter accuracy on an unseen colon, the winning team could predict local 3D geometry extremely accurately. This robust generalization to a new scene within the same domain is a promising step towards real-world applications. The generalizability to a new domain remains an open research question and has not been addressed for the depth prediction task in this challenge. To test the applicability to real colonoscopy, new hardware facilitating datasets consisting of real colonoscopy frames with corresponding ground truth depth is required. While synthetic, phantom, and Structure-from-Motion-based data sources all have their own limitations, a thorough evaluation on all three modalities could paint a more holistic picture of model performance in the meantime.

In comparison to depth prediction, predicting pose is a less well-studied problem, and accordingly the task is not yet fully solved. One main concern remains drift, which could be addressed by future work. Interestingly, both depth loss (L_1) and pose loss (RTE) increase roughly three-fold between the know scenes (I and II) versus the unseen scene (III). Therefore, it is crucial for future work to delve deeper into investigating the generalizability of models across different scenes, both within the same domain and across domains. While this challenge was the first one to evaluate generalizability from synthetic pose prediction to real procedures, the evaluation is limited by the quality of the COLMAP labels and their visual verification.

To have an impact on patient outcomes, accurate depth and pose predictions are a first step. Future work should tackle the challenge of achieving robust global reconstructions from local pose and depth predictions based on which unscreened colon mucosa can be identified and visualized. Such a framework will have to work in real-time and should be seamlessly integrate into clinical practice.

Acknowledgments

The authors would like to thank the EndoVis team for facilitating this challenge and providing guidance and support.

This work was supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145Z/16/Z]; Engineering and Physical Sciences Research Council (EPSRC) [EP/P027938/1, EP/R004080/1, EP/P012841/1]; The Royal Academy of Engineering Chair in Emerging Technologies scheme; and the EndoMapper project by Horizon 2020 FET (GA 863146). At the time of the challenge, the three corresponding authors were affiliated with University College London.

The contribution of team CVML from the University of Central Lancashire was supported by the Science and Technology Facilities Council [ST/S005404/1].

The synthetic data supporting this work is openly available under a CC BY license at `https://www.ucl.ac.uk/interventional-surgical-sciences/simcol3d-data`.

Anita Rau, Sophia Bano, Yueming Jin, and Danail Stoyanov organized this challenge. Pablo Azagra, Javier Morlana, and José M.M. Montiel curated the real dataset and generated COLMAP labels. Rawen Kader and Laurence B. Lovat clinically motivated and validated the project. All other authors were participants in the challenge. All co-authors helped write this manuscript.

References

Alhashim, I., Wonka, P., 2018. High quality monocular depth estimation via transfer learning. arXiv preprint arXiv:1812.11941 .

Ali, S., Zhou, F., Bailey, A., Braden, B., East, J.E., Lu, X., Rittscher, J., 2021. A deep learning framework for quality assessment and restoration in video endoscopy. Medical image analysis 68, 101900.

Araghi, M., Soerjomataram, I., Jenkins, M., Brierley, J., Morris, E., Bray, F., Arnold, M., 2019. Global trends in colorectal cancer mortality: projections to the year 2035. International journal of cancer 144, 2992–3000.

Azagra, P., Sostres, C., Ferrandez, Á., Riazuelo, L., Tomasini, C., Barbed, O.L., Morlana, J., Recasens, D., Batlle, V.M., Gómez-Rodríguez, J.J., et al., 2022. Endomapper dataset of complete calibrated endoscopy procedures. arXiv preprint arXiv:2204.14240 .

Bae, G., Budvytis, I., Yeung, C.K., Cipolla, R., 2020. Deep multi-view stereo for dense 3d reconstruction from monocular endoscopic video, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 774–783.

Bian, J., Li, Z., Wang, N., Zhan, H., Shen, C., Cheng, M.M., Reid, I., 2019. Unsupervised scale-consistent depth and ego-motion learning from monocular video. Advances in neural information processing systems 32.

Bian, J.W., Zhan, H., Wang, N., Li, Z., Zhang, L., Shen, C., Cheng, M.M., Reid, I., 2021. Unsupervised scale-consistent depth learning from video. International Journal of Computer Vision 129, 2548–2564.

