

A Reference-based Approach for Tumor Size Estimation in Monocular Laparoscopic Videos

Seyed Amir Mousavi^{1,2*}, Francesca Tozzi^{3,5*}, Homin Park^{1,2}, Esla Timothy Anzaku^{1,2}, Matthias Van Liefferinge⁵, Nikdokht Rashidian^{4,5}, Wouter Willaert^{3,5}, and Wesley De Neve^{1,2}

¹ IDLab, ELIS, Ghent University, Ghent, Belgium

² Center for Biosystems and Biotech Data Science, Ghent University Global Campus, Incheon, Korea

{seyedamir.mousavi, homin.park, eslatimothy.anzaku, wesley.deneve}@ghent.ac.kr

³ Department of GI Surgery, Ghent University Hospital, Ghent, Belgium

⁴ Department of HPB Surgery & Liver Transplantation, Ghent University Hospital, Ghent, Belgium

⁵ Department of Human Structure and Repair, Ghent University, Ghent, Belgium
{francesca.tozzi, matthias.vanliefferinge, nikdokht.rashidian, wouter.willaert}@ugent.be

Abstract. Laparoscopic exploration of the abdominal cavity is routinely performed for the diagnosis, assessment, and staging of peritoneal metastasis (PM). Accurately measuring tumor size during this procedure is crucial for prognosis and treatment planning. As conventional approaches for tumor size measurement rely on subjective manual assessments during or after surgery, they stand to benefit from computer assistance. This study proposes a new method for measuring tumor size in laparoscopic monocular videos. Specifically, we introduce a novel mathematical equation that connects the intrinsic parameters of a monocular camera, the surface area of target and reference objects, and their distances to the camera. Furthermore, we combine this equation with an object segmentation model (Mask2Former) and a depth estimation model (MiDaS), creating an end-to-end framework that automates tumor size measurement in monocular laparoscopic videos. We evaluate the proposed method using a laparoscopy dataset comprising 18 videos depicting 76 tumor biopsies, with tumor size measured by surgeons who are experts in laparoscopic surgery. When estimating the size of the various tumors in this dataset, we obtain a Mean Absolute Error (MAE) of 2.44 mm \pm 0.23 mm, demonstrating that the newly proposed method accurately predicts intraoperative tumor size. Our code and the evaluation dataset are publicly available on <https://github.com/amiiiiirrrr/TSEMLV>.

Keywords: Laparoscopy · Monocular Vision · Tumor Size Estimation.

* These authors have contributed equally to this work.

1 Introduction

Peritoneal metastasis (PM) consists of the development of malignant tumors within the peritoneal layer of the abdominal cavity. In controlling the spread of cancer in the peritoneum, conventional approaches such as chemotherapy and surgery may prove ineffective. Laparoscopic exploration of the abdominal cavity is considered the gold standard for diagnosing PM and assessing its extension within the abdomen. Accurate evaluation of the extension of PM enables proper treatment planning and follow-up for this oncological condition [2].

The Peritoneal Cancer Index (PCI) assesses tumoral load, allowing a semi-quantitative evaluation of the evolution of PM. PCI is routinely used during laparoscopic exploration [8,11,13]. Additionally, the effectiveness of chemotherapy can be measured by the accurate determination of tumor size during laparoscopic exploration before and after treatment. Inaccurate tumor size estimation could compromise the evaluation of the effectiveness of chemotherapy in targeting PM, potentially leading to an inadequate treatment approach and unnecessary treatments. This might result in an increased risk of side effects with no additional therapeutic benefit. Moreover, due to the absence of a universally accepted method for determining the size of abdominal tumors, and the fact that clinicians are currently compelled to rely on subjective estimations, this approach stands to benefit from automation, for example, through the adoption of modern computer vision techniques. In this paper, we introduce a comprehensive solution to the problem of tumor size measurement, describing an end-to-end method that objectively assesses tumor size in laparoscopic videos that have been obtained from a monocular camera. In particular, our contributions can be summarized as follows:

