



Neural Radiance Field Based 3D View Reconstruction for Gastrointestinal Tract Surgery Planning

Xiaohong W. Gao¹ · Annisa Ristya Rahmanti² · Barbara Braden³

Received: 3 September 2025 / Accepted: 21 May 2026
© The Author(s) 2026

Abstract

Gastrointestinal (GI) tract is a 9-meter-long food passage that transports food and nutrition to the human body. Due to the narrow space (~2 cm in diameter) of the GI tract, a lesion removal surgery through endoscopic procedures faces limited view of the lesion, leading to the challenge of maximum removal of diseased lesion and minimum sacrifice of healthy tissues. This paper evaluates the application of the state-of-the-art AI technique of neural radiance field to reconstruct 3D models of a lesion based on 2D endoscopic videos. In this way, a full 3D view of concerned GI lesion can be presented to surgeons to allow comprehensive surgical planning. The advanced AI techniques, neural radiance field (NeRF), structure-from-motion and multi-view stereo are applied. This system is implemented based on COLMAP and Nerfstudio libraries. The system is evaluated using two sets of video clips containing 2600 images. Initial results illustrate that this end-to-end deep learning architecture, i.e. from 2D video input to 3D model output, presents considerable potential for reconstruction of GI lesions. The similarity measures of SSIM, PSNR and LPIPS between original (ground truth) and rendered images are of 19.46 ± 2.56 , 0.70 ± 0.054 , and 0.49 ± 0.05 respectively. This work contributes to the solution towards one of the long-standing challenges in computer vision field, which is to construct 3D views from 2D videos in the field of GI surgery planning and has achieved a promising performance. Future work includes enlarging datasets and removal of ghostly artefact from rendered images.

Keywords Deep learning · 3D view reconstruction · NeRFs · Gastrointestinal tract · SfM · 2D endoscopic video

Introduction

This paper is the extension of the work published at BIO-IMAGING 2025 [1]. Gastrointestinal tract (GI) cancers (oesophagus, stomach, bowel) were responsible for 26.3% of cancer cases and 35.4% of deaths worldwide in 2018 [2]. As the 2nd largest death caused by cancer in the world (after lung cancer), GI cancers have very low 5-year survival rate

(< 20%). At present, the only curative and most effective treatment for early GI cancer or lesion is the removal of concerned lesion endoscopically, especially, for a lesion confined to the mucosal layer, the surface columnar epithelium and the first of the four layers on the GI wall. In this procedure, a substance is injected first under the target to act as a cushion. Then a surgical plan by marking cutting lines is conducted. Finally, the dissection takes place at submucosal layer under the concerned lesion following the planned boundary. The key to success of this surgery is that the endoscopists have a clear view of the lesion, the planned lines and surroundings from varying view angles throughout in order to perform precise dissection.

The challenges here are that all the views are confined into a narrow (~ 2 cm in diameter) cylindrical food path or tube along where the endoscopic camera travels in one direction, resulting in some concerned tissues, anatomy and planned boundary being invisible. In addition, in this complex surgical scene, the endoscopist/surgeon/clinician has to contend with various motions coming from respiration,

✉ Xiaohong W. Gao
x.gao@mdx.ac.uk

Annisa Ristya Rahmanti
annisaristya@ugm.ac.id

Barbara Braden
barbara.braden@ukmuenster.de

¹ Middlesex University, London NW4 4BT, UK

² Universitas Gadjah Mada, Yogyakarta, Indonesia 55281

³ Medical Department B, University of Münster, Münster, Germany

Fig. 1 An illustration of endoscopic procedure for inspecting oesophagus, the upper GI tract. **a** Probe tool to be inserted into the mouth. **b** The diagram showing the passageway of oesophagus. **c** A endoscopic video images displayed in montage form. The top tow row images are collected when the white light is shining. The bottom two rows, also called narrow band images, are the images when the light is shone at wavelengths of 415 nm (green) and 540 nm (blue)

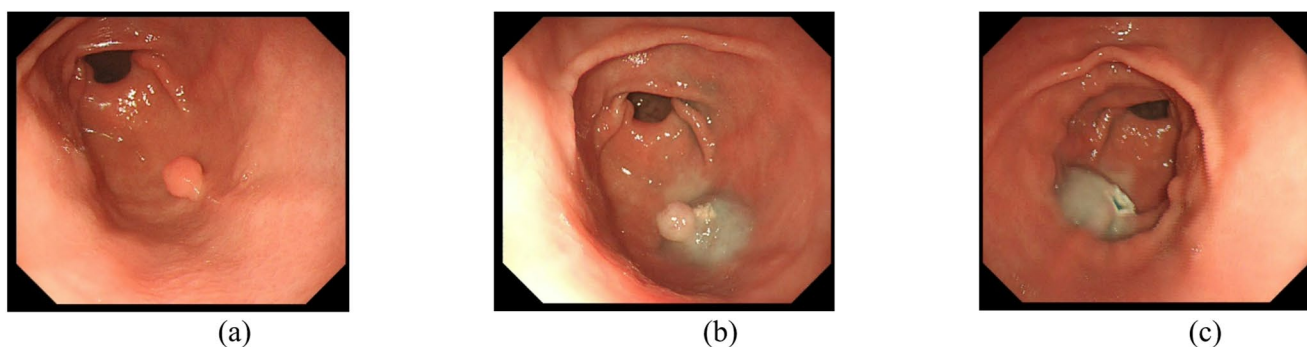
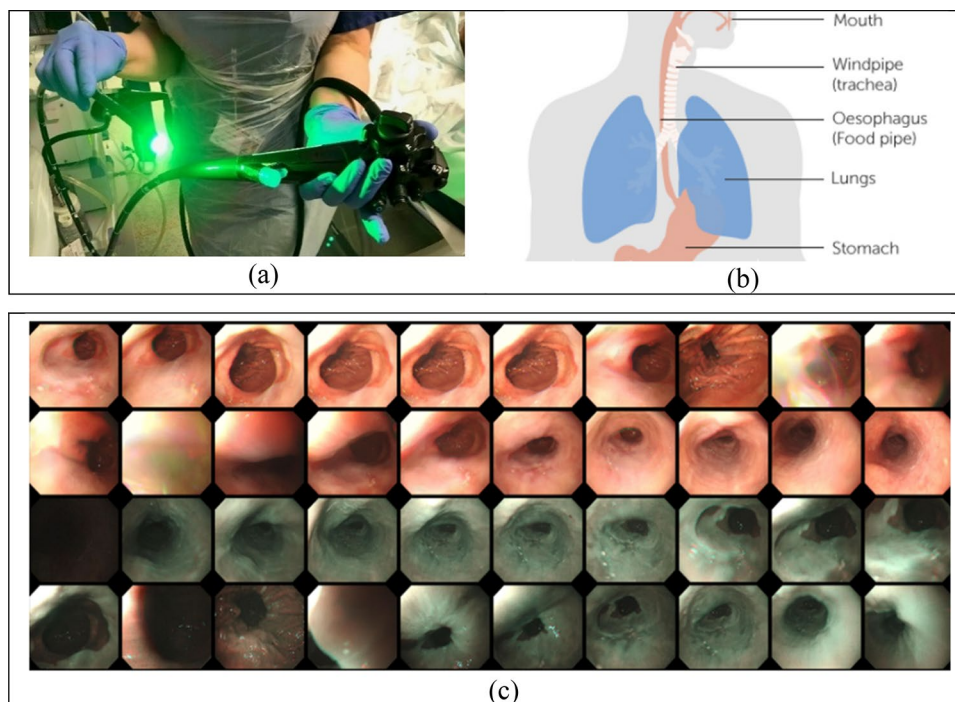