Bianchi, F., Ciuti, G., Koulaouzidis, A., Arezzo, A., Stoyanov, D., Schostek, S., Oddo, C.M., Menciassi, A., Dario, P., 2017. An innovative robotic platform for magnetically-driven painless colonoscopy. Annals of translational medicine 5.

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

Butterly, L., Robinson, C.M., Anderson, J.C., Weiss, J.E., Goodrich, M., Onega, T.L., Amos, C.I., Beach, M.L., 2014. Serrated and adenomatous polyp detection increases with longer withdrawal time: results from the new hampshire colonoscopy registry. Official journal of the American College of Gastroenterology— ACG 109, 417–426.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2023. Swin-unet: Unet-like pure transformer for medical image segmentation, in: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III, Springer. pp. 205–218.

Carvalho, M., Le Saux, B., Trouvé-Peloux, P., Almansa, A., Champagnat, F., 2018. On regression losses for deep depth estimation, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE. pp. 2915–2919.

Chadebecq, F., Lovat, L.B., Stoyanov, D., 2023. Artificial intelligence and automation in endoscopy and surgery. Nature Reviews Gastroenterology & Hepatology 20, 171–182.

Cheng, K., Ma, Y., Sun, B., Li, Y., Chen, X., 2021. Depth estimation for colonoscopy images with self-supervised learning from videos, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 119–128.

Corley, D.A., Jensen, C.D., Marks, A.R., Zhao, W.K., Lee, J.K., Doubeni, C.A., Zauber, A.G., de Boer, J., Fireman, B.H., Schottinger, J.E., et al., 2014. Adenoma detection rate and risk of colorectal cancer and death. New england journal of medicine 370, 1298–1306.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Dowrick, T., Chen, L., Ramalhinho, J., Puyal, J.G.B., Clarkson, M.J., 2023. Procedurally generated colonoscopy and laparoscopy data for improved model training performance, in: MICCAI Workshop on Data Engineering in Medical Imaging, Springer. pp. 67–77.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems 27.

Freedman, D., Blau, Y., Katzir, L., Aides, A., Shimshoni, I., Veikherman, D., Golany, T., Gordon, A., Corrado, G., Matias, Y., et al., 2020. Detecting deficient coverage in colonoscopies. IEEE Transactions on Medical Imaging 39, 3451–3462.

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

Gordon, A., Li, H., Jonschkowski, R., Angelova, A., 2019. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8977–8986.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

He, X., Zemel, R., Carreira-Perpinan, M., 2004. Multiscale conditional random fields for image labeling, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., pp. II–II. doi:10.1109/CVPR.2004.1315232.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134.

Itoh, H., Oda, M., Mori, Y., Misawa, M., Kudo, S.E., Imai, K., Ito, S., Hotta, K., Takabatake, H., Mori, M., et al., 2021. Unsupervised colonoscopic depth estimation by domain translations with a lambertian-reflection keeping auxiliary task. International Journal of Computer Assisted Radiology and Surgery 16, 989–1001.

Ji, G.P., Chou, Y.C., Fan, D.P., Chen, G., Fu, H., Jha, D., Shao, L., 2021. Progressively normalized self-attention network for video polyp segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer. pp. 142–152.

Kaminski, M.F., et al., 2010. Quality indicators for colonoscopy and the risk of interval cancer. New England Journal of Medicine 362, 1795–1803.

Kim, D., Ga, W., Ahn, P., Joo, D., Chun, S., Kim, J., 2022. Global-local path networks for monocular depth estimation with vertical cutdepth. arXiv preprint arXiv:2201.07436 .

Kurniawan, N., Keuchel, M., 2017. Flexible gastro-intestinal endoscopy—clinical challenges and technical achievements. Computational and structural biotechnology Journal 15, 168–179.

Liu, F., Shen, C., Lin, G., Reid, I., 2015. Learning depth from single monocular images using deep convolutional neural fields. IEEE transactions on Pattern Analysis and Machine Intelligence 38, 2024–2039.

Loshchilov, I., Hutter, F., 2016. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 .

Ma, R., Wang, R., Pizer, S., Rosenman, J., McGill, S.K., Frahm, J.M., 2019. Real-time 3d reconstruction of colonoscopic surfaces for determining missing regions, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 573–582.

