Vision-Based Neurosurgical Guidance: Unsupervised Localization and Camera-Pose Prediction ⋆

Gary Sarwin[1][0000−0002−9011−4882], Alessandro Carretta[2,3], Victor Staartjes[2], Matteo Zoli[3], Diego Mazzatenta[3], Luca Regli[2], Carlo Serra[2], and Ender Konukoglu[1]

[1] Computer Vision Lab, ETH Zurich, Switzerland
[2] Department of Neurosurgery, University Hospital of Zurich, Zurich, Switzerland
[3] Department of Biomedical and Neuromotor Sciences (DIBINEM), University of Bologna, Bologna, Italy

Abstract. Localizing oneself during endoscopic procedures can be problematic due to the lack of distinguishable textures and landmarks, as well as difficulties due to the endoscopic device such as a limited field of view and challenging lighting conditions. Expert knowledge shaped by years of experience is required for localization within the human body during endoscopic procedures. In this work, we present a deep learning method based on anatomy recognition, that constructs a surgical path in an unsupervised manner from surgical videos, modelling relative location and variations due to different viewing angles. At inference time, the model can map unseen video frames on the path and estimate the viewing angle, aiming to provide guidance, for instance, to reach a particular destination. We test the method on a dataset consisting of surgical videos of pituitary surgery, i.e. transsphenoidal adenomectomy, as well as on a synthetic dataset. An online tool that lets researchers upload their surgical videos to obtain anatomy detections and the weights of the trained YOLOv7 model are available at: https://surgicalvision.bmic.ethz.ch.

Keywords: Neuronavigation · Unsupervised Embedding · Anatomical Recognition · Endoscopic Surgeries · Surgical AI · Surgical Vision

1 Introduction

The environment during an endoscopic procedure poses numerous challenges to the surgeons. Successfully navigating this environment requires extensive experience combined with an extremely high level of anatomical understanding from the video feed. Challenges stem from the inherent nature of the human anatomy, such as non-rigid deformations, the absence of obvious boundaries between anatomical structures, and adverse events like bleeding that can happen during surgery, as well as the imaging device, such as limited field of view and light reflection.

A variety of methods have been developed to assist neurosurgeons orient themselves during neurosurgeries. While computer-assisted neuronavigation has been a crucial tool and long-term research focus in the community [1,2], it relies on preoperative imaging and brain shift hampers this reliability [3].

Additional real-time anatomical guidance can be achieved through interoperative MRI [4,5], ultrasound [6,7], the use of fluorescent substances [8,9], awake surgery [10], and electrophysiological neuromonitoring [11,12]. These techniques are efficient, relying on physical traits, however, also costly as they demand proficiency in a new imaging modality and, more importantly, require temporary surgery halts or instrument retractions for intraoperative information [13].

The pursuit of a more cost-effective real-time solution, independent of additional machinery, coupled with advancements in deep learning techniques, has propelled the development of vision-based localization methods. Various approaches, including structure from motion and simultaneous localization and mapping (SLAM) [14–17], aim for 3D map reconstruction based on feature correspondence. [15]. Many vision-based localization methods rely on distinctive landmark positions and tracking them across frames for localization. Factors inherent to endoscopic neurosurgical videos, such as low texture, a lack of distinguishable features, non-rigid deformations, and disruptions [15], degrade their performance. Consequently, alternative solutions are imperative to tackle these challenges.

Despite recent progress, the task of detecting or segmenting anatomical structures, which could serve as a foundation for an alternative approach to neuronavigation, remains under-explored and poses an ongoing challenge. It is important to note that recognizing anatomy in surgical videos is more challenging than detecting surgical tools, given the absence of clear boundaries and variations in color or texture between anatomical structures.

Interest in applying machine learning to neurosurgery has increased, especially in pituitary surgery, and first works have explored the possibility of detection and segmentation of anatomical structures [18,19].

[18] reported promising results of anatomical structure detection, and additionally, demonstrated a way to use the detections and their constellation to construct a common surgical path in an unsupervised manner, allowing relative localization during surgery. Furthermore, in [19] the authors proposed a multi-task network to identify critical structures during the sellar phase of pituitary surgery. Their model, PAINet, jointly predicts segmentation of the two largest structures and centroids of the smaller and less frequently occurring structures.