Fig. 2 Demonstration of lesion resection endoscopically. **a** A lesion is detected; **b** cushion is injected; **c** successful removal of the concerned lesion [1]

heartbeat, camera and muscles. Because of this, at present, all these operations can only be conducted by expert endoscopists, which puts significant amount of pressure on health care systems. Figure 1 illustrates the endoscopic procedure when a probe (a) consisted of LED light, camera and tools for collecting biopsy samples is inserted into the mouth (b). Figure 1(c) presents the montage of endoscopic video images with the top 2 rows collected under conventional white light whereas the bottom two rows are under the blue (415 nm) and green (540 nm) lighting conditions, showing the narrow band (NB) images.

Figure 2 demonstrates the process of resection of a polyp in the stomach endoscopically. While the lesion in Fig. 2

is benign, if left untreated, it could progress into cancer. Hence resection is in need. Once the target is confirmed (Fig. 2 (a)), an injection of a dedicated substance is carried out to highlight and alleviate the lesion whereby a surgical planning can be made (Fig. 2(b)). Then the lesion can be removed safely (Fig. 2(c)).

Hence reconstruction of 3D view of concerned lesions plays an important role in endoscopic surgical planning and lesion removal. 3D reconstruction for GI from endoscopic videos has been studied by a number of researchers. For example, Ali et al. [3] has established a physical model of oesophagus that is applied to develop an AI system to

quantify Barrett's oesophagus. The researchers as detailed in [4] has established the shape of oesophagus to allow visualization of additional contextual and geometrical information of oesophagus, from panorama image to 3D points cloud then to regular triangulation mesh. Recently, 3D shape reconstruction of whole stomach based on structure-from-motion (SfM) [5] is investigated by spreading indigo carmine (IC) dye on the stomach surface to present colour texture of the stomach. This is because endoscopic videos present weak texture of GI surface. On the other hand, the approach of SfM appears to present robust results when it comes to 3D reconstruction from 2D videos. Further study for deep multi-view stereo for dense 3D reconstruction [6] is also conducted based on SfM and is consisted of 3 steps, which are sparse reconstruction via SfM, monocular depth estimation and embedding vector generation via patch embedding network.

The steps to construct a 3D model from 2D images usually include image collection, feature or/and keypoint detection and extraction, key point tracking and matching, structure from motion to determine camera intrinsic, extrinsic and orientation parameters, and key point-cloud reconstruction, such as mesh reconstruction, mesh refinement and mesh texturing.

With the current advances of state-of-the-art (SOTA) deep learning (DL) techniques, many innovative approaches have been developed towards reconstructing 3D deformable objects.

In this study, 3D scene reconstruction based on 2D endoscopic videos is conducted based on SOTA neural radiance field (NeRF) so that a lesion can be viewed from all viewing angles, allowing correct recognition of concerned lesions,

tissue types and related anatomy, leading to an assistant system in an operative room allowing multiple views while performing lesion removal.

NeRF [7] addresses the long-standing problem in computer vision field, which is to reconstruct a 3D representation of a scene from sparse 2D images. NeRF method synthesizes a new view by directly optimising parameters of a continuous 5D scene representation to minimize the error of rendering a set of captured images. NeRF represents a scene using fully connected deep network with an input of a single continuous 5D coordinate (i.e. 3D spatial + 2D view direction angles). The output of this network is the volume density and view-dependent emitted radiance at that spatial location. More recently, NeRF has been enhanced into Uncertainty-aware Conditional NeRF (UC-NeRF) for the application to GI tract for addressing the challenge of endoscopic sparse views [8]. This work has improved the scale consistency through the sparse SfM depth constraints significantly. Since the modelling process takes place on still images, UC-NeRF is limited to static or semi-static surgical scenes.

In this study, the training of enhance NeRF takes into account of both spatial and temporal information hence has potential to be applied in real time.