Mahmood, F., Durr, N.J., 2018. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. Medical Image Analysis 48, 230–243.

Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., et al., 2022. Surgical data science–from concepts toward clinical translation. Medical image analysis 76, 102306.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., et al., 2020. Bias: Transparent reporting of biomedical image analysis challenges. Medical image analysis 66, 101796.

Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al., 2017. Surgical data science for next-generation interventions. Nature Biomedical Engineering 1, 691–696.

Mathew, S., Nadeem, S., Kumari, S., Kaufman, A., 2020. Augmenting colonoscopy using extended and directional cyclegan for lossy image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4696–4705.

Nandamuri, S., China, D., Mitra, P., Sheet, D., 2019. Sumnet: Fully convolutional model for fast segmentation of anatomical structures in ultrasound volumes, in: 2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019), IEEE. pp. 1729–1732.

Nichol, A.Q., Dhariwal, P., 2021. Improved denoising diffusion probabilistic models, in: International Conference on Machine Learning, PMLR. pp. 8162–8171.

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

Pickhardt, P.J., Nugent, P.A., Mysliwiec, P.A., Choi, J.R., Schindler, W.R., 2004. Location of adenomas missed by optical colonoscopy. Annals of internal medicine 141, 352–359.

Pittiglio, G., Barducci, L., Martin, J.W., Norton, J.C., Avizzano, C.A., Obstein, K.L., Valdastri, P., 2019. Magnetic levitation for soft-tethered capsule colonoscopy actuated with a single permanent magnet: A dynamic control approach. IEEE robotics and automation letters 4, 1224–1231.

Puyal, J.G.B., Brandao, P., Ahmad, O.F., Bhatia, K.K., Toth, D., Kader, R., Lovat, L., Mountney, P., Stoyanov, D., 2022. Polyp detection on video colonoscopy using a hybrid 2d/3d cnn. Medical Image Analysis 82, 102625.

Rau, A., Bhattarai, B., Agapito, L., Stoyanov, D., 2022. Bimodal camera pose prediction for endoscopy. arXiv preprint arXiv:2204.04968 .

Rau, A., Bhattarai, B., Agapito, L., Stoyanov, D., 2023. Task-guided domain gap reduction for monocular depth prediction in endoscopy, in: MICCAI Workshop on Data Engineering in Medical Imaging, Springer. pp. 111–122.

Rau, A., Edwards, P.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. International journal of computer assisted radiology and surgery 14, 1167–1176.

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

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 234–241.

Sanderson, E., Matuszewski, B.J., 2022. Fcn-transformer feature fusion for polyp segmentation, in: Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings, Springer. pp. 892–907.

Saxena, A., Chung, S., Ng, A., 2005. Learning depth from single monocular images, in: Weiss, Y., Schölkopf, B., Platt, J. (Eds.), Advances in Neural Information Processing Systems, MIT Press. URL: https://proceedings.neurips.cc/paper/2005/file/17d8da815fa21c57af9829fb0a869602-Paper.pdf.

Saxena, A., Sun, M., Ng, A.Y., 2008. Make3d: Learning 3d scene structure from a single still image. IEEE transactions on pattern analysis and machine intelligence 31, 824–840.

Wang, R., Pizer, S.M., Frahm, J.M., 2019. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5555–5564.

Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B., 2018. High-resolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8798–8807.

Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2022. Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media 8, 415–424.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems 34, 12077–12090.

Ye, M., Wang, H., Deng, N., Yang, X., Yang, R., 2014. Real-time human pose and shape estimation for virtual try-on using a single commodity depth camera. IEEE transactions on visualization and computer graphics 20, 550–559.

Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P., 2022. New crfs: Neural window fully-connected crfs for monocular depth estimation. arXiv preprint arXiv:2203.01502 .

Zhang, S., Zhao, L., Huang, S., Ye, M., Hao, Q., 2020. A template-based 3d reconstruction of colon structures and textures from stereo colonoscopic images. IEEE Transactions on Medical Robotics and Bionics 3, 85–95.

Zhao, X., Wu, Z., Tan, S., Fan, D.J., Li, Z., Wan, X., Li, G., 2022. Semi-supervised spatial temporal attention network for video polyp segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV, Springer. pp. 456–466.