PHOTONICS Research

Monocular depth estimation based on deep learning for intraoperative guidance using surface-enhanced Raman scattering imaging

Aniwat Juhong,[1,2] Bo Li,[1,2] Yifan Liu,[1,2] Cheng-You Yao,[2,3] Chia-Wei Yang,[2,4] A. K. M. Atique Ullah,[2,4] Kunli Liu,[2,4] Ryan P. Lewandowski,[5] Jack R. Harkema,[5] Dalen W. Agnew,[5] Yu Leo Lei,[6] Gary D. Luker,[7] Xuefei Huang,[2,3,4] Wibool Piyawattanametha,[2,8] and Zhen Qiu[1,2,3,*]

[1]Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan 48824, USA
[2]Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, Michigan 48824, USA
[3]Department of Biomedical Engineering, Michigan State University, East Lansing, Michigan 48824, USA
[4]Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, USA
[5]Department of Pathobiology and Diagnostic Investigation, College of Veterinary Medicine, Michigan State University, East Lansing, Michigan 48824, USA
[6]Department of Periodontics and Oral Medicine, University of Michigan, Ann Arbor, Michigan 48104, USA
[7]Departments of Radiology and Biomedical Engineering, University of Michigan, Ann Arbor, Michigan 48109, USA
[8]Department of Biomedical Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok 10520, Thailand
*Corresponding author: qiuzhen@msu.edu

Received 29 July 2024; revised 16 November 2024; accepted 8 December 2024; posted 9 December 2024 (Doc. ID 536871); published 31 January 2025

Imaging of surface-enhanced Raman scattering (SERS) nanoparticles (NPs) has been intensively studied for cancer detection due to its high sensitivity, unconstrained low signal-to-noise ratios, and multiplexing detection capability. Furthermore, conjugating SERS NPs with various biomarkers is straightforward, resulting in numerous successful studies on cancer detection and diagnosis. However, Raman spectroscopy only provides spectral data from an imaging area without co-registered anatomic context. This is not practical and suitable for clinical applications. Here, we propose a custom-made Raman spectrometer with computer-vision-based positional tracking and monocular depth estimation using deep learning (DL) for the visualization of 2D and 3D SERS NPs imaging, respectively. In addition, the SERS NPs used in this study (hyaluronic acid-conjugated SERS NPs) showed clear tumor targeting capabilities (target CD44 typically overexpressed in tumors) by an *ex vivo* experiment and immunohistochemistry. The combination of Raman spectroscopy, image processing, and SERS molecular imaging, therefore, offers a robust and feasible potential for clinical applications.

https://doi.org/10.1364/PRJ.536871

1. INTRODUCTION

Surgical resection of a tumor is a standard of care therapy for most solid tumors. The ultimate goal of surgical resection is to remove the entire tumor with minimal damage to adjacent tissue, an outcome that strongly correlates with reduced tumor recurrence and improved survival [1,2]. Tumor margins in numerous aggressive cancers are typically indistinct due to the primary tumor's propensity to invade adjacent healthy tissue areas. As a result, defining appropriate margins for surgical resection remains challenging [3]. There are several modalities used in the clinic to visualize tumors and facilitate tumor removal, such as magnetic resonance imaging (MRI), positron emission tomography (PET), and computed tomography (CT) [4–7]. However, these imaging modalities lack sufficient resolution needed to identify and remove microscopic sites of cancer invasion from the main tumor mass. To achieve precise tumor delineation and complete resection, a suitable intraoperative tool should meet the following requirements: high sensitivity and specificity, short acquisition time for real-time or near-real-time intraoperative detection, and high spatial resolution. With regards to imaging modalities, optical imaging exhibits distinct advantages compared to the previously mentioned non-optical imaging modalities in several aspects, such as lack of ionizing radiation, high sensitivity, and excellent spatiotemporal resolution [8–11]. Recently, surface-enhanced Raman scattering (SERS) nanoparticles (NPs) imaging has increasingly been recognized as a promising molecular imaging technique for clear delineation of tumor margins and

tumor surgical resection due to its exceptional sensitivity, distinctive Raman signature (fingerprint), multiplexing detection capability [12–18], and lack of autofluorescence and photobleaching problems associated with fluorescence imaging. SERS NPs are composed of a gold core, Raman active dye, and silica shell, which have been developed to function as tumor-targeting beacons showing substantially strong signals due to the surface plasmon resonance (SPR) effect [19] of the metallic core (gold). In addition, they can be effortlessly conjugated with various tumor-targeting ligands and fabricated with different Raman-active dyes. Each Raman dye emits a unique Raman spectrum, called "flavor", facilitating multiplexing. Several research groups, as well as our group, have demonstrated encouraging results of SERS NPs imaging for *ex vivo*, *in vivo*, and image-guided surgery experiments [20–24]. However, Raman spectroscopy predominantly provides spectral data, lacking the capability to co-register and visually represent anatomic features, limiting applications for image-guided surgery.

To overcome this problem, we propose a custom-made Raman spectroscopy system with computer-vision-based positional tracking and DL-based techniques to visualize 2D and 3D SERS NPs imaging, respectively. Specifically, the traditional template matching algorithm [25] is employed for probe tracking, and the affine transformation [26] is then used to co-register a 2D SERS image (reconstructed by using the multiplexing algorithm [27,28]) and a sample photograph. For 3D imaging, the image is reconstructed based on a deep-learning monocular depth estimation (distance relative to the camera) of each given pixel in the input image. Multiple depth estimation accuracy with a single network (MiDaS) is a promising DL technique that estimates depth from an arbitrary input image. MiDaS utilizes a conventional encoder-decoder structure to generate the depth map images. The legacy MiDaS V2.1 model [29] uses a residual network as the backbone for feature extraction as this network structure is invulnerable to vanishing gradients and allows MiDaS to extract multi-channel feature maps from input tensors. The vision transformer (ViT) [30] is the state-of-the-art model employed in computer vision tasks. It can surpass convolutional neural networks (CNNs)-based models across various domains and settings. Therefore, the latest MiDaS versions (3.0 [31] and 3.1 [32]) replace the CNNs backbone with vision transformer networks, showing superior results. In this work, we directly utilized the pre-trained MiDaS 3.1 to reconstruct a 3D mouse image co-registering with the SERS image, as shown in Section 3.D.

2. METHODS

A. Raman Spectrometer

A schematic of the proposed Raman system is illustrated in Fig. 1. A 785 nm laser (iBeam Smart 785, Toptica Photonics, Munich, Germany) is employed for the excitation source; the custom-made fiber bundle Raman catheter (Fiber Guide Industries, Caldwell, ID, USA) is used for the laser illumination and the Raman spectra collection. A proximal end of the probe is made up of one single mode fiber (780HP, 4.4 μm core diameter) for 785 nm laser illumination and 36 multimode fibers (AFS200/220T, 200 μm core) for the Raman spectra collection as shown in Fig. 1(b). The single mode fiber for illumination is centrally positioned with the probe and encompassed by the 36 multimode fibers for Raman spectra acquisition. In addition, a fused silica plano-convex lens (L1, f = 6.83 mm, PLCS-4.0-3.1-UV, CVI Laser Optics, Albuquerque, NM, USA) is placed in front of the probe to collimate the 785 nm laser illumination with a beam diameter of 1 mm and power of 30 mW on the sample. For the distal end, it is arranged in a vertical array or linear array for effectively coupling the light to the spectrometer (Kymera 193i-A, Andor Technology, Belfast, UK) by using optical relay lenses (L2, f = 100 mm, AC254-100-B and L3, f = 80 mm, AC254-080-B, Thorlabs Inc., Newton, NJ, USA). In addition,