In this work, we extend the model proposed in [18] to include the *viewing direction of the endoscope* in the construction of the surgical path and mapping on the path. The viewing direction plays an important role in navigation during surgery since it heavily influences the structures and their constellations viewed by the endoscope. Thus, including it in the model has the potential to yield better surgical paths and more accurate guidance. Importantly, we include the viewing direction, represented as rotational angles around the x and y axis (i.e.,

pitch and yaw), in the model also in an unsupervised way without assuming the presence of any camera parameters for training.

The unsupervised learning of the surgical path and viewing direction is facilitated by an Autoencoder (AE) architecture using a training set of videos. The AE is trained to reconstruct bounding boxes of a given frame based on a constrained latent representation. At inference time, the latent representation of a new frame provides relative positioning and viewing angles. Unlike approaches reconstructing a 3D environment and relying on landmarks for localization, our method aims to construct a common surgical roadmap and localizing within that map relying on bounding box detections. Relying on semantic bounding box detections eliminates the need for tracking arbitrary landmarks, facilitating handling disruptions during surgery, such as bleeding, flushing and retractions. The learned mapping relies on the principle that the visibility, relative sizes and constellations of anatomical structures, which can be inferred from bounding box detections, strongly correlate with the position along the surgical trajectory and the viewing angle of the endoscope.

The proposed approach is demonstrated on the transsphenoidal adenomectomy procedure, chosen for its relatively one-dimensional surgical path. This choice makes it well-suited for proving the concept of our suggested method.

2   Methods

2.1   Problem Formulation and Approach

Let $\mathbf{S}_t$ denote a sequence of endoscopic video frames $\mathbf{x}_{t-s:t}$. Here, s denotes the sequence length, and $\mathbf{x}_t \in \mathbb{R}^{w \times h \times c}$ represents the t-th frame with w, h, and c indicating the width, height, and number of channels, respectively. Our primary goal is to embed the sequence $\mathbf{S}_t$ into a 3D latent dimension represented by the variable $\mathbf{z} = \left[z^1, z^2, z^3\right]$.

Our approach involves identifying anatomical structures in the sequence $\mathbf{S}_t$ and mapping the frame $\mathbf{x}_t$ to the latent space using the identified structures. Notably, z^1 serves as an *implicit* surgical atlas, signifying a path from the beginning of the procedure until the final desired anatomy. It is implicit because position information along the surgical path is unavailable for constructing the latent space. z^2 and z^3 represent pitch and yaw angles, respectively, forming a rotation matrix to predict the endoscope's viewing direction. Note that depth and camera pose information is rarely available in standard configurations, either because it is inaccessible or the functionality is missing. Extracting and collecting this data from medical devices can be cumbersome, or not possible due to the highly regulated environment and restrictions concerning modifications to these devices. Therefore, these are modelled and learned in an unsupervised way.

To achieve this, object detection is performed on all frames $\mathbf{x}_{t-s:t}$ in $\mathbf{S}_t$, resulting in a sequence of detections $\mathbf{c}_{t-s:t}$ denoted as $\mathbf{C}_t$. A detection $\mathbf{c}_t \in \mathbb{R}^{n \times 5}$ includes binary variables $\mathbf{y}_t = [y_t^0, \dots, y_t^n] \in {0,1}^n$ indicating presence of structures in the t-th frame and bounding box coordinates $\mathbf{b}_t = [\mathbf{b}_t^0, \dots, \mathbf{b}_t^n]^T \in \mathbb{R}^{n \times 4}$.