Methodology

Image Pre-process to Remove Artefact of Colour Misalignment

Due to the confinement of narrow space of GI, it is quite common that as many as a quarter of video frames contain several types of artefacts, such as water bubbles, instrument, and saturation. While these artefacts are present most of the time, the concerned image features are still visible. However, the artefact of colour misalignment, where coloured frames of red, green and blue are acquired at different locations resulted from the combination of movements when an endo camera travels, should be removed not only because most of interested contents are not present but also the presence of these artificial colours will influence the detection of camera light of rays, hence affecting the accuracy of training. Figure 3 depicts a number of artefacts presented in a video clip, where the top row of artefacts (green arrows) still contains visible GI features but bottom row of colour misalignment (whole images) mis-represents GI contents.

To detect artefacts, many deep learning-based systems are developed offering promising performance [9–11]. In this study, the processing time also plays an important role for the future development of real-time 3D systems. Hence the real-time system, real-time instance segmentation system,

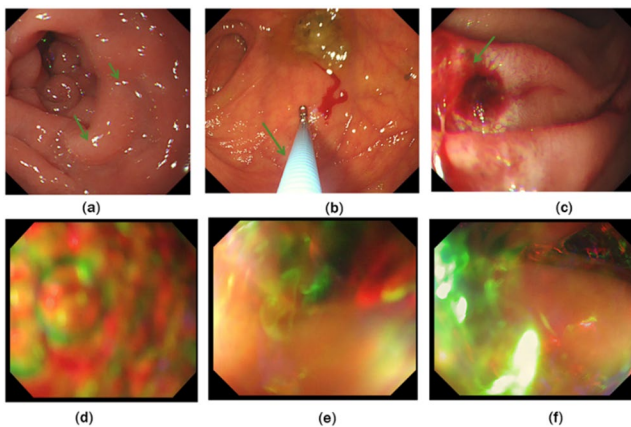
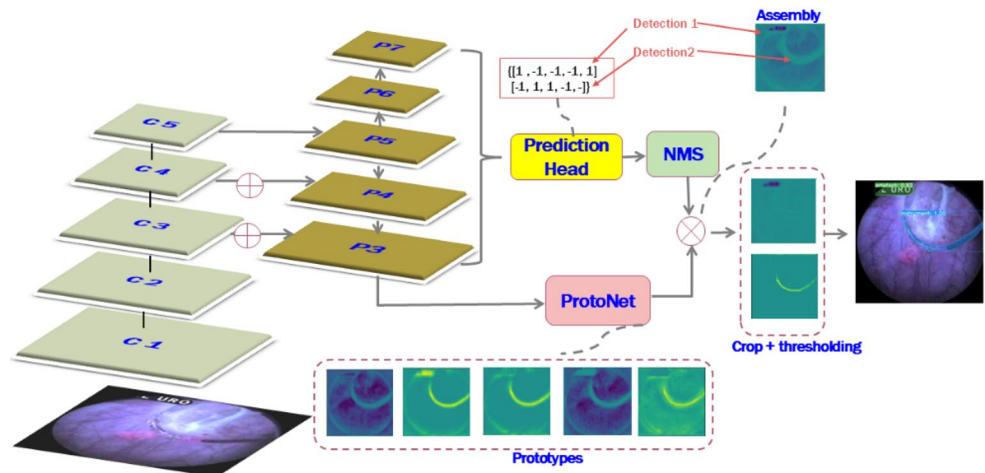


Fig. 3 Examples of frames with artefacts (arrows). **a** specularity; **b** instrument; **c** bleeding; **d–f** colour misalignment (all over images)

Fig. 4 The architecture of YOLACT network for detection and segmentation of artefact. The artefacts to distinguish are instrument, bubbles, saturation, and colour misalignment (CMA). The mask or ground truth of frames with CMA is the whole frame



YOLACT [12, 13] is enhanced and applied for artefact classification. Figure 4 presents the architecture of the network to classify frames with normal or artefact features (instrument, bubbles, saturation or colour misalignment (CMA)). In this pre-processing stage, only frames with CMA will be discarded. This is because the subsequent 3D modelling and reconstruction of lesioned GI regions depends on the information of colour and intensity attributes, i.e. structure from motion and neutral radiance fields.

For this end-to-end detection system of YOLACT [12] (Fig. 4), the basic underline model employs ResNet101 to extract initial feature maps. The object segmentation is accomplished through two parallel subnets (*ProtoNet* and *Prediction Head*), which generate a set of prototype masks and predict per-object mask coefficients respectively.

More specifically, *ProtoNet* employs a fully connected network (FCN) accommodating the largest pyramid feature layer (P_3), to produce a set of image-sized *prototype masks*. These k mask prototypes ($k = 32$ in this study) are then applied to deliver predictions for the entire image in relation to classification, segmentation and detection [13].

On the other hand, *Prediction Head* contains three branches, which are c *class confidence* ($c = 5$ for ‘Instrument’, ‘Bubbles’, ‘Saturation’, ‘artefact-text’, ‘CMA’), 4 *bounding box regressors* ($= [x_{top-left-corner}, y_{top-left-corner}, width, height]$), and a vector of mask coefficients, one for each prototype to be processed in parallel. Subsequently, the branch of ‘*Crop+thresholding*’ in Fig. 4 delivers a vector size of $4 + c + k$ for each anchor or region of interest (RoI). As a result, for each instance, one or more masks will stem from that instance by linearly combining (plus or minus) the work from both *prototype* and *mask coefficient* strands, leading to the production of final masks (M) by a sigmoid nonlinearity as formulated in Eq. (1).

$$M = \sigma (PC^T) \quad (1)$$

where P is an $h \times w \times k$ matrix of prototype masks and C is a $n \times k$ matrix of mask coefficients for n instances that have passed *score thresholding* and initial *NMS*, the maxima suppression technique [11]. NMS determines whether an instance should be kept or discarded. For example, duplicated detections are suppressed not only for the same class but also for cross-class boundary boxes depending on the probability of boxes, i.e. the box with higher probability suppresses the one with lower probability. In Eq. (1), C^T indicates the transpose of C Matrix.