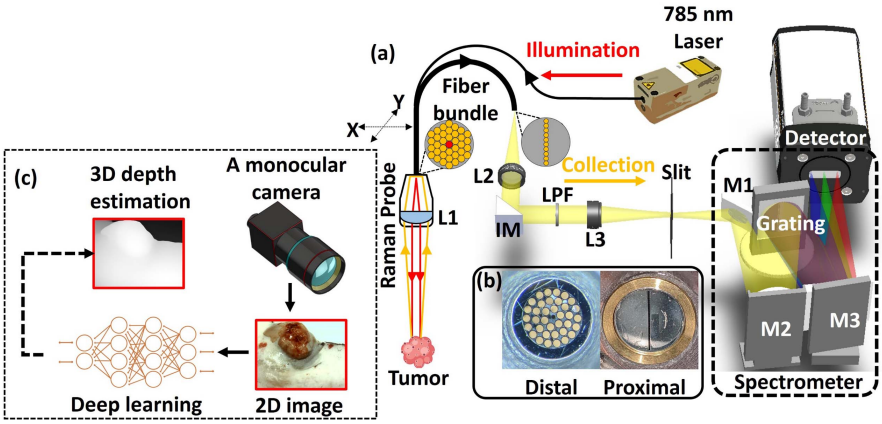

Fig. 1. Schematic of the custom-made Raman imaging system, together with the visualization system. (a) The optical diagram of the Raman spectroscopy system. A 785 nm laser is used to illuminate the sample through a single mode fiber and collimated by a plano-convex lens (L1). The scattered light is then collected by the Raman probe, coupled into the spectrometer using the relay optics (L2 and L3 lenses) with an interchangeable mirror (IM) and a long-pass filter (LPF) in between. The spectrometer consists of a rotatable grating, three mirrors (M1, reflection mirror; M2, collimating mirror; and M3, focusing mirror), and a back-illuminated deep-depletion CCD. To perform 2D Raman imaging, the Raman probe is translated by a two-axis motorized stage. (b) The photographs of the distal and proximal ends of the custom-made fiber bundle. (c) Schematic of the visualization system for generating the 2D and 3D co-registered SERS images.

the Rayleigh scattering from the collected light is filtered out by a long-pass filter (LPF, $\lambda_c = 830$ nm; BLP01-830R-25, Semrock, Rochester, NY, USA), placed between the relay lenses. As a result, the light that traverses the spectrometer is solely subjected to Stokes-Raman scattering. The Stokes-Raman scattering light from the spectrometer is then collected by a cooled deep-depletion spectroscopic charge-coupled device (CCD) array (1024×256 pixels with a pixel size of $26\ \mu m \times 26\ \mu m$; DU920P Bx-DD, Andor Technology, Belfast, UK) with a wavelength range of 835–912 nm (Raman shift of 770–1777 cm^{-1}). To achieve raster scanning, a two-axis translation stage is constructed by joining two linear stages in an orthogonal manner (DDS050, Thorlabs Inc., Newton, NJ, USA). Furthermore, a color monocular camera (ELP 5–50 mm, with Sony IMX323 chip, Shenzhen, China) is applied to track the Raman probe position and capture the sample photographs to reconstruct the 2D and 3D co-registered SERS images.

B. SERS NPs Synthesis

SERS NPs were synthesized using the tris-based assisted synthesis protocol with Au NPs formation at elevated temperature, as shown in Fig. 2(a). First, the sodium citrate reduction approach was employed to prepare 17 nm Au-NP seeds. The seeds were then mixed with tris at 98°C, followed by adding gold chloride for seed-mediated growth to obtain 50 nm Au NPs. The Raman dye was promptly added after the formation of 50 nm Au NPs, and the solution was stirred for one minute, followed by cooling in an ice bath. To functionalize SERS NPs with biomolecules, particularly hyaluronic acid (HA) and polyethylene glycol (PEG), thiol groups were employed for the attachment of these biomolecules to Au NPs via gold-thiol interaction [33–37]. S420 SERS NPs were mixed with thiolated-HA and this mixture solution was then incubated overnight at 4°C. After that, unbounded HA was removed by repeated centrifugation. Likewise, the procedure to conjugate PEG with S481 SERS NPs is the same as the HA conjugation. The size and shape of synthesized SERS NPs were characterized by a transmission electron microscope (TEM; 2200FS, JEOL Ltd., Tokyo, Japan) and a dynamic light scattering particle analyzer (DLS; Zetasizer Nano ZS, Malvern Panalytical Ltd., Malvern, UK). SERS NPs are homogenous spheres approximately 50 nm in diameter, as shown in Fig. 2(b). The DLS result was also applied to validate the distribution size with a measurement of 56 nm, as shown in Fig. 2(c). The comprehensive synthesis protocol and characterization of SERS NPs are demonstrated in our previous work [23]. The normalized Raman spectra (acquired by our custom-made Raman spectrometer) of S420 and S481 SERS NPs with a concentration of 500 pM (1 M = 1 mol/L) are demonstrated in Fig. 2(d).

C. Position Tracking and Image Co-registration Algorithms

Before processing the data acquired by a low-cost camera, a camera calibration [38,39] was applied to correct the image distortion due to the lens quality and optical alignment. The template matching algorithm [40] is then used to determine the precise position of a Raman probe image (the template image) in a large surgery area image (the input image). The concept of this algorithm is to slide the template image over the input image, akin to a 2D convolutional operation, followed by a comparison of the template and the corresponding patch of the input image, which can be done by several methods. In this work, we employed a normalized cosine coefficient (TM_CCOEF_NORMED) implemented in Python using the OpenCV Library [41] to calculate the template matching for the Raman probe detection. With the Raman probe position, the scanning position can be easily estimated during data acquisition. In addition, to accurately overlay the SERS image (X) and surgery area image (Y), an image co-registration algorithm is required by calculating the geometric transformation matrix (T), as shown in the equations below:

$$Y = T \cdot X, \tag{1}$$

$$X = \begin{bmatrix} x_1' & x_2' & \cdots & x_n' \\ y_1' & y_2' & \cdots & y_n' \\ 1 & 1 & \cdots & 1 \end{bmatrix}, \tag{2}$$

$$Y = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \\ 1 & 1 & \cdots & 1 \end{bmatrix}, \tag{3}$$

$$T = \begin{bmatrix} m_{00} & m_{01} & m_{02} \\ m_{10} & m_{11} & m_{12} \\ 0 & 0 & 1 \end{bmatrix}, \tag{4}$$

where (x_n', y_n') and (x_n, y_n) are the corresponding positions (n is the number of corresponding positions) in the input image X and the reference image (Y), respectively, and m_{ij} are the simplified transformation matrix parameters derived from the rotation, scaling, shearing, and translation matrices, as shown in the equation below:

$$T = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \mathrm{sh}_x & 0 \\ \mathrm{sh}_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \cos(\theta) & \sin(\theta) & 0 \\ -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{5}$$

In the translation matrix, t_x and t_y are the displacement along the x and y axes, respectively; in the scaling matrix, s_x and s_y are the scale factors along the x and y axes, respectively; in the shear matrix, sh_x and sh_y are the shear factors along the x and y axes, respectively; in the rotation matrix, θ is the angle of rotation. Indeed, the T matrix can be estimated by using corresponding points together with the minimized least square error (ε^2) as shown below:

$$\varepsilon^2 = \|TX - Y\|^2, \tag{6}$$

$$\frac{d\varepsilon^2}{dT} = -2X^T(Y - TX) = 0, \tag{7}$$

$$X^T Y = X^T T X, \tag{8}$$

$$T = (X^T X)^{-1}(X^T Y). \tag{9}$$

To obtain a more accurate co-registration result (2D co-registered SERS image), the estimated transformation matrix (T) is then applied to the reconstructed SERS image (X) derived