- We derived a novel mathematical equation that establishes a relationship between the intrinsic parameters of a monocular camera, the surface area of target and reference objects, and their distances to the camera.
- We combined the derived equation with state-of-the-art object segmentation (through Mask2Former [5]) and depth estimation (through MiDaS [4]). This resulted in an end-to-end method that takes laparoscopic videos as input and outputs tumor size measurements, thereby automating the entire process.
- Mask2Former is trained on 30 videos, 2309 frames, annotated by medical experts for 27 abdominal cavity organs, tumors, and surgical instruments.
- The newly introduced approach is validated using a dataset comprising 18 videos depicting 76 biopsies, with a ground truth created by skilled surgeons.

2 Related Work

The authors of [6] introduce a method for computing 3-D affine measurements from a single perspective view using minimal geometric information, such as the vanishing line of a reference plane and the vanishing point for a non-parallel direction. Andalo *et al.* [3] focus on height measurements in standalone images,

employing a vanishing point detector and demonstrating how to estimate the vertical direction and detect the ground plane vanishing line. However, these methods are effective for man-made environments and struggle with scenes lacking parallel lines or lines not aligned with the coordinate system axis. Indeed, in PM, the complex abdominal cavity structure poses challenges for accurate linear segment extraction, making traditional computer vision techniques impractical for estimating vanishing points in medical imagery.

In the evolving field of GastroIntestinal (GI) endoscopy, recent developments in tumor size measurement include a method proposed by Zhou *et al.* that is using Video Capsule Endoscopy (VCE) frames. Their approach, outlined in [17], employs RGB channels and a Support Vector Machine (SVM) to detect and determine polyp sizes. Additionally, in [12], Oka *et al.* introduce a novel lesion size measurement system that combines an inexpensive optical device with a conventional endoscope. This system measures target lesion sizes by displaying a grid scale on endoscopic images, with the grid width adjusting in real-time based on the distance between the endoscope tip and the lesion. [10] presents a novel system for contactless measurement in endoscopy, specifically focusing on wireless capsule endoscopes (WCE). It employs a deep convolutional image registration method combined with a multi-layer feed-forward neural network to accurately measure lesion locations and sizes within the gastrointestinal tract.

Another study by [7] introduces a virtual tape-measure prototype utilizing a laser line to accurately measure polyp sizes during colonoscopies. While the prototype yields highly accurate results, challenges arise from the quality of the projected laser line. Additionally, [14] presents a novel Structured Light (SL) laser probe embedded into a conventional endoscope for one-shot size measurement of polyps during flexible endoscopy of the stomach. The proposed probe significantly reduces errors in polyp size estimation compared to visual inspection. Indeed, lesions can be precisely measured via laser or light projection techniques, whereas the necessity for supplementary instruments to be attached to the endoscope restricts their suitability for specific surgical procedures.

To function effectively, the aforementioned studies require additional devices or specific conditions in a controlled environment and are not open-sourced for reproduction. In contrast, our method is the first to automate tumor size measurement in PM using only a standard monocular camera typically employed in these surgeries. This eliminates the need for extra equipment or special setups suited for the dynamic environment of the abdominal cavity, which features complex anatomical structures, tissue deformation, and variable lighting conditions, making accurate measurements challenging, especially when combined with irregular camera movements.

3 Methodology

As shown in Fig. 1, this paper proposes a novel method to calculate the surface area of a tumor, our target object, based on the known diameter of a Surgical Instrument (SI), our reference object. In more detail, we employed Mask2Former