The calculation of the loss function is the same as for YOLACT [13]. Three loss functions are utilised to train this end-to-end detection model as formulated in Eq. (2), which are classification loss ($\mathcal{L}_{class}$), box regression loss ($\mathcal{L}_{box}$) and mask loss ($\mathcal{L}_{mask}$) where the weights of 1, 1.5, and 1.5 are applied for them respectively, which are obtained empirically.

$$\mathcal{L} = \mathcal{L}_{class} + 1.5 \mathcal{L}_{box} + 1.5 \mathcal{L}_{mask} \quad (2)$$

In particular,

$$\mathcal{L}_{mask} = BCE(M, M_{gt}) \quad (3)$$

where the binary cross entropy *BCE* is formulated using Eq. (4).

$$BCE(p, y) = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (4)$$

where y represents the label and p is the predicted probability of the point being a label for all N points. M and M_{gt} are calculated in Eq. (1).

Empirically, if a frame contains an accumulated artefact occupying 60% or more region of a frame, it will be removed. The remaining ones are applied to train the 3D model takes place based on NeRF.

3D View Reconstruction for Concerned Lesioned GI Based on NeRF

For 3D scene modelling, neural radiance fields (NeRFs) [7] take 5-degree coordinates as an input. The 5D refers to each 3D point at (x, y, z) when viewing with a camera ray of light emitting direction at (θ, ϕ) . Hence, NeRF enables learning novel view synthesis, scene geometry and reflectance properties by optimising a deep fully-connected neural network as a multilayer perceptron (MLP). As such, NeRF represents this 5D function by regressing from a single 5D coordinate (x, y, z, θ, ϕ) to a single volume density and view-dependent RGB colour. As a result, to render NeRF from a specific viewpoint, camera rays are marched through the scene, and a neural network produces colours and densities for 3D points, which are then accumulated into a 2D image.

Figure 5 schematically illustrates the process of representing scenes as neural radiance fields for view synthesis.

Firstly, a scene (Fig. 5(a)) is represented using a 5D vector-valued function with an input of a 3D location $\mathbf{x} = (x, y, z)$ and 2D viewing direction $\mathbf{d} = (\theta, \phi)$. The output of this function is an emitted colour $\mathbf{c} = (r, g, b)$ and volume density (Fig. 5(b)) at each ray. This 5D function is approximated by applying a multilayer perceptron (MLP) network $F_{\Theta} : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$ to optimise its weights in order to map each input 5D coordinate to its corresponding volume density and directional emitted colour. Finally, based on the classical volume rendering approach [14], the colour of any light ray passing through the scene is rendered (Fig. 5(c)) whereas the volume density $\sigma(\mathbf{x})$ is interpreted as the differential probability of a ray terminating at location $\mathbf{x}$. Hence the expected colour $C(\mathbf{r})$ of camera ray light $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, starting at the original location $\mathbf{o}$, with near and far bounds t_n and t_f , is expressed in Eq. (5) [6].

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t)) c(\mathbf{r}(t), \mathbf{d}) dt \quad (5)$$

where

$$T(t) = \exp\left(-\int_{t_n}^t \sigma(\mathbf{r}(s)) ds\right) \quad (6)$$

As pointed out by Mildenhall et al. [7], operating a network F_{Θ} directly on $xyz\theta\phi$ can result in poorly rendering performance when colour and geometry have high-frequency variations. Hence, F_{Θ} is formulated as a composite of two functions, i.e., $F_{\Theta} = F'_{\Theta} \circ \gamma$, where F'_{Θ} is learned applying a regular MLP network, from which the estimated colour $\hat{C}_c(\mathbf{r})$ can be expressed in Eq. (7) as a weighted sum of all sampled colours c_i along the ray.

$$\hat{C}_c(\mathbf{r}) = \sum_{i=1}^{N_c} w_i c_i \quad (7)$$

where

$$w_i = T_i(1 - \exp(-\sigma_i \delta_i)) \quad (8)$$

$$T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) \quad (9)$$

and

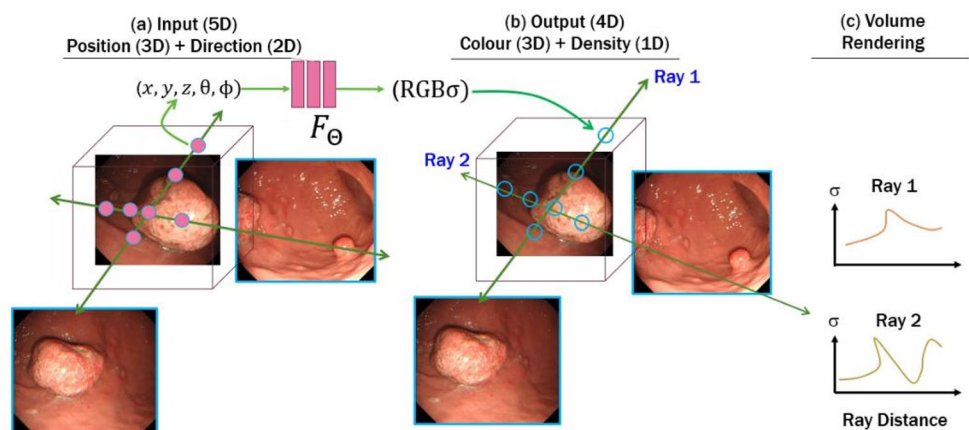
$$\delta_i = t_{i+1} - t_i \quad (10)$$

δ_i in Eq. (10) refers to the distance between adjacent samples.