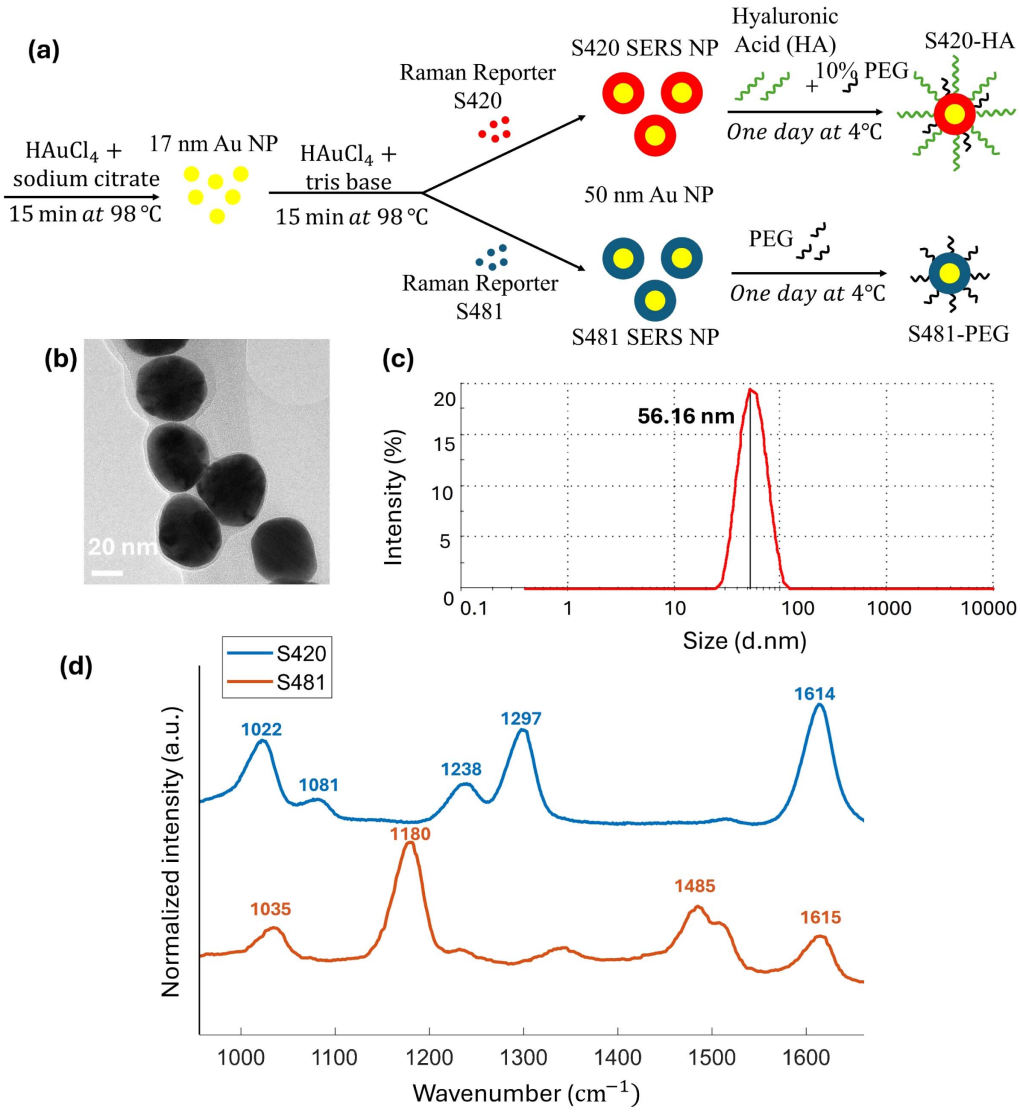

Fig. 2. Synthesis of the SERS NPs. (a) SERS NPs synthesis and HA/PEG conjugation procedure. First, 17 nm gold seeds (Au NPs) are formed. Second, the NPs further grow to 50 nm; meanwhile different Raman reporters (S420 and S481) are attached to the gold surface. Lastly, the SERS NPs are functionalized with HA or PEG. (b) TEM image of the SERS NPs with diameter of approximately 50 nm. (c) DLS result of the corresponding SERS NPs. The measured size is 56.16 nm in diameter. (d) Normalized Raman spectra of the stock SERS NPs solution of both flavors (S420 and S481).

from the demultiplexing algorithm. In our case, the raster scan was applied to reconstruct the SERS image and the fiducial landmarks (four corners of the scanning area) were marked on the sample. Thus, the four corners of the SERS image were used as the corresponding points to the four fiducial points on the samples for the image co-registration.

D. Depth Estimation Using DL

MiDaS is considered as a promising model for performing monocular depth estimation, and the original MiDaS V 2.1 [29] is based on a CNN backbone; however the newer versions (MiDaS V 3.0 [31] and V 3.1 [32]) employ transformer architectures as their backbones, which can significantly outperform the original version. The training protocols of the MiDaS V 2.1, 3.0, and 3.1 models are analogous. Briefly, the MiDaS models were trained by using 12 mixing datasets, multi-objective optimization [42] with Adam [43], and a scale-and-shift-invariant loss [44]. The encoder and decoder weights were updated by applying the learning rates of 10^{-5} and 10^{-4}, respectively. The models were initially pre-trained on a subset of the datasets for 60 epochs, followed by training for another 60 epochs on the full dataset. The complete training details are elucidated in the original MiDaS V 2.1 paper. All DL models demonstrated in this work were implemented on a personal computer equipped with an 11th Gen Intel core i7-11700k CPU, 64 GB, and an NVIDIA RTX 3090 graphic processing unit (GPU). Indeed, all MiDaS models are built using encoder and decoder structures. Each MiDaS model differs in the backbone of the encoder part (a variant of CNNs and Transformer architectures), while the rest of the model remains consistent. Since the latest MiDaS V 3.1 provides the best result compared to other versions, it is used in this study. Bidirection encoder

representation from image transformers (BEiT) [45] is used as the backbone of MiDaS V 3.1, as shown in Figs. 3(a) and 3(b). BEiT is a state-of-the-art architecture that enables self-supervised pretraining of vision transformer (ViT) to surpass supervision pretraining. The pre-training task in BEiT is the masked image modeling (MIM) head, as shown in Fig. 3(b). The concept of MIM is to recover the original visual tokens based on the corrupted image patches. In other words, MIM uses two views for each image to train the model. First, the 2D image with a size of $H \times W \times C$ is divided into a sequence of HW/P^2 patches for each channel, where (H, W) is the image size, C is the number of channels, and (P, P) is the patch size. All the patches are then flattened into vectors and linearly projected. Second, an image tokenizer converts the image into a sequence of discrete tokens rather than using raw pixels. The discrete variational autoencoder (dVAE) [46,47] is directly used to train this image tokenizer. Indeed, the image tokenizer is a readily trained token generator for the input patches.