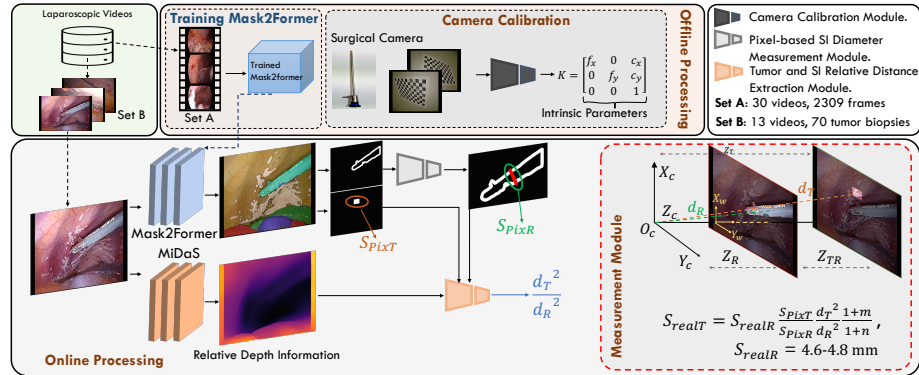


Fig. 1. Overview of the proposed method: Offline processing consists of calibrating the camera to obtain intrinsic parameters and using an annotated Set A to train a segmentation model. Online processing entails the use of a module that computes the relative distance, a module that measures the SI diameter in pixels, and a measurement module that integrates all data to estimate the size of a tumor.

to extract the pixel surface area of the tumor, and then generate a bounding box (S_{PixT}) from the segmentation mask of the tumor that encompasses it. MiDaS is used to determine the relative distance between tumors and the SI. Determining the pixel thickness of the SI (S_{PixR}) in videos is a critical step that is illustrated by the red line in Fig. 1. To do so using OpenCV, the pixels of the SI are first obtained using Mask2Former, and then its contour is generated. Next, the center point of the contour is found. By finding the elongation angle of the SI contour, it is possible to derive a line that passes through the center point of the contour and is perpendicular to the line along the elongation direction of the SI. This line represents the pixel thickness of the SI. In what follows, we establish a relationship between the real-world surfaces of the reference and target objects, their pixel area, and their relative distances to the camera by adopting pinhole camera imaging.

3.1 A Reference-Based Object Size Measurement Model

The authors of [16] introduce a novel method for absolute localization estimation of a target using monocular vision. Specifically, they calculate the absolute distance between the camera and the target by mapping 3-D points in the world to a 2-D image generated through pinhole imaging. The research effort presented in this paper broadens the scope of the method presented in [16], facilitating dimension measurement of the target object using a reference object. To that end, we also rely on pinhole imaging, which maps 3-D objects to a 2-D plane, given its adequacy for measuring objects in world coordinates via camera calibration. We assume the reader has some basic awareness of both pinhole imaging [9] and the method discussed in [16].

Reference object - As shown in Fig. 2, (X_R, Y_R, Z_R) and (X_{wr}, Y_{wr}, Z_{wr}) are two points on the reference plane that correspond to the camera and world coordinate system, respectively. The world coordinate system is positioned on the surface of the reference object, wherein the optical axis meets the reference object plane [16]. That way, the rotation $R = \text{diag}(1, 1, 1)$ and translation $T = (0, 0, Z_R)^T$ matrices are attainable. The conversion of world coordinates to pixel coordinates can be mathematically represented as follows:

$$Z_R \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ O^T & 1 \end{bmatrix} \begin{bmatrix} X_{wr} \\ Y_{wr} \\ Z_{wr} \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x & c_x Z_R \\ 0 & f_y & c_y & c_y Z_R \\ 0 & 0 & 1 & Z_R \end{bmatrix} \begin{bmatrix} X_{wr} \\ Y_{wr} \\ Z_{wr} \\ 1 \end{bmatrix} \quad (1)$$

Note that f denotes the focal length of the camera and (u, v) represents the pixel coordinate frame. In this paper, m_x and m_y are defined as the number of pixels per unit distance along the x and y axes of the pixel coordinates, respectively. $f_x = f m_x$ and $f_y = f m_y$ denote the focal length of the camera in the x and y directions, respectively, and (c_x, c_y) represents the principal point in the image plane that intersects the optical axis. Fig. 2 illustrates that the reference object can be divided into N rectangles along the X_w axis within the scene, creating roughly rectangular pieces. It should also be evident that $Z_w = 0$ for the reference object, and through simplification of Eq. 1, it is feasible to obtain the real surface area S_{realR} of the reference object as follows:

$$\sum_{i=1}^N (P_{2y}^i - P_{1y}^i)(P_{1x}^{i+1} - P_{1x}^i) = \frac{Z_R^2}{f_x f_y} \sum_{i=1}^N (v_2^i - v_1^i)(u_1^{i+1} - u_1^i) = S_{\text{Pix}R} \frac{Z_R^2}{f_x f_y} \quad (2)$$

where $S_{\text{Pix}R}$ is the pixel-wise representation of the area of the reference object on the image plane, and P_{1x}^i and P_{1y}^i represent the coordinates of point P_1^i in the X_w and Y_w directions, respectively, in the world coordinate system. Following this, the absolute distance between the camera and the reference object, $d_R = \|\mathbf{P}_{FR} - \mathbf{O}_c\|$, can be obtained through the establishment of the relationship between a point in the image plane (u_{FR}, v_{FR}) , which is the midpoint of the shape of the reference object in the image, and its corresponding point $P_{FR} = (X_{FR}, Y_{FR})$ in the world plane. This can be expressed as follows:

$$d_R = Z_R \sqrt{(X_{FR})^2 + (Y_{FR})^2 + 1} \quad , \quad \begin{bmatrix} X_{FR} \\ Y_{FR} \\ 1 \end{bmatrix} = Z_R \begin{bmatrix} (u_{FR} - c_x)/f_x \\ (v_{FR} - c_y)/f_y \\ 1/Z_R \end{bmatrix} \quad (3)$$

$$S_{realR} = \frac{d_R^2}{1 + e_r} \frac{S_{\text{Pix}R}}{f_x f_y} \quad , \quad e_r = \left(\frac{u_{FR} - c_x}{f_x}\right)^2 + \left(\frac{v_{FR} - c_y}{f_y}\right)^2 \quad (4)$$

Target object - In what follows, we replicate the above approach for the target object, targeting the derivation of a formula that connects the surface area

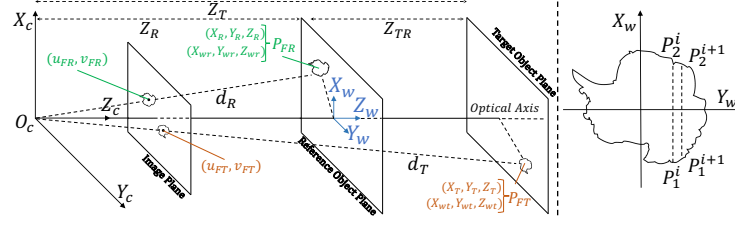


Fig. 2. Geometry for calculating the surface area of the target object from the surface area of the reference object and their respective distances to the projection center.

of the target object to the absolute distance d_T between the target object and the camera center. Assume the points (X_T, Y_T, Z_T) and (X_{wt}, Y_{wt}, Z_{wt}) of the target object belong to the camera and world coordinate system, respectively. These points are transformed into pixel coordinates (u, v) using the pinhole camera model as follows:

$$Z_T \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_T \\ Y_T \\ Z_T \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ O^T & 1 \end{bmatrix} \begin{bmatrix} X_{wt} \\ Y_{wt} \\ Z_{wt} \\ 1 \end{bmatrix} \quad (5)$$

In this context, we maintain $R = \text{diag}(1, 1, 1)$ and $T = (0, 0, Z_R)^T$, given that these matrices are exclusively employed for the purpose of converting world coordinates to camera coordinates, and the world coordinate system was previously fixed on the reference object plane. In this case, however, it should be evident that $Z_{wt} = Z_{TR}$, where Z_{TR} is the distance between the target and reference planes. This is an important difference with the approach followed for the reference object. Consequently, Eq. 5 can be reformulated as follows:

$$Z_T \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x & c_x Z_R \\ 0 & f_y & c_y & c_y Z_R \\ 0 & 0 & 1 & Z_R \end{bmatrix} \begin{bmatrix} X_{wt} \\ Y_{wt} \\ Z_{wt} \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x(Z_{TR} + Z_R) \\ 0 & f_y & c_y(Z_{TR} + Z_R) \\ 0 & 0 & Z_{TR} + Z_R \end{bmatrix} \begin{bmatrix} X_{wt} \\ Y_{wt} \\ 1 \end{bmatrix} \quad (6)$$