On the other hand, γ is not learnt but a mapping from $\mathbb{R}$ space into a higher dimensional space $\mathbb{R}^{2L}$ as computed in Eq. (11).

$$\gamma(p) = (\sin(2^0 \pi p), \cos(2^0 \pi p), \dots, \sin(2^{L-1} \pi p), \cos(2^{L-1} \pi p)) \quad (11)$$

Fig. 5 The process flow of NeRF. **a** 5D input; **b** 4D output; **c** volume rendering at a new viewing direction [1]



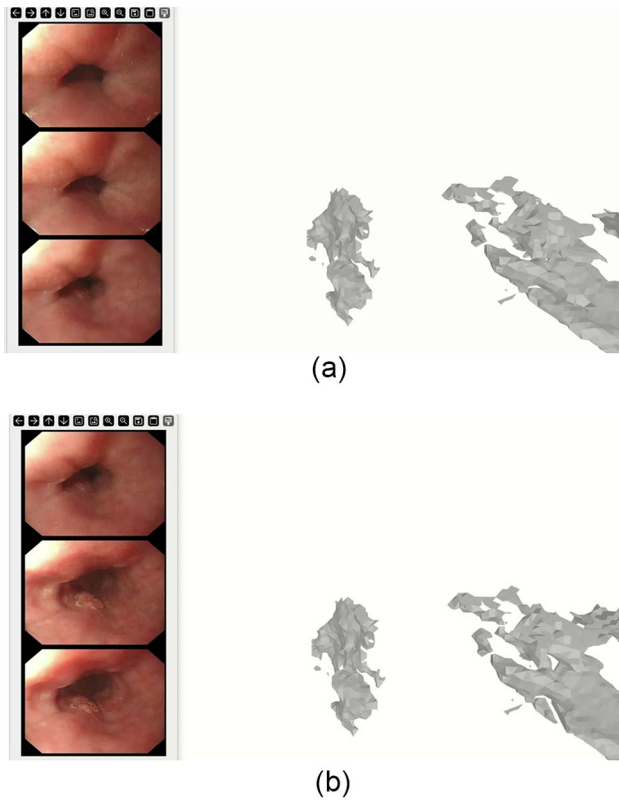


Fig. 6 Demonstration of two processes (a, b) of generating point cloud points from every 3 consecutive video frames for concerned lesions

where $p = x, y, z, \theta, \phi$ respectively. $L = 10$ when $p = x, y, z, \theta, \phi$ and $L = 4$ when $p = \theta, \phi$.

Hence the loss between ground true colour $C(r)$ and predicted pixel colours for both coarse $\hat{C}_c(r)$ and fine $\hat{C}_f(r)$ renderings is calculated in Eq. (12).

$$\mathcal{L} = \sum_{r \in \mathcal{R}} [\|\hat{C}_c(r) - C(r)\|_2^2 + \|\hat{C}_f(r) - C(r)\|_2^2] \quad (12)$$

where $\mathcal{R}$ is the set of rays in each batch.

Implementation

To model 3D view of endoscopic view based on NeRFs, Nerfstudio [15] is implemented. Nerfstudio tools are a repository collecting a family of simple python-based extensive application programming interface (API) functions to allow the visualisation of modelling scenes based on neural radiance fields (NeRFs). By providing a simplified end-to-end process of creating, training and testing NeRFs, these APIs supports viewing and interacting with these processes through an internet browser.

In this study, the input data set is a clip of endoscopic video containing RGB frames. After removal of artefact of CMA as explained in “Image Pre-process to Remove

Artefact of Colour Misalignment”, these frames/images are analysed to extract needed information, including ground truth information such as camera intrinsic and extrinsic data. To obtain endoscopic camera information, COLMAP system is employed. COLMAP [16] is the structure-from-motion (SfM) package that can be employed to estimate the ground truth information regarding to camera poses, camera intrinsic parameters, and scene boundaries. In addition, this pre-processing stage converts those input images into a format that is compatible with NerfStudio¹, i.e. a JSON format.

Then training takes place based on pre-processed images to create a configuration file and a model.

Similarity Measurements

Three common measures are employed to calculate the similarity between original (ground truth) and rendered images, which are structural similarity (SSIM) [17] (Eq. (13)), peak signal-to-noise ratio (PSNR) (Eq. (14)), and more recently Learned Perceptual Image Patch Similarity (LPIPS) (Eq. (15)).

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{x,y} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (13)$$

In Eq. (13) of SSIM, μ_x, μ_y are the averages of x, y , with σ_x^2, σ_y^2 being the variances of x, y respectively and $\sigma_{x,y}$ the covariance of x and y . The variables of c_1 and c_2 are applied to stabilize the division when a small denominator occurs and are set to be $(0.01L)^2$ and $(0.03L)^2$ respectively, whereby L stands for the dynamic intensity range of an image, e.g. $L=255$ for an 8-bit image. x, y refer to original and rendered images respectively.

For PSNR, Eq. (14) is calculated.

$$PSNR = 20\log_{10}MAX_I - 10\log_{10}MSE \quad (14)$$

where MAX_I refers to the maximum possible value of the image (e.g. 255 for 8-bit) and MSE the mean squared error between two concerned images.

In addition, LPIPS [18] metric refers to Learned Perceptual Image Patch Similarity and is formulated in Eq. (15). LPIPS is calculated by comparing the activations of two image patches using pre-defined neural network features. Specifically, it computes the distance between the feature representations of the patches. The lower the LPIPS score is, the more perceptually similar the patches are. The distance between reference and rendered patches x, x_0 with

¹ <https://docs.nerf.studio>.