The outputs from the tokenizer and MIM are used to determine the loss value to update the learnable parameters, allowing the network to obtain a deep understanding of underlying image patterns without the explicit labels. It is important to note that BEiT was initially designed for an image classification problem and does not provide depth estimation functionality. To assemble MiDaS V 3.1, BEiT is used as a feature extractor and must be appropriately connected to the depth decoder. Regarding the encoder-decoder in MiDaS, the input is progressively processed for each encoder stage, similar to the decoder stage. Thus, the BEiT backbone can be integrated by placing appropriate hooks, meaning a tensor computed in the encoder is taken and available as input for the decoder at one of its stages. This requires a reassembling process to reshape the tensors to fit the decoder, as shown in Figs. 3(c) and 3(d). Essentially, the input image is embedded as the tokens, which are passed through serval BEiT stages. At each stage, the tokens are reassembled into image-like representation with different resolutions. After that, the fusion module is employed to fuse and upsample these image-like representations in order to generate an exquisite prediction. The final prediction is then fed to a task-specific output head to generate the depth map image. The depth map image generated by the MiDaS model is considered as a disparity-like image (inversely

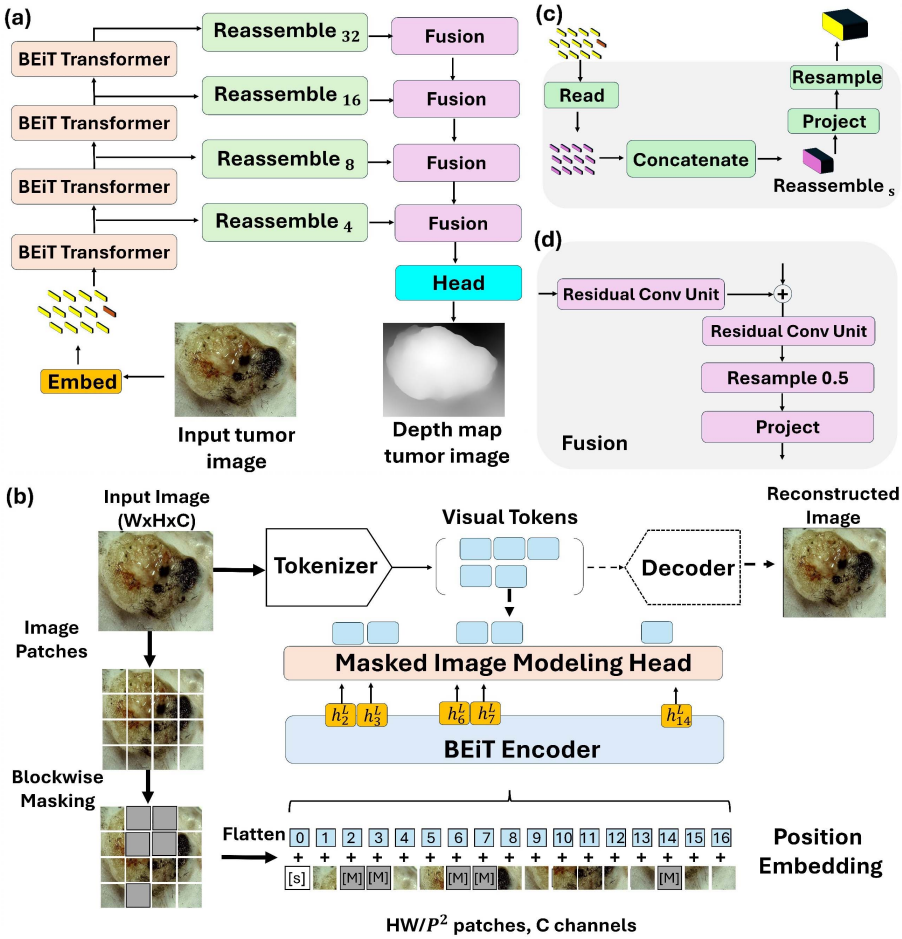

Fig. 3. (a) Overview of the MiDaS V 3.1 architecture. The input image is embedded with a positional embedding and a patch-independent readout token (orange) is included. These patches are fed to four BEiT stages. At each BEiT, the output tensor is passed through the Reassemble and Fusion blocks to predict the encoder outputs for each stage. (b) BEiT transformer architecture used in the encoder part in (a). (c) Reassemble block applied to assemble the tokens into feature maps with $1/s$ the spatial resolution of the input image. (d) Fusion block used to combine the features and upsample the feature maps by two times.

proportional to the depth map intensity), which is then projected into 3D space using the reprojectImageTo3D function in OpenCV [41]. Lastly, the color of each pixel in the 2D co-registered SERS image is mapped onto the corresponding positions (x–y plane) in the 3D space of the depth map image to obtain the final 3D SERS image.

3. RESULTS AND DISCUSSION

A. Phantom Characterizations

The step-wedge with a height of 9.5 mm of each step, which was constructed from the standard mounting bases (BA1S, Thorlabs Inc., Newton, NJ, USA), was used as a phantom to characterize the depth estimation DL models. The camera captured this phantom photograph and it was used as the input for the three different MiDaS models (CNN, ViT, and BEiT) to estimate the depth and compare the performance of each model. To quantify the performance of each model, the depth map intensities from step 4 to step 1 (along with the white-dashed line) were plotted, as illustrated in Fig. 4(a). The absolute errors were then calculated from the intensity profiles of each model and the ground truth (the black line). Table 1 shows the average absolute error ± standard deviation results of each model. It shows that the MiDaS model based on BEiT architecture can surpass other models with the lowest average absolute error of 0.0485 ± 0.1737.

Furthermore, a 3D-printed tumor phantom was utilized for thorough characterization of the MiDaS models, as depicted in Fig. 4(b). The distance between the phantom and the camera varied from 5 cm to 11 cm with an increment of 2 cm. The phantom depth map images were then generated by the MiDaS models. The quality images captured at the out-of-focus distances (5 cm and 7 cm) are unsatisfactory, leading to deterioration of depth map quality, as the models cannot correctly recognize some poor resolution areas to generate the depth map image, especially the CNN MiDaS model. Nevertheless, the BEiT model can still generate somewhat decent-quality depth map images. Table 2 shows four evaluation metrics (average value from all distances ± standard deviation): IoU, F1-score, recall, and precision, of the depth map images and their corresponding masks. This evaluation shows the overall performance of the MiDaS models for generating depth map images of the same object with different image quality (in-focus and out-of-focus images); particularly, the BEiT MiDaS model can surpass other models with the promising scores of all evaluation metrics. In addition, the complexity and average execution time for one input image were evaluated to assess the feasibility for intraoperative guidance applications. Although we implemented MiDaS on a moderate-budget GPU (an NVIDIA RTX 3090 GPU), the execution time is feasible for intraoperative guidance applications. Indeed, the execution time can be improved by using more powerful GPUs currently available on the market.

In addition to the depth map image characterization, the intensity of Raman spectra of the same sample at various distances from the Raman catheter was also characterized by using the step-wedge phantom from Fig. 4(a) and S420 SERS NPs