By applying Eq. 2 and Eq. 3 to the target object, we can establish a connection between the distance of the target object to the camera center ($d_T = \|\mathbf{P}_{FT} - \mathbf{O}_c\|$), its actual surface size (S_{realT}), its surface size in pixels (S_{pixT}), and the intrinsic parameters of the camera:

$$d_T = Z_T \sqrt{(X_{FT})^2 + (Y_{FT})^2 + 1} \quad , \quad \begin{bmatrix} X_{FT} \\ Y_{FT} \\ 1 \end{bmatrix} = Z_T \begin{bmatrix} (u_{FT} - c_x)/f_x \\ (v_{FT} - c_y)/f_y \\ 1/Z_T \end{bmatrix} \quad (7)$$

$$S_{realT} = \frac{d_T^2}{1 + e_t} \frac{S_{PixT}}{f_x f_y}, \quad e_t = \left(\frac{u_{FT} - c_x}{f_x}\right)^2 + \left(\frac{v_{FT} - c_y}{f_y}\right)^2 \quad (8)$$

where (u_{FT}, v_{FT}) is a point on the target image plane that represents the center of the shape of the target object in the picture, and $P_{FR} = (X_{FR}, Y_{FR})$ is its equivalent point on the target world plane. Finally, through the integration of Eq. 4 and Eq. 8, we can derive a formula that describes a connection between the real-world surface sizes of the reference and the target objects, alongside their corresponding pixel surfaces and distances to the camera:

$$S_{realT} = S_{realR} \frac{S_{PixT} d_T^2}{S_{PixR} d_R^2} \frac{1 + e_r}{1 + e_t} \quad (9)$$

By applying this equation in conjunction with (i) a segmentation model to determine the pixel surface areas of the objects and (ii) a depth estimator model that supplies the relative distances between the SI and the tumor to the camera, we are able to measure the size of a tumor.

4 Experimental Results

To calibrate the surgical camera (5 mm, 30° scopes, Olympus, Hamburg, Germany), multiple checkerboard images affixed to the wall of the surgical room were captured from various angles [15]. Afterwards, the intrinsic camera parameters were calculated using OpenCV: $f_x = 489$, $f_y = 529$, $c_x = 369$, and $c_y = 277$. Three laparoscopic biopsy forceps with diameters of 4.6 mm, 4.7 mm, and 4.8 mm were used. Thirty videos, recorded between June 2020 and March 2023, were selected from two university hospitals and subsequently annotated. The training, validation, and test sets for Mask2Former were composed of 18, 6, and 6 videos containing 1304, 433, and 572 frames, respectively. The entire dataset comprised 27 distinct classes of organs and anatomical structures of the abdominal cavity, including, but not limited to, the liver, stomach, gallbladder, diaphragm, spleen, and bowels. The obtained Overall Accuracy, Mean IoU, Mean Accuracy, Mean Precision, and Mean Recall values across all classes were 77.61, 52.27, 63.94, 70.40, and 63.94, respectively. MiDaS, without any fine-tuning, was solely utilized as an inferencer to provide relative distances. Two highly experienced GI surgeons analyzed 18 laparoscopic videos, each featuring 1-4 biopsy procedures for PM. 76 frames containing a SI and a tumor were randomly chosen. We provide several images, along with corresponding depth information, segmentation outcomes, and the visual output of the newly developed measurement module, in Fig. 3. Fig. 4 displays the estimates produced by the surgeons and our method, with tumor sizes ranging from 1.5 mm to 37.5 mm. Surgeons measured the length of a tumor either along the diagonal, horizontal, or vertical dimension of the bounding box, whereas the proposed method provided measurements for all three dimensions, enabling a comprehensive comparison. The obtained results demonstrate that our method is reliable, yielding an MAE [10] of $2.44 \text{ mm} \pm 0.23 \text{ mm}$.