An autoencoder architecture is employed to embed $\mathbf{C}_t$ into $\mathbf{z}_t = \left[z_t^1, z_t^2, z_t^3\right]$. The encoder maps $\mathbf{C}_t$ to $\mathbf{z}_t$, and the decoder reconstructs $\hat{\mathbf{c}}_t = \left(\hat{y}_t, \hat{b}_t\right)$, representing detections for the last frame in a given sequence. z_t^1 is used to generate $\hat{y}_t$ and $\hat{b}_t^I = \left[\hat{b}_t^{I,0}, \ldots, \hat{b}_t^{I,n}\right]^T$, the bounding box location reconstruction assuming pitch and yaw angles are zero, i.e., in a *centred view*. z_t^2 and z_t^3 are used to build a rotation matrix $\mathbf{R}_t$, as shown in Figure 1, to model variability due to differences in viewing angles. The final bounding box reconstructions, $\hat{b}_t$, are obtained by rotating the bounding-box centre coordinates of $\hat{b}_t^I$ with $\mathbf{R}_t$ and keeping the same bounding-box size in the rotated position.

The model parameters are updated to ensure that the prediction with the rotated centre coordinates, $\hat{\mathbf{c}}_t$, fits $\mathbf{c}_t$ in a training set, as elaborated in the following.

2.2   Object Detection

Our method involves the identification of anatomical structures by detecting them as bounding boxes in video frames. For this purpose, we utilize the YOLOv7 network in the object detection phase of our pipeline [20]. The network is trained on endoscopic videos from a training set, where frames are sparsely labelled with bounding boxes. Subsequently, the trained network is applied to all frames of the training videos to generate detections for these classes on every frame. These detections serve as the input for training the subsequent autoencoder, which models the embedding. For further processing, the instrument class was omitted as this class does not necessarily correlate with the position on the surgical path.

2.3   Embedding and Camera-Pose

The encoder of the AE comprises multi-head attention layers followed by fully connected layers, ultimately reducing the input to 3 latent dimensions, where z_t^1 represents the position on the surgical path, and z_t^2 and z_t^3 represent pitch and yaw angles, which represent the rotation angles of the camera with respect to a centred view. A transformer-based encoder is employed to encode the temporal information in the sequence of detections.

The decoder consists of two fully connected networks, the first generating the class probabilities $\hat{\mathbf{y}}_t$ of $\hat{\mathbf{c}}_t$, and the second generating the corresponding bounding boxes $\hat{\mathbf{b}}_t^I$ from z_t^1 in the centred view. To obtain bounding box reconstructions in the observed view, the centre coordinates of the predicted bounding boxes $\hat{\mathbf{b}}_t^I$ are rotated by multiplying them with the rotation matrix $\hat{\mathbf{R}}_t$, as explained in Section 2.1, to obtain $\hat{\mathbf{b}}_t$, the second component of $\hat{\mathbf{c}}_t$. The AE is designed to reconstruct only the last frame $\mathbf{c}_t$ in $\mathbf{C}_t$ since z_t^1 is intended to correspond to the current position. However, it considers s previous frames to provide additional information in determining the latent representation $\mathbf{z}_t$ of $\mathbf{x}_t$.

The loss function consists of a classification loss and a bounding box loss, the latter being calculated only for the classes present in the ground truth. In the

current setup, the fully connected network that produces the bounding boxes can classify any view as the centred view, since the bounding box coordinates can be rotated to fit the input even if the centred view is not at pitch and yaw angles of zero. Therefore, for increased interpretability, we enforce that the output is centred by feeding the output of the bounding box network together with $\mathbf{y}_t$ once again into the encoder by stacking the output s-times to achieve the same input size, and add the predicted $\hat{\mathbf{z}}_t^2$ and $\hat{\mathbf{z}}_t^3$, which should both be zero if the view is centred, to the loss function. This leads to the objective to minimize for the t-th frame in the m-th training video:

$$
\begin{aligned}
\mathcal{L}_{m,t} = &-\sum_{i=1}^{n} \left(y_{m,t}^i \log\left(\hat{y}_{m,t}^i\right) + \left(1 - y_{m,t}^i\right) \log\left(1 - \hat{y}_{m,t}^i\right) \right) \\
&+ \sum_{i=1}^{n} y_{m,t}^i \left| \mathbf{b}_{m,t}^i - \hat{\mathbf{b}}_{m,t}^i \right| + \sqrt{\left(\hat{z}_{m,t}^2\right)^2} + \sqrt{\left(\hat{z}_{m,t}^3\right)^2},
\end{aligned}
$$

where $|\cdot|$ denotes the l_1 loss.