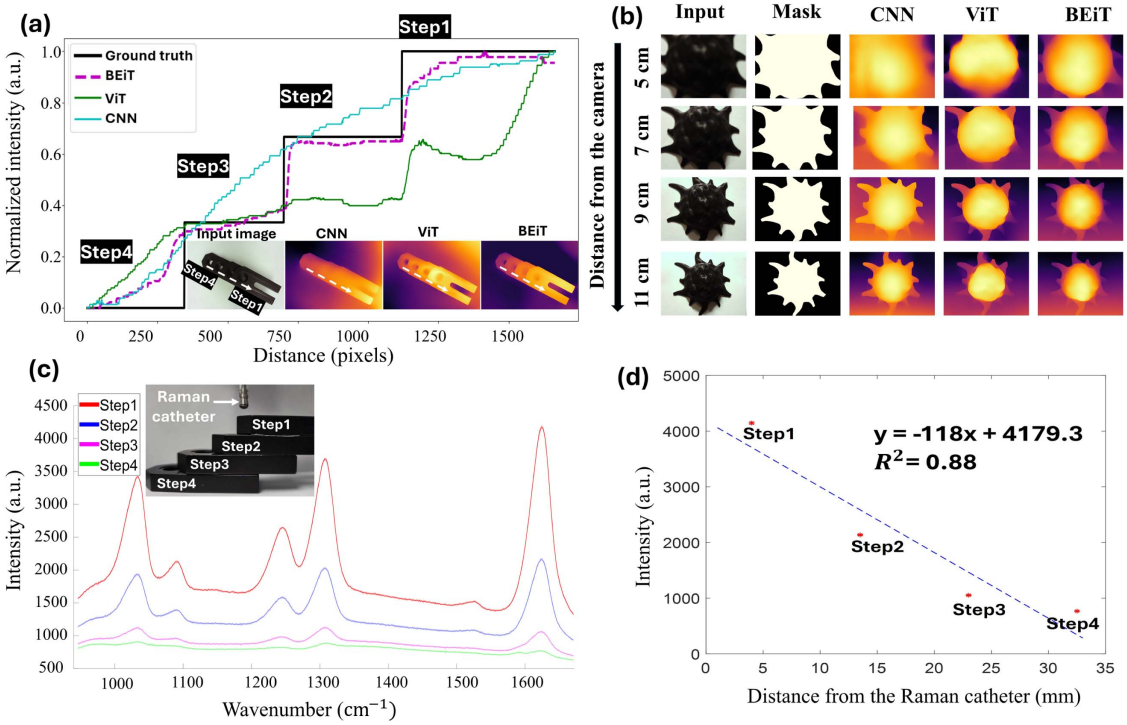

Fig. 4. Validation of depth map imaging and Raman spectra at different distances from a camera and a Raman catheter, respectively. (a) Depth map imaging of a step-wedge phantom generated by MiDaS models based on three different backbones (CNN, ViT, and BEiT) and the comparison of the depth map intensity profiles of each model. (b) Depth map imaging of a tumor phantom with different distances from the camera. (c) Raman spectra of S420 SERS NPs characterization at different distances from the Raman catheter by using the step-wedge phantom. (d) Linearity plot of the highest intensity of S420 (1614 cm^{-1}) versus the distances from the Raman catheter.

Table 1. Depth Map Intensity Characterization Result (Average Absolute Error ± Standard Deviation) of MiDaS Models with Three Different Architectures: CNN, ViT, and BEiT

Step Number	CNN	ViT	BEiT
Step 1	0.074 ± 0.560	0.318 ± 0.140	**0.051 ± 0.560**
Step 2	0.070 ± 0.046	0.252 ± 0.010	**0.032 ± 0.040**
Step 3	0.135 ± 0.088	**0.018 ± 0.016**	0.024 ± 0.012
Step 4	0.092 ± 0.077	0.161 ± 0.100	**0.087 ± 0.083**
Average	0.0927 ± 0.070	0.1872 ± 0.0665	**0.0485 ± 0.1737**

Table 2. Tumor Phantom Characterization Result of the Three Different MiDaS Models

Model	CNN	ViT	BEiT
IoU	0.139 ± 0.026	0.241 ± 0.018	**0.272 ± 0.033**
F1-score	0.244 ± 0.041	0.389 ± 0.024	**0.426 ± 0.042**
Recall	0.262 ± 0.024	0.370 ± 0.027	**0.402 ± 0.029**
Precision	0.234 ± 0.058	0.421 ± 0.074	**0.466 ± 0.088**
Execution time (s)	**0.861**	0.998	1.175
Number of parameters	**1.05 × 10^8**	3.34 × 10^8	3.45 × 10^8

solution with a concentration of 500 pM, as shown in Fig. 4(c). The SERS NPs solution was dropped on each step with a volume of 20 μL, followed by acquiring the Raman spectra using 30 mW laser power and one second exposure time. The linearity plot of the highest peak of S420 (1614 cm^{-1}) and the distance between the Raman catheter and sample is illustrated in Fig. 4(d). The distance between the catheter and the sample is inversely proportional to the intensity of the Raman spectra. Thus, this has to be addressed to enhance the accuracy of clinical applications.

B. *Ex vivo Experiment*

To validate the targeting capability of the conjugated-HA SERS NPs, we performed an *ex vivo* experiment on tumor tissue and spleen connective tissue (control) harvested from the MUC1 breast tumor mouse model [48]. All procedures used in experiments conducted on animals were approved by the Institutional Animal Care & Use Committee (IACUC) of Michigan State University. SERS NPs used in this experiment were also published in our previous work [23]. First, we scanned the background signal from all the tissues. Second, all tissues were incubated with the mixture solution of S420-HA and S481-PEG SERS NPs with a concentration of 250 pM for 15 minutes. The S481-PEG was used as a control SERS NPs solution (non-targeting). In the next step, all the tissues were rinsed with phosphate-buffered saline (PBS) four to five times, followed by acquiring the Raman spectra and reconstructing the image using the demultiplexing algorithm [27,28]. This algorithm is based on the direct classical least squares (DCLS) method, using measured Raman spectra, reference spectra of SERS NPs of each flavor (spectra of a pure SERS NPs solution at a high concentration), and background spectra as inputs to estimate the weight of a specific flavor.

Ideally, by rinsing tissues after incubation, the non-targeting NPs (S481-PEG) should be removed from the incubated tissues, and the majority of targeting NPs (S420-HA) should remain on the tumor with overexpressed CD44. However, in the practical experiment, we detected signals from both S420-HA and S481-PEG in both the tumor and normal tissues, as shown in Figs. 5(a1), 5(a2), and 5(a3), due to tissue texture and non-specific binding. Therefore, the Raman ratiometric image of S420-HA and S481-PEG was applied to evaluate the targeting of the NPs, as shown in Fig. 5(a4). According to the ratiometric result, the ratio of targeting NPs (S420-HA) on the tumor tissue is significantly stronger than the ratio on the control tissue, which is encouraging and promising. Furthermore, the H&E and IHC of CD44 of the corresponding tissues were prepared, and the results are shown in Figs. 5(b1) and 5(b2), respectively. CD44 is labeled as brown areas, and they are intense (overexpressed) in the tumor tissue, as shown in Fig. 5(c). This is also consistent with the ratiometric result.

C. Image-Guided Surgery Experiment

In this experiment, we would like to validate the capability of the proposed Raman system and SERS NPs and closely replicate the clinical conditions of human surgery. A 5-month-old female C57BL6 double transgenic mouse with breast cancer was used for this experiment. First, the operative surgery area (tumor area) was defined, followed by acquiring the Raman signal as the background signal. The mouse was then intratumorally injected with the S420-HA solution with a concentration of 500 pM, a volume of 100 μL, and a depth of injection of approximately 2–3 mm. 42 hours after the injection, the mouse was euthanized by using a table-top research anesthesia machine (V300PS, Parkland Scientific, USA) with 10 L/min of oxygen flow and 1.5% of anesthetic agent vapor in oxygen during the image-guided surgery imaging. The tumor skin was then cut open, followed by rinsing the tumor area with PBS four to five times and acquiring Raman spectra. After that, the Raman image (weight of S420-HA) of the scanned area was reconstructed and the tumor was also gradually resected following the white boundaries, as shown in Fig. 6. It is important to note that the deeper the resection is performed, the weaker the signal of SERS NPs is. This is due to the effective working distance of the Raman probe. Therefore, the depth of information on the operative area is essential for providing additional insights and guidance for more effective surgery, and we also demonstrate the concept of the 3D SERS NPs imaging in the next section.