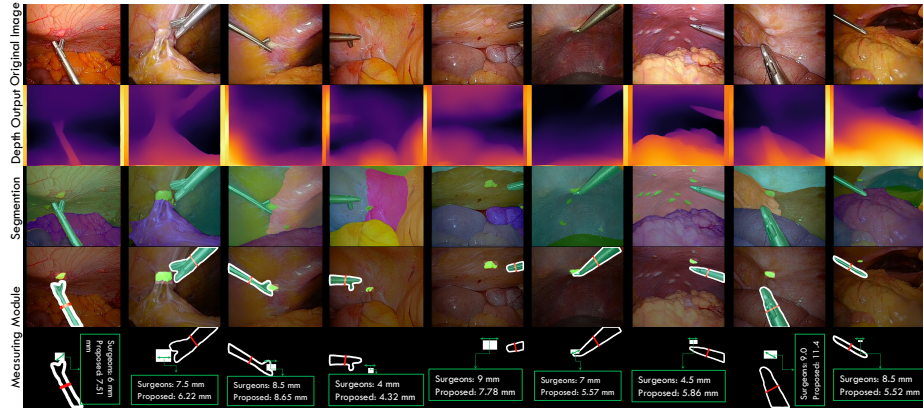


Fig. 3. Visual results: upon acquiring the output of MiDaS and Mask2Former, the measurement module determines the horizontal, vertical, and diagonal tumor length.

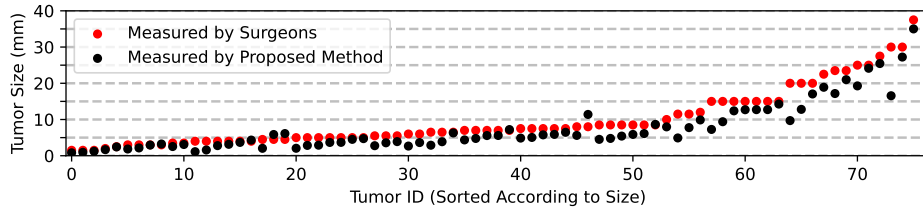


Fig. 4. Differences in tumor size estimates made by surgeons and our method.

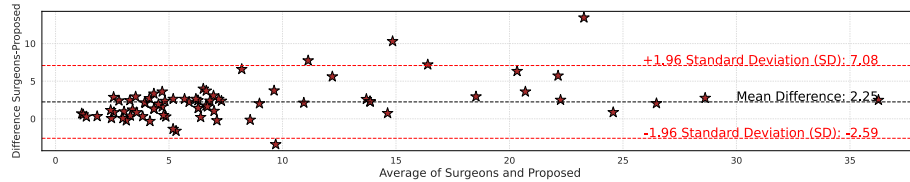


Fig. 5. Agreement between the tumor size estimates made by surgeons and our method.

We also show a Bland-Altman plot [1] in Fig. 5 to assess the agreement between the two employed measurement methods. Over 95% of the data points are within the agreement limits (mean difference ± 1.96 times the standard deviation of the differences). 76.32% of the points are in the mean difference ± 2 range, and there are just four outliers (out of mean difference $\pm 1.96SD$) among the 76 images. This means that the two ways of measuring are strongly in agreement.

5 Conclusions and Future Research

In this paper, we introduced a novel method for automatic size measurement of abdominal tumors in monocular laparoscopic videos. We achieved this by defining a mathematical relationship between target and reference objects, object segmentation, and depth estimation. The obtained experimental results indicate that the newly proposed method for tumor size measurement is accurate, with a MAE of $2.44 \text{ mm} \pm 0.23 \text{ mm}$.

In future research, we aim at determining depth information in the abdominal cavity by adopting self-supervised monocular depth estimation, as state-of-the-art depth estimators are trained on natural images. A second topic for future research is training a deep attention-based convolutional model to measure tumor size at various distances and viewpoints without any reference object.