The total training loss is then the sum of $\mathcal{L}_{m,t}$ over all frames and training videos. The proposed loss function can be viewed as maximizing the joint likelihood of a given $\mathbf{y}$ and $\mathbf{b}$ with a probabilistic model utilizing a mixture model for the bounding boxes.

3   Experiments and Results

3.1   Datasets

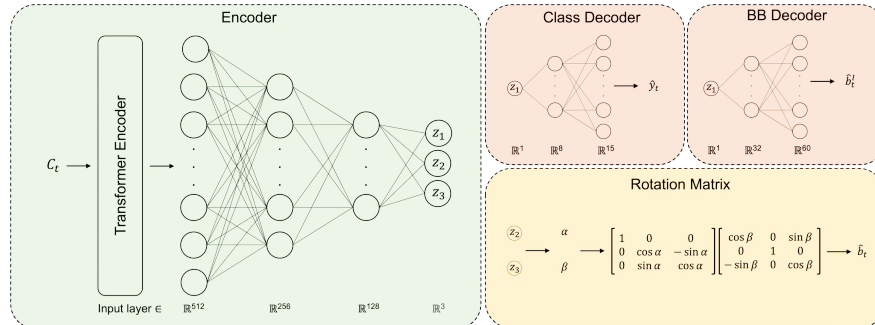In this work two datasets were utilized, a medical dataset and a synthetic dataset.

Fig. 1. The model comprises an encoder and a decoder that consists of two fully connected networks. The encoder takes $\mathbf{C}_t$ as input and embeds this sequence into a 3D latent representation. The decoder consists of two fully connected networks to generate the class probabilities $\hat{\mathbf{y}}_t$ and the bounding box coordinates $\hat{\mathbf{b}}_t^I$ from z_t^1. Furthermore, the encoder outputs z_t^2 and z_t^3 that are used to construct a rotation matrix to rotate the predicted bounding boxes around the pitch and yaw axes.
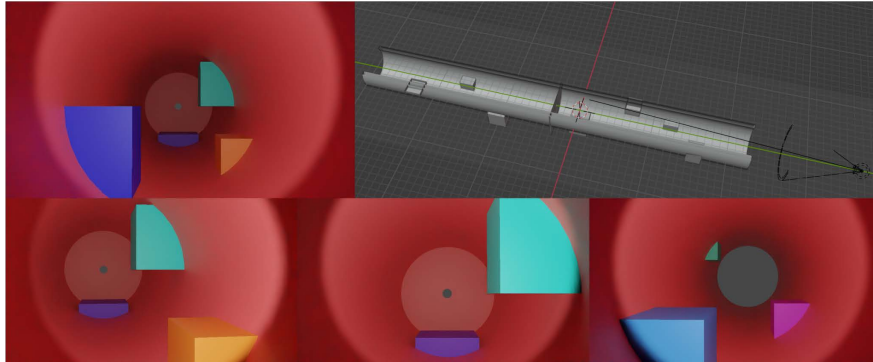
Fig. 2. An overview of the model used for the creation of the synthetic dataset.

Medical Dataset: The dataset utilized for object detection contains 166 videos documenting transsphenoidal adenomectomy procedures in 166 patients. These videos, captured using a variety of endoscopes across multiple facilities over 10 years, were made accessible under general research consent. Expert neurosurgeons annotated the videos, encompassing 16 distinct classes, namely, 15 anatomical structure classes and one class for surgical instruments. The dataset encompasses approximately 19,000 labelled frames. Each class has one instance per video, except for the instrument class since various instruments are categorized under the same class. Among the 166 videos, 146 were allocated for training and validation purposes, while the remaining 20 were purposed for testing. Despite the utilization of data from various centres, it is important to acknowledge potential biases introduced by the geographical vicinity of these centres.