D. 2D Tracking and 3D SERS Imaging

In addition to the image-guided surgery and *ex vivo* experiments, we demonstrate our custom-made Raman system and monocular depth estimation based on DL to visualize the SERS NPs signal on the sample in 2D and 3D surfaces in the physical world. To simplify the experiment, the S420-HA solution with a concentration of 500 pM was directly dropped on the cut-open tumor of another breast tumor mouse with an incubation time of 15 minutes, followed by rinsing with PBS four to five times and acquiring Raman spectra. Before applying this S420-HA solution, the background Raman signal was also acquired as it is one of the input variables for the SERS

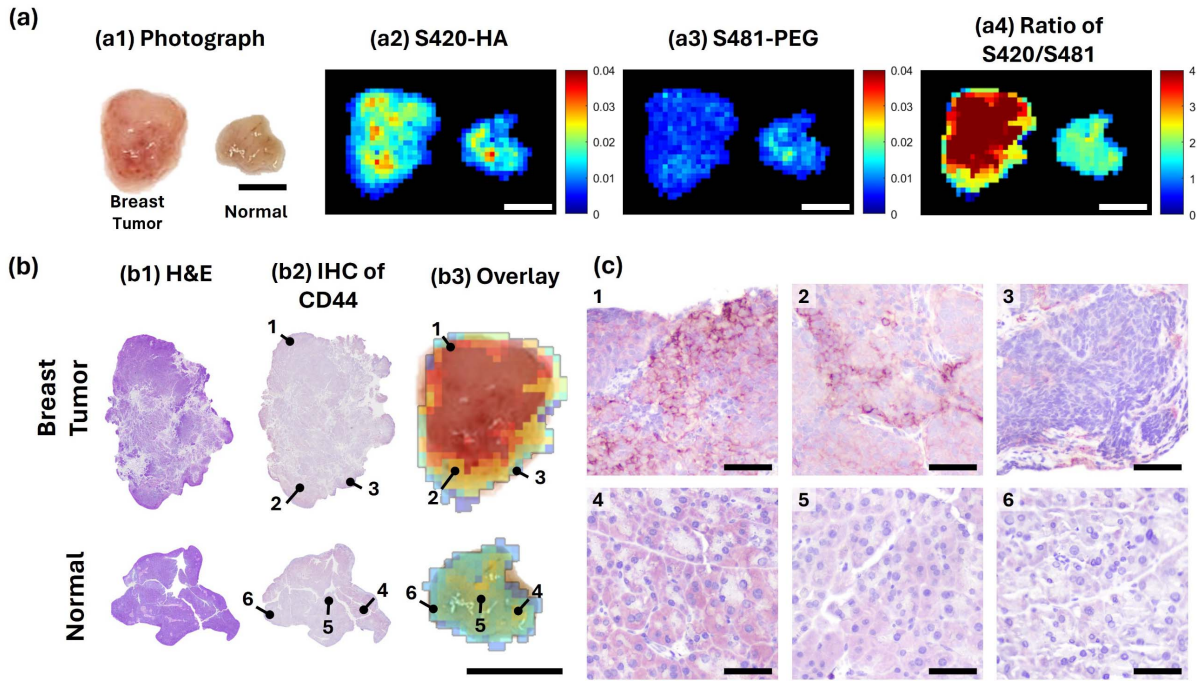

Fig. 5. (a) Multiplexed Raman images of tissues topically stained with the mixture of SERS-HA (CD44 targeting) and SERS-PEG (control) solutions. (a1) Photographs of the mouse tumor tissue and spleen connective tissue (control), and (a2)–(a4) Raman images of individual channels and ratiometric result. (b) H&E and IHC-CD44 images of the corresponding tissues. (c) Representative enlarged IHC images in (b) of the breast tumor and normal tissues. Scale bars in (a), (b) and (c) are 5 mm and 50 μm, respectively.

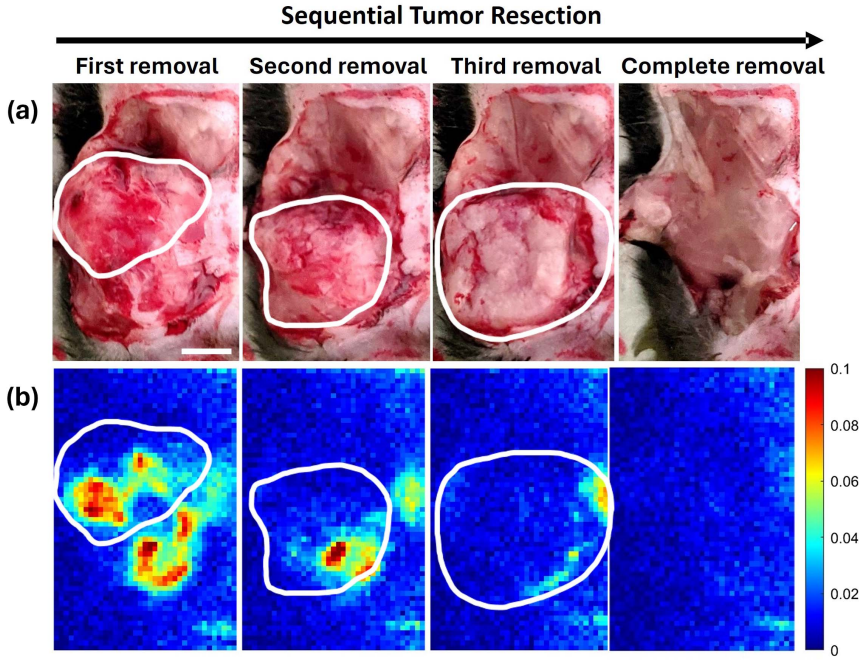

Fig. 6. SERS-image-guided surgery for resection of a mouse with a breast tumor. (a) Photographs of the tumor during the intraoperative SERS-image-guided surgery from the first removal to the complete removal. (b) Corresponding SERS images (weight of S420-HA) reconstructed by the demultiplexing algorithm. The scale bar is 5 mm, and the white boundaries depict the resection regions.

image reconstruction. A color camera was used to record the video of the scanning area and capture the photograph of the sample to generate the 2D SERS mapping video and the 3D SERS image. To generate the 2D SERS mapping video, the template matching algorithm was applied to track the Raman catheter position to estimate the scanning positions. After that, the SERS signals (the weights of S420-HA) were then generated on these estimated scanning positions, as shown in