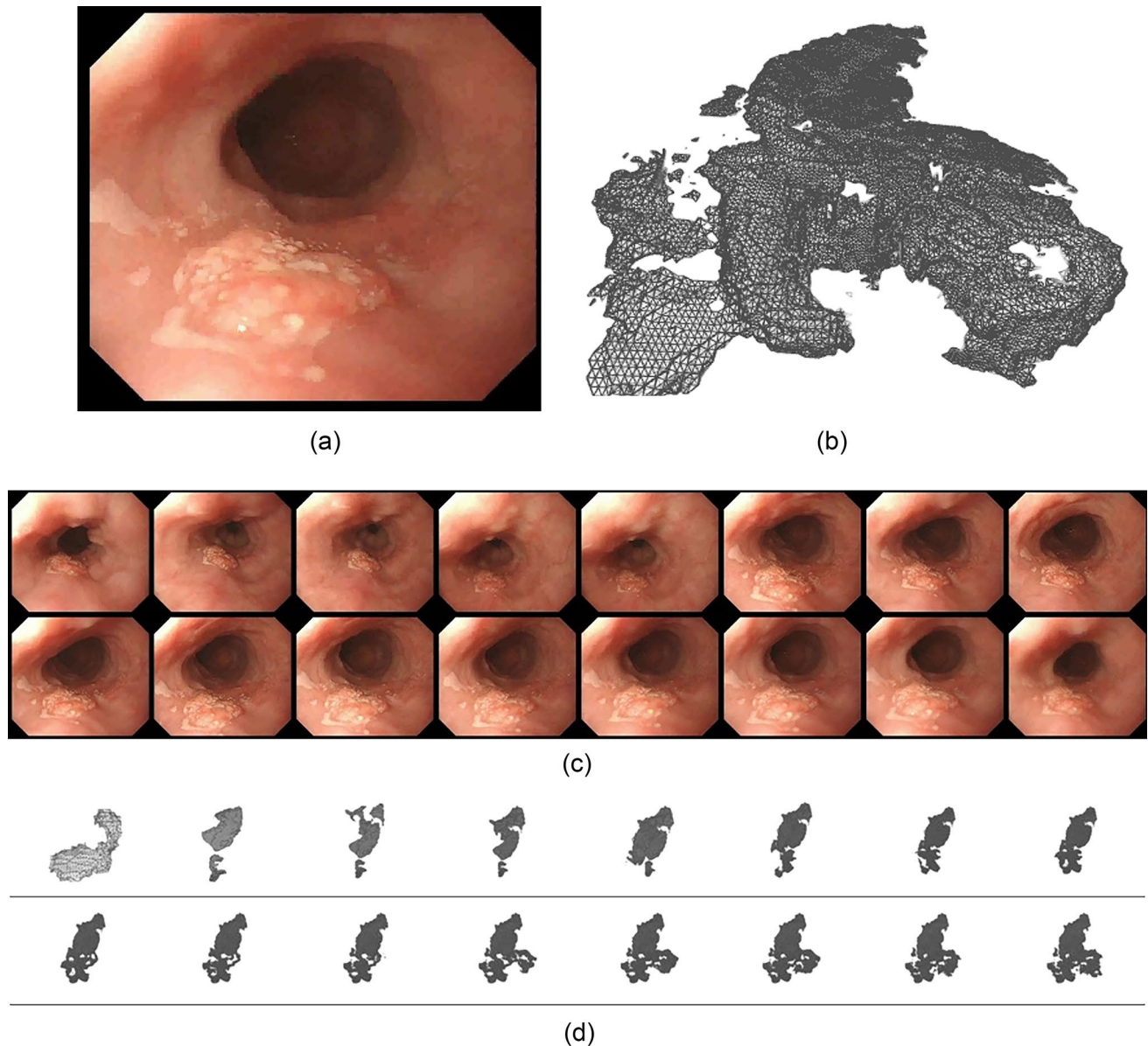


Fig. 7 The modelling processing of NeRF. **a** A frame with a lesion (arrow). **b** The modeled mesh for **a**. **c** A clip of video showing the lesion. **d** The modelled point clouds or mesh

network $\mathcal{F}$ is calculated in Eq. (15) where H , W , C refer to image patch high, width, and channel respectively.

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} w_l \odot (\hat{y}_{hw}^l - \hat{y}_{0hw}^l)^2 \quad (15)$$

The feature stacks are extracted from layer L where unit-normalization ($\hat{y}$, $\hat{y}_0$ corresponding to x , x_0 respectively) in the channel dimension takes place, and for layer l , $\hat{y}^l, \hat{y}_0^l \in \mathbb{R}^{H_l \times W_l \times C_l}$. The weight w_l is performing element-wise multiplication ($\odot$), which is equivalent to computing cosine distance [18].

Results

Figure 6 presents a process of generating 3D mesh and point clouds for the concerned lesion. At each time, the system takes consecutive 3 frames and generates the mesh points, and afterwards 3D shapes.

In Fig. 7 the formation of the lesion with regards to mesh and point clouds are illustrated. The interface can show point cloud or mesh with their visualization being interchangeable. In Fig. 7(a), a frame showing a cancer (arrow), that is modelled applying the NeRf technique as displayed in mesh at 7(b). Figure 7(c) shows a clip of the video around

Fig. 8 The illustration of training process. **a**, **b** Training takes place at varying camera rays. **c** An individual training sample selected from **b** with a specific camera ray direction [1]



the lesion with their modelling mesh or point clouds provided in 9(d).

Figure 8 demonstrates the training process implemented via Nerfstudio. The training processing is visualised live on a browser, which allows users to select any camera ray direction as exemplified at Fig. 8 (a)(b) to view detailed training (8(a)) as well as each individual training image (8(b)) with a specific camera location (Fig. 8(c)).

Figure 9 demonstrates the montage of training data set (a) and the new view angle (b) rendered using the trained model. The needed view path can be selected, defined or requested through the visualisation tool presented in Fig. 8.

While the rendered data (Fig. 9(b)) have some background information missing due to the lack of sufficient data, the concerned lesion can be rendered and viewed at any needed viewing angle, which is important in a clinical setting as an endoscopic camera can only travel in one direction within the narrow food path whereas clinicians need to know all the surrounding information for a surgical planning.