The assumption that a tumor lies on a plane fronto-parallel to the image plane is a limitation, as it simplifies the problem by treating the tumor as if it is always viewed from a perpendicular angle. In practice, tumors may be viewed from skewed angles due to the positioning of the camera and the anatomy of the abdominal cavity. This can distort the shape and size of the tumor at hand, leading to inaccurate size estimation. Therefore, a final topic for future research involves developing methods to account for these skewed perspectives.

References

1. Altman, D.G., Bland, J.M.: Measurement in medicine: the analysis of method comparison studies. *Journal of the Royal Statistical Society Series D: The Statistician* **32**(3), 307–317 (1983)
2. Alyami, M., Hübner, M., Grass, F., Bakrin, N., Villeneuve, L., Laplace, N., Passot, G., Glehen, O., Kepenekian, V.: Pressurised intraperitoneal aerosol chemotherapy: rationale, evidence, and potential indications. *The Lancet Oncology* **20**(7), e368–e377 (2019)
3. Andaló, F.A., Taubin, G., Goldenstein, S.: Efficient height measurements in single images based on the detection of vanishing points. *Computer Vision and Image Understanding* **138**, 51–60 (2015)
4. Birkl, R., Wofk, D., Müller, M.: MiDaS v3. 1—A Model Zoo for Robust Monocular Relative Depth Estimation. arXiv preprint arXiv:2307.14460 (2023)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1290–1299 (June 2022)
6. Criminisi, A., Reid, I., Zisserman, A.: Single view metrology. *International Journal of Computer Vision* **40**, 123–148 (2000)
7. Goldstein, O., Segol, O., Gross, S.A., Jacob, H., Siersema, P.D.: Novel device for measuring polyp size: an ex vivo animal study. *Gut* pp. gutjnl–2017 (2018)
8. Harmon, R.L., Sugarbaker, P.H.: Prognostic indicators in peritoneal carcinomatosis from gastrointestinal cancer. In: *International Seminars in Surgical Oncology*. vol. 2, pp. 1–10. BioMed Central (2005)
9. Hartley, R., Zisserman, A.: *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK; New York, USA, 2nd edn. (2003)

10. Iakovidis, D.K., Dimas, G., Karargyris, A., Bianchi, F., Ciuti, G., Koulaouzidis, A.: Deep endoscopic visual measurements. *IEEE journal of biomedical and health informatics* **23**(6), 2211–2219 (2018)
11. Jacquet, P., Sugarbaker, P.H.: Clinical research methodologies in diagnosis and staging of patients with peritoneal carcinomatosis. *Peritoneal carcinomatosis: principles of management* pp. 359–374 (1996)
12. Oka, K., Seki, T., Akatsu, T., Wakabayashi, T., Inui, K., Yoshino, J.: Clinical study using novel endoscopic system for measuring size of gastrointestinal lesion. *World Journal of Gastroenterology: WJG* **20**(14), 4050 (2014)
13. Sugarbaker, P.H., Jablonski, K.A.: Prognostic features of 51 colorectal and 130 appendiceal cancer patients with peritoneal carcinomatosis treated by cytoreductive surgery and intraperitoneal chemotherapy. *Annals of surgery* **221**(2), 124 (1995)
14. Visentini-Scarzanella, M., Kawasaki, H., Furukawa, R., Bonino, M.A., Arolfo, S., Secco, G.L., Arezzo, A., Menciassi, A., Dario, P., Ciuti, G.: A structured light laser probe for gastrointestinal polyp size measurement: a preliminary comparative study. *Endoscopy International Open* **6**(05), E602–E609 (2018)
15. Zhang, Z.: A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence* **22**(11), 1330–1334 (2000)
16. Zhang, Z., Han, Y., Zhou, Y., Dai, M.: A novel absolute localization estimation of a target with monocular vision. *Optik* **124**(12), 1218–1223 (2013)
17. Zhou, M., Bao, G., Geng, Y., Alkandari, B., Li, X.: Polyp detection and radius measurement in small intestine using video capsule endoscopy. In: 2014 7th International Conference on Biomedical Engineering and Informatics. pp. 237–241. IEEE (2014)