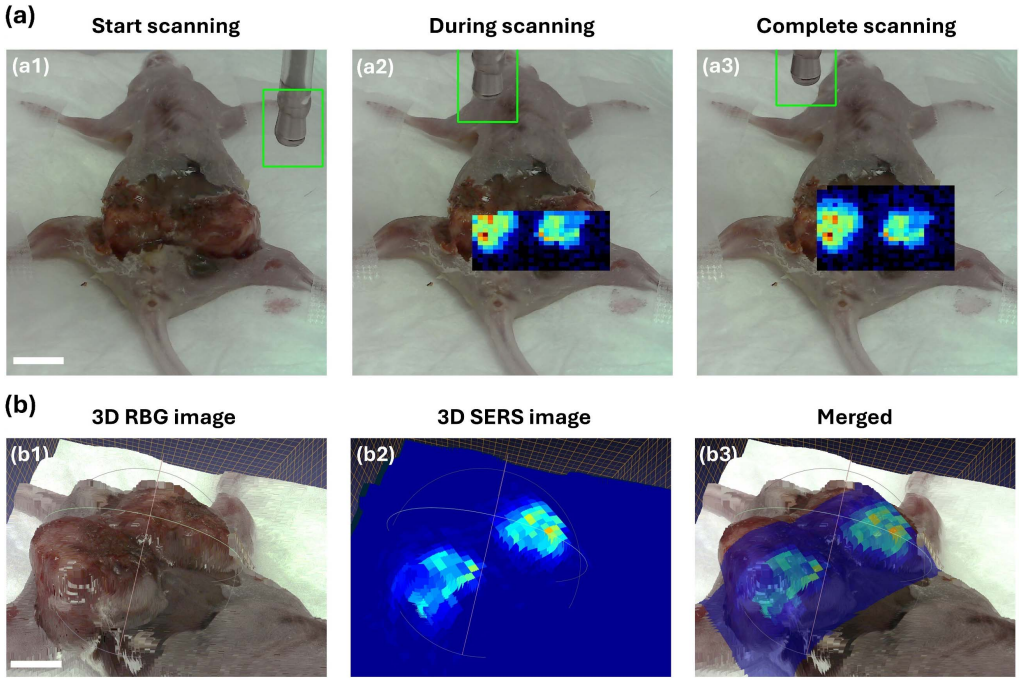

Fig. 7. (a) 2D SERS image during Raman spectra acquisition (see Visualization 1): (a1) before scanning, (a2) during scanning, and (a3) complete scanning. (b) 3D image of the sample, SERS, and co-registered SERS reconstructed by using affine transformation and MiDaS V 3.1 DL model with the BEiT backbone architecture (see Visualization 2). The scale bars of (a1) and (b1) are 10 mm and 8 mm, respectively.

Fig. 7(a) and Visualization 1. After completing the scanning, the image co-registration algorithm was applied to co-register the 2D SERS image with the sample photograph and the MiDaS DL based on BEiT was utilized to generate the depth map image. With these 2D co-registered SERS and depth map images, the 3D co-registered SERS image was reconstructed and projected as point clouds in the 3D space, as shown in Fig. 7(b) and Visualization 2. Since Fig. 7(a) shows the Raman catheter tracking with real-time 2D SERS image reconstruction, a large field of view (FOV) was needed to acquire the image for covering the catheter and scanning area images. Nevertheless, the smaller FOV was employed to illustrate greater detail in the 3D SERS image shown in Fig. 7(b). According to these promising results, the proposed method can facilitate 2D and 3D SERS imaging through the utilization of a Raman catheter system and a simple camera, which can immeasurably improve the visualization and precision of SERS NPs distribution, leading to more efficient clinical applications. Specifically, it is beneficial for image-guided surgery by assisting surgeons to locate solid tumors and achieve more precise resections. However, there is an obvious artifact pattern in 3D SERS imaging. It is caused by the large excitation laser (approximately 1 mm). This could be resolved by improving the optic design of the Raman system to reduce the beam size and adding a scanner to maintain the acquisition speed, which could be our future work.

4. DISCUSSION AND CONCLUSIONS

Intraoperative imaging systems, in tandem with exogenous contrast agents, play a crucial role in tumor resection by assisting a surgeon to identify tumor areas with a high degree of sensitivity and specificity. However, traditional imaging systems commonly encounter poor tumor margin visualization, particularly the weak signal of a tumor at deeper layers. Without depth information, these weak signals might be neglected, leading to ineffective tumor resection. Therefore, the whole tumor might not be completely removed, causing tumor recurrence. In recent years, SERS NPs imaging has been increasingly recognized as an encouraging molecular imaging technique due to its remarkable sensitivity, multiplexing detection capability, and photostability. In addition, it has demonstrated significant potential in cancer detection and enhancing delineation of tumor margin, as SERS NPs can be easily conjugated with various biomarkers.

In this work, we propose an approach to visualize 2D and 3D SERS imaging. A step-wedge phantom and a tumor phantom were used to evaluate the depth map estimation performance of MiDaS models with three different backbone architectures: CNN, ViT, and BEiT. MiDaS based on BEiT can outperform other models; thus, it was employed for the 3D visualization of SERS NPs. HA-conjugated SERS NPs were evaluated by *ex vivo* and image-guided surgery experiments using the traditional 2D SERS image reconstruction showing promising results. Nevertheless, it lacks the depth information for practical clinical applications, affecting surgery outcomes. Therefore, the proposed approach combines the use of a custom-made Raman spectrometer with computer-vision-based positional tracking for 2D SERS imaging and monocular depth estimation based on the MiDaS model for 3D SERS imaging. This combination can overcome the disadvantage of the conventional Raman system, which only provides spectral information and is unsuitable for clinical applications. The 2D

and 3D image co-registration between the Raman images and the sample photographs in the physical world enables better performance and efficiency of tumor resection, potentially leading to its implementation in human clinical trials in the near future. Essentially, the proposed method shows a proof-of-concept study of image-guided surgery using 3D and 2D SERS imaging. However, there are some limitations that need to be improved in the future, particularly the resolution of SERS imaging. The excitation laser beam diameter in the proposed system is somewhat large (roughly 1 mm), causing the artifact in 3D and 2D image reconstruction, which is unsuitable for small tumor resection. Therefore, the optics part should be re-designed to obtain a smaller beam size for enhanced resolution. In addition, the depth map estimation using MiDaS can be influenced by the resolution of an input image acquired at an out-of-focus distance. Thus, auto-focus approaches, such as resolution enhancement deep learning or a hardware-based approach, should be considered to avoid this problem. The proposed method may be more feasible for future clinical applications as a result of these improvements.

Funding. National Science Foundation (1808436, 1918074, 2306708, 2237142-CAREER); U.S. Department of Energy (234402).

Acknowledgment. We would like to thank Amy Porter, Investigative Histopathology Laboratory, Michigan State University, for preparing the H&E and IHC slides.

Disclosures. The authors declare no conflicts of interest related to this article.

Data Availability. The data are available from the corresponding author on reasonable request.