The accuracy of rendered images using the aforementioned three metric measures is provided in Table 1, which is based on two sets of data. The initial video frames with 10 min each containing over 30,000 frames per video. After pre-processing for detection of colour misalignment artefact, 6000 frames are kept. When extracting camera ray information using COLMAN approach, only 2600 images are found related due to poor image quality, e.g. blurry with floating objects or noise/artefact.

The overall mAP of 63% is considered very promising in comparison with the EAD2020 artifact detection competition [19] where an overall mAP of 28% is observed using 2%, 50% and 75% IoU thresholds.

Discussion and Conclusion

This study investigates the feasibility of reconstruction of 3D scene of GI from 2D endoscopic videos as an end-to-end process, i.e. from an input 2D video to an output of a 3D

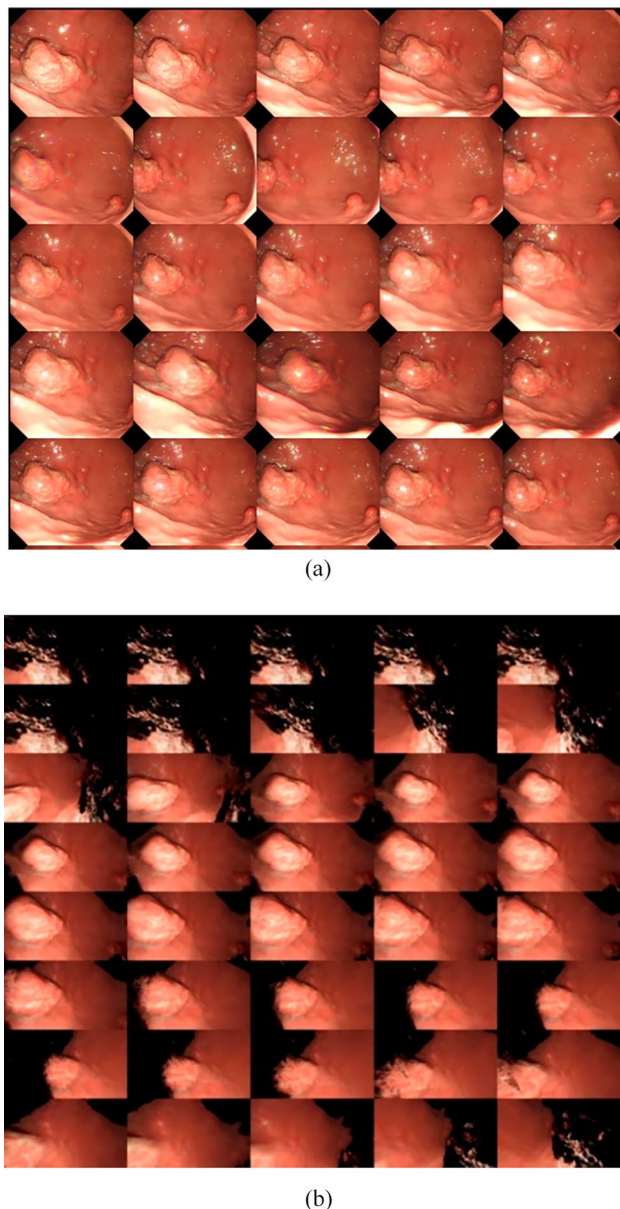


Fig. 9 Demonstration of training data (a) and rendered data (b) at a specific view direction for a polyp in the stomach [1]

Table 1 The measurement of similarity between ground truth (original) and rendered images for PSNR and SSIM, higher value implies more similar whereas lower LPIPS referring more similar between the two [1]

Data set	Image no	PSNR	SSIM	LPIPS
2	2600	19.46 ± 2.56	0.70 ± 0.054	0.49 ± 0.05

Table 2 The segmentation results for detection of artefacts based on the network shown in Fig. 4 [18]

IoU	0.50	0.70	0.95	All
mAP	99.00	87.0	1.00	63.00
box	77.85	61.64	0.99	47.42
mask	76.91	61.51	0.87	46.17

mAP Mean Averaged Precision

model, without the prior knowledge of camera information of location, intrinsic and extrinsic data. The developed system consists of two phases. The first phase accounts for data pre-processing to removal the frames that contain most of the artefact. Then SOTA NeRF approach is applied.

With regards to artefact detection, YOLACT-based real time processing (Fig. 4) network is employed, which detects, classify and segments artefacts. Similar to EAD2020 competition [20], 5 classes of artefacts are considered, including specularity, bubbles, instruments, artefacts, and bleeding. The overall mAP is 63% for segmentation, which is considered to be good (Table 2). In particular, when IoU is set to be 50% overlapping, 99% mAP can be achieved. The backbone model employs Resnet101.

With regard to modelling of 3D shapes, the challenges include the acquired endoscopic video being with less texture information on the GI surface in comparison with none-medical images, e.g. plants, and limited camera positional information that should be extracted from videos requiring images with varying view angles. Hence the ground truth images after pre-processing only contain 10% of the original input after removal of frames with 60% artefact.

However, even with only 2600 images as a training data set, the 3D model is able to render high quality images with various viewing angles. In the future, more datasets will be evaluated. Post processing includes removing noises or ghostly artefact and more training data set. In comparison with the work of NeRF for natural images [7], our results appear to be less performed where their PSNR, SSIM and LPIPS for real forward-facing images are 26.50, 0.811 and 0.250 respectively in comparison with our results of 19.46, 0.70 and 0.49 (Table 1). However, in [7], around 100 views are acquired for each filmed object with known camera information. In our study, this information needs to be extracted from endoscopic videos themselves with much less viewing angles due to the constraints of viewing space in the food-pipe passage, leading to less image frames are employed.

Due to the limitation of dataset size, this 3D modelling system proceeds with the detection of artefact first to elude misalignment. In the future when large number of annotated data is in place, it is anticipated that this step will be redundant as the modelling system will be trained to articulate frames with artefacts.