Synthetic Dataset: For quantitative analysis, a synthetic dataset was created in Blender [21] with ground truth. A 3D environment was built to represent a surgical path with various structures. Ground-truth object detection labels could be extracted from the software. Eight different objects were modelled, akin to different anatomical structures, with a single instance per object to emulate the medical setting. To train the AE, a video moving through the environment was recorded with random viewing directions, moving forward and backwards several times. In total, the data consists of 16502 frames with corresponding ground-truth object detection labels. The model is depicted in Figure 2.

3.2   Implementation Details

The YOLO network was trained with identical parameters and implementation as in [18], which follows standard implementation as reported in [20].

The AE integrates a transformer encoder comprising six transformer encoder layers, each with five heads, and an input size of $s \times 15 \times 5$, where s is established as 64 frames. Following this, the output dimension of the transformer encoder undergoes reduction through three fully connected layers to sizes of 512, 256,

and 128, respectively, employing rectified linear unit (ReLU) activation functions between layers. Subsequently, the final fully connected layer reduces the dimensionality to $\mathbf{z}_t \in \mathbb{R}^3$, employing a sigmoid activation function to obtain z_t^1, and a tanh activation function for z_t^2 and z_t^3. Moreover, the decoder composed of two fully connected networks, namely the class decoder and bounding box decoder, have two fully connected layers, elevating the dimensionality of the latent variable z_t^1 from 1D to 8, to $\mathbf{n}$, and from 1D to 32, to $\mathbf{n} \times 4$, correspondingly. The initial layer of both of those networks is succeeded by a ReLU activation function, while the final layer adopts a sigmoid activation function. Furthermore, for z_t^2 and z_t^3, an output of -1 and 1 represents a -90 and 90 degrees rotation around the respective axis, these outputs are then converted to radians for the construction of the rotation matrix.

For the AE's training, the AdamW optimizer [22] is employed in conjunction with a warm-up scheduler, which linearly increases the learning rate from 0 to 1×10^{-4} over a span of 60 epochs. The model is trained for a total of 2500 epochs for the synthetic, and 270 epochs for the medical dataset. The model has approximately 4.6M parameters in the setup for the medical dataset. The inference times for the YOLO network and our encoder are 13 ms (1920 pixels, NVIDIA A100-40GB), and 20 ms (Intel Core i7-6700K), respectively. These times suggest that the combined inference time of 33 ms is suitable for real-time applications.

3.3   Results

Anatomical Detection: The YOLOv7 trained on the medical dataset reaches a mean average precision 53.4% at a 0.5 intersection over union threshold as also was reported in [18].
Quantitative Assessment of the Embedding: Due to the absence of all camera parameters, the exact modeling of rotation is an ill-posed task. However, we show that even though we introduce significant simplifications and substitute the homography matrix with a straightforward rotation matrix, we can still approximate the angles of rotation. We test on 1022 sequences which were recorded in Blender under random viewing angles moving through the model. We report a mean error in angle predictions of 0.43, and 0.69 degrees with a standard deviation of 2.38 and 1.74 degrees, for the pitch and yaw angle, respectively.

Additionally, we examine the correlation between the predicted location along the surgical path with the real depth of the synthetic model for a video travelling through the model. We do this for both our AE and the model proposed in [18]. We expect our AE that takes rotation into account to embed space more representative compared to an AE that does not consider rotation. More specifically, our AE should be able to map depth to the surgical path and account for different views by taking the rotation into account, whereas the previous model that only maps to the surgical path without rotation would need to occupy more space in the 1D latent space to describe various views of the same location and thereby describing depth less accurately. Our AE achieved a Pearson correlation coefficient of 0.97 whereas the previous model proposed in [18] achieved 0.94.
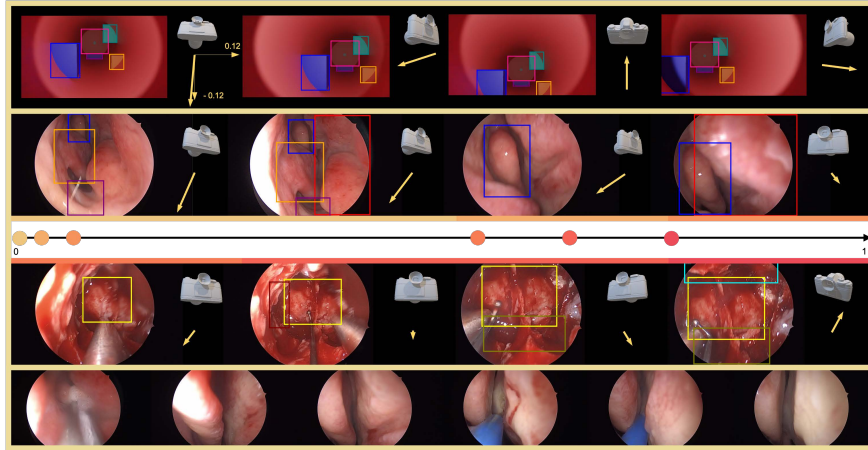**Qualitative Assessment of the Embedding:** Of greater importance is that

Fig. 3. Results are shown that depict the predicted viewing direction of the model for the sequences in the synthetic dataset (row 1), as well as the medical dataset (rows 2,3). Additionally, for the medical dataset, their predicted location along the surgical path is shown. The depicted cameras are for illustrative purposes only. Finally, in the bottom row images are shown that are mapped to the same location along the surgical path by the AE. We can see the same anatomical location from different points of view and during different stages of the surgery.

the model can tell the surgeon whether the viewing direction is right, or whether the camera should be pointed in another direction, and additionally can tell the surgeon whether in that direction should be looked more or less. In Figure 3 sequences are shown for the medical dataset, together with the predictions $\mathbf{z}_t$. The arrows plot the negative predictions of z_t^2 and z_t^3 to visualize the general direction the camera is pointed in and the magnitude. For illustrative purposes, a camera model is depicted to visualize the camera's orientation. The sequences visualize the ability of the model to extract various viewing directions from the images which correspond to the movement of the camera between the images in one sequence and that the predicted viewing directions are in line with the movement of visual landmarks between images within a sequence. The idea is to provide a reference visualization, such as an arrow, indicating the desired or planned viewing direction needed to locate a specific anatomical structure or for orientation purposes. This allows surgeons to adjust the endoscope's current orientation by referring to the arrow, as shown in Figure 3. They do not need to rely on exact angles; instead, they can determine the direction in which the endoscope should be pointed, adjusting it more or less as needed. Videos visualizing the results are supplied as supplementary material.

Finally, in the bottom row of Figure 3, we depict images that are mapped to the same location by our AE. These images show the same anatomical position during different stages of the surgery as well as from different viewpoints. When encoding the same images using the AE proposed in [18], these images

are mapped within a range of 0.81% of the latent space instead of a single point. This demonstrates that the proposed AE that takes rotation into account can embed space more coherently and confirms the results of quantitative correlation experience.

4   Conclusion

In this study, we introduced an approach to neuronavigation leveraging deep learning techniques. Our proposed method is image-based and utilizes bounding box detections of anatomical structures to orient itself within a surgical path learned from a dataset of surgical videos, and additionally provides feedback on the direction the camera, i.e. endoscope is pointing in. This is facilitated through an AE architecture trained without supervision. Our approach enables the localization and prediction of anatomical structures along the surgical trajectory, both forward and backwards, similar to functionalities seen in mapping applications.

However, our work also comes with certain limitations. Primarily, we have confined our focus to a single surgical procedure in this preliminary investigation. Extending to other surgical procedures is a goal for future research. Additionally, our proposed method can be integrated with SLAM techniques, as well as leveraging guidance from pre- or intra-operative MRI. Both of these extensions constitute areas of future exploration. Another limitation lies in the fact that the latent dimension only offers relative positional and angular encoding. To surpass this limitation, additional labelling of the actual positions along the surgical path may be necessary, as well as camera parameters.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Härtl, R., Lam, K.S., Wang, J., Korge, A., Kandziora, F., Audigé, L.: Worldwide survey on the use of navigation in spine surgery. World neurosurgery **79**(1), 162–172 (1 2013)
2. Orringer, D.A., Golby, A