REFERENCES

1. E. Y. Lukianova-Hleb, Y.-S. Kim, I. Belatsarkouski, *et al.*, "Intraoperative diagnostics and elimination of residual microtumours with plasmonic nanobubbles," Nat. Nanotechnol. **11**, 525–532 (2016).
2. T. Wang, D. Wang, H. Yu, *et al.*, "A cancer vaccine-mediated postoperative immunotherapy for recurrent and metastatic tumors," Nat. Commun. **9**, 1532 (2018).
3. N. Anup, A. Gadeval, and R. K. Tekade, "A 3D-printed graphene BioFuse implant for postsurgical adjuvant therapy of cancer: proof of concept in 2D-and 3D-spheroid tumor models," ACS Appl. Bio Mater. **6**, 1195–1212 (2023).
4. H. Aydın, I. Sillenberg, and H. von Lieven, "Patterns of failure following CT-based 3-D irradiation for malignant glioma," Strahlentherapie und Onkologie **177**, 424–431 (2001).
5. R. W. Gao, N. T. Teraphongphom, N. S. van den Berg, *et al.*, "Determination of tumor margins with surgical specimen mapping using near-infrared fluorescence," Cancer Res. **78**, 5144–5154 (2018).
6. X. Gao, Q. Yue, Z. Liu, *et al.*, "Guiding brain-tumor surgery via blood–brain-barrier-permeable gold nanoprobes with acid-triggered MRI/SERRS signals," Adv. Mater. **29**, 1603917 (2017).
7. S. Kunjachan, J. Ehling, G. Storm, *et al.*, "Noninvasive imaging of nanomedicines and nanotheranostics: principles, progress, and prospects," Chem. Rev. **115**, 10907–10937 (2015).
8. M. F. Kircher, U. Mahmood, R. S. King, *et al.*, "A multimodal nanoparticle for preoperative magnetic resonance imaging and intraoperative optical brain tumor delineation," Cancer Res. **63**, 8122–8125 (2003).
9. S. Pal, A. Ray, C. Andreou, *et al.*, "DNA-enabled rational design of fluorescence-Raman bimodal nanoprobes for cancer imaging and therapy," Nat. Commun. **10**, 1926 (2019).
10. J. Qi, J. Li, R. Liu, *et al.*, "Boosting fluorescence-photoacoustic-Raman properties in one fluorophore for precise cancer surgery," Chem **5**, 2657–2677 (2019).
11. A. M. Zysk, K. Chen, E. Gabrielson, *et al.*, "Intraoperative assessment of final margins with a handheld optical imaging probe during breast-conserving surgery may reduce the reoperation rate: results of a multicenter study," Ann. Surg. Oncol. **22**, 3356–3362 (2015).
12. S. Laing, L. E. Jamieson, K. Faulds, *et al.*, "Surface-enhanced Raman spectroscopy for *in vivo* biosensing," Nat. Rev. Chem. **1**, 0060 (2017).
13. J. Langer, D. J. de Aberasturi, J. Aizpurua, *et al.*, "Present and future of surface-enhanced Raman scattering," ACS Nano **14**, 28–117 (2019).
14. M. Li, S. K. Cushing, and N. Wu, "Plasmon-enhanced optical sensors: a review," Analyst **140**, 386–406 (2015).
15. M. Li, H. Lin, S. K. Paidi, *et al.*, "A fluorescence and surface-enhanced Raman spectroscopic dual-modal aptasensor for sensitive detection of cyanotoxins," ACS Sens. **5**, 1419–1426 (2020).
16. X. Pan, L. Li, H. Lin, *et al.*, "A graphene oxide-gold nanostar hybrid based-paper biosensor for label-free SERS detection of serum bilirubin for diagnosis of jaundice," Biosens. Bioelectron. **145**, 111713 (2019).
17. B. Shan, Y. Pu, Y. Chen, *et al.*, "Novel SERS labels: rational design, functional integration and biomedical applications," Coord. Chem. Rev. **371**, 11–37 (2018).
18. Y. Wang, S. Kang, A. Khan, *et al.*, "Quantitative molecular phenotyping with topically applied SERS nanoparticles for intraoperative guidance of breast cancer lumpectomy," Sci. Rep. **6**, 21242 (2016).
19. A. Liang, Q. Liu, G. Wen, *et al.*, "The surface-plasmon-resonance effect of nanogold/silver and its analytical applications," TrAC Trends Anal. Chem. **37**, 32–47 (2012).
20. R. M. Davis, J. L. Campbell, S. Burkitt, *et al.*, "A Raman imaging approach using CD47 antibody-labeled SERS nanoparticles for identifying breast cancer and its potential to guide surgical resection," Nanomaterials **8**, 953 (2018).
21. H. Gao, "Progress and perspectives on targeting nanoparticles for brain drug delivery," Acta Pharmaceutica Sin. B **6**, 268–286 (2016).
22. R. Huang, S. Harmsen, J. M. Samii, *et al.*, "High precision imaging of microscopic spread of glioblastoma with a targeted ultrasensitive SERRS molecular imaging probe," Theranostics **6**, 1075 (2016).
23. K. Liu, A. A. Ullah, A. Juhong, *et al.*, "Robust synthesis of targeting glyco-nanoparticles for surface enhanced resonance Raman based image-guided tumor surgery," Small Sci. **4**, 2300154 (2024).
24. C. L. Zavaleta, B. R. Smith, I. Walton, *et al.*, "Multiplexed imaging of surface enhanced Raman scattering nanotags in living mice using noninvasive Raman spectroscopy," Proc. Natl. Acad. Sci. USA **106**, 13511–13516 (2009).
25. R. Brunelli, *Template Matching Techniques in Computer Vision: Theory and Practice* (Wiley, 2009).
26. K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," Int. J. Comput. Vis. **60**, 63–86 (2004).
27. E. Garai, S. Sensarn, C. L. Zavaleta, *et al.*, "High-sensitivity, real-time, ratiometric imaging of surface-enhanced Raman scattering nanoparticles with a clinically translatable Raman endoscope device," J. Biomed. Opt. **18**, 096008 (2013).
28. C. L. Zavaleta, E. Garai, J. T. Liu, *et al.*, "A Raman-based endoscopic strategy for multiplexed molecular imaging," Proc. Natl. Acad. Sci. USA **110**, E2288–E2297 (2013).
29. R. Ranftl, K. Lasinger, D. Hafner, *et al.*, "Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer," IEEE Trans. Pattern Anal. Mach. Intell. **44**, 1623–1637 (2020).
30. A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: transformers for image recognition at scale," arXiv, arXiv:2010.11929 (2020).
31. R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 12179–12188.

32. R. Birkl, D. Wofkm, and M. Müller, "Midas v3.1: a model zoo for robust monocular relative depth estimation," arXiv (2023).
33. O. Gotov, G. Battogtokh, D. Shin, *et al.*, "Hyaluronic acid-coated cisplatin conjugated gold nanoparticles for combined cancer treatment," J. Ind. Eng. Chem. **65**, 236–243 (2018).
34. H. Lee, K. Lee, I. K. Kim, *et al.*, "Synthesis, characterization, and *in vivo* diagnostic applications of hyaluronic acid immobilized gold nanoprobes," Biomaterials **29**, 4709–4718 (2008).
35. M.-Y. Lee, J.-A. Yang, H. S. Jung, *et al.*, "Hyaluronic acid–gold nanoparticle/interferon α complex for targeted treatment of hepatitis C virus infection," ACS Nano **6**, 9522–9531 (2012).
36. X. Li, H. Zhou, L. Yang, *et al.*, "Enhancement of cell recognition *in vitro* by dual-ligand cancer targeting gold nanoparticles," Biomaterials **32**, 2540–2545 (2011).
37. Y. Xue, X. Li, H. Li, *et al.*, "Quantifying thiol–gold interactions towards the efficient strength control," Nat. Commun. **5**, 4348 (2014).
38. A. Juhong, B. Li, C.-Y. Yao, *et al.*, "Cost-effective near infrared fluorescence wide-field camera for breast tumor imaging," IEEE Photon. Technol. Lett. **35**, 813–816 (2023).
39. Z. Zhang, "A flexible new technique for camera calibration," IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1330–1334 (2000).
40. F. Jurie and M. Dhome, "A simple and efficient template matching algorithm," in *Proceedings Eighth IEEE International Conference on Computer Vision (ICCV)* (2001), pp. 544–549.
41. G. Bradski, "The OpenCV Library," Dr. Dobb's J. of Softw. Tools **120**, 122–125 (2000).
42. O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," Adv. Neural Inf. Process. Syst. **31**, 525–536 (2018).
43. D. P. Kingma and J. Ma, "Adam: a method for stochastic optimization," arXiv, arXiv:1412.6980 (2014).
44. Z. Li and N. Snavely, "Megadepth: learning single-view depth prediction from internet photos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 2041–2050.
45. H. Bao, L. Dong, S. Piao, *et al.*, "BEiT: BERT pre-training of image transformers," arXiv, arXiv:2106.08254 (2021).
46. A. Ramesh, M. Pavlov, G. Goh, *et al.*, "Zero-shot text-to-image generation," in *International Conference on Machine Learning (PMLR)* (2021), pp. 8821–8831.
47. J. T. Rolfe, "Discrete variational autoencoders," arXiv, arXiv:1609.02200 (2016).
48. N. Stergiou, N. Gaidzik, A.-S. Heimes, *et al.*, "Reduced breast tumor growth after immunization with a tumor-restricted MUC1 glycopeptide conjugated to tetanus toxoid," Cancer Immunology Res. **7**, 113–122 (2019).