In addition, post processing will be conducted including to remove noises or ghostly artefact as recommended more recently by Warburg et al. [21]. Specifically, to make use as many video frames as possible, especially for medical applications with limited dataset, a new algorithm will be developed to establish camera information based on the existing available but less clear images through the application of human vision models. While many frames appear

burry with regard to motion tracking, human vision can still perceive these motions easily and clearly. In this way, the developed system will also become more transparent, such as applying attention models.

Dataset plays a key role in training robust models. In the future, the open public datasets C3VD [22] will be investigated and applied for this reconstruction modelling.

Significantly, to allow real-time application, 3D Gaussian Splatting-based approach [23] for real-time radiance field rendering will be evaluated in the future to increase the processing speed.

Acknowledgements This project is funded by the British Council as Women in STEM Fellowship (2023–2024). In part, this project is also funded by the Cancer Research UK under AI-enhanced Early Cancer Diagnosis Scheme (2019–2021). Their financial support is greatly acknowledged.

Author Contributions The corresponding authors contributed mainly on both writing and data processing. The last author performed data collection and data annotation. The 2nd author conducted data cleaning and analysis.

Funding The project is funded by both the British Council and the Cancer Research UK (CRUK) as detailed in the Acknowledgements Section.

Data Availability Due to concerns of patients' privacy, at the moment, the data are not available to the public. However, specific requests can be made to the authors, which will be considered based on case by case basis.

Declarations

Conflict of interest No competing interests are identified.

Research involving human and/or animals No animals were involved. All patients' personal information is removed before leaving the hospital.

Informed consent All data were collected with informed patient consent. The data collection constitutes part of the project of Endo.AI (funded by CRUK) and has reference of *REC Ref: 16/YH/0247* [12].

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Gao XW, Rahmanti A, Braden B. 3D view reconstruction from endoscopic videos for gastrointestinal tract surgery planning. In: Proceedings of the 18th international joint conference on biomedical engineering systems and technologies (BIOSTEC 2025), vol. 1. 2025. pp 221–228.
2. Lu L, Mullins CS, Schafmayer C, Zeißig S, Linnebacher M. A global assessment of recent trends in gastrointestinal cancer and lifestyle-associated risk factors. *Cancer Commun (Lond)*. 2021;41(11):1137–51.
3. Ali S, Bailey A, Ash S et al. A pilot study on automatic three-dimensional quantification of barrett's esophagus for risk stratification and therapy monitoring. *Gastroenterology*. 2021;161:865–878.
4. Prinzen M, Trost J, Bergen T, Nowack S, Wittenberg T. 3D Shape reconstruction of the esophagus from gastroscopic video. In: Handels H, et al. editors. *Image processing for medicine*. Berlin: Springer; 2015. pp. 173–8.
5. Widya A, Monno Y, Imahori K et al. 3D reconstructor of whole stomach from endoscope video using structure-from-motion, arXiv:1905.12988v1 (2019).
6. Bae G, Budvytis I, Yeung CK, Cipolla R. Deep multi-view stereo for dense 3D reconstruction from monocular endoscopic video, MICCAI 2020. *Lecture Notes in Computer Science*. Vol. 12263. Springer; 2020.
7. Mildenhall B, Srinivasan PP, Tancik M et al. NeRF: representing scenes as neural radiance fields for view synthesis. In: *ECCV 2020*. 2020.
8. Guo J, Wang J, Wei R, et al. UC-NeRF, uncertainty-aware conditional neural radiance fields from endoscopic sparse views. *IEEE Trans Med Imaging*. 2025;44(3):12841296.
9. Ali S, Dmitrieva M, Ghatwary N, et al. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. *Med Image Anal*. 2021;70:102002.
10. Ali S, Zhou F, Braden B et al. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Nat Sci Reports*. 2020;10:2748.
11. Bissoonauth-Daiboo P, Khan MMH, Auzine MM, Baichoo S, Gao XW, Heetun Z. Endoscopic image classification with vision transformers. In: *ICAAI 2023*. 2023. pp. 128–132.
12. Bolya D, Zhou C, Xiao F, Lee YJ. YOLACT: real-time instance segmentation. In: *Proceedings of the ICCV 2019*. 2019.
13. Gao XW, Taylor S, Pang W, Hui R, Lu X, Braden B. Fusion of colour contrasted images for early detection of oesophageal squamous cell dysplasia from endoscopic videos in real time. *Inform Fusion*. 2023;92:64–79.
14. Kajiya JT, Herzen BPV. Ray tracing volume densities, *Computer Graphics, SIGGRAPH* (1984).
15. Tancik M, Weber E, Ng E et al. Nerfstudio: a modular framework for neural radiance field development. arXiv:2302.04264. 2023.
16. Schonberger JL, Frahm JM. Structure-from-motion revisited. In: *CVPR'2016*. 2016.
17. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13(4):600–12.
18. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: *CVPR'2018*. 2018.
19. Ali S, Dmitrieva M, Ghatwary N et al. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, *Med Image Anal*. 2021;70(2021):102002.
20. Gao XW, Braden B. Artefact detection and segmentation based on a deep learning system. In: *Proceedings of the 2nd international workshop and challenge on computer vision in endoscopy (EndoCV2020)*, Iowa, USA, 3rd April, 2020.

21. Warburg F, Weber E, Tancik M, Hołyński A, Kanazawa A. ICCV'2023. Nerfbusters: Removing Ghostly Artifacts from Casually Captured NeRFs; 2023.
22. Bobrow TL, Golhar M, Vijayan R, et al. Colonoscopy 3D Video Dataset (C3VD). *Med Image Anal.* 2023;90:102956.
23. Kerbl B, Kopanas G, Leimkuehler T, Drettakis G. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans Graph.* 2023;42(4):1–14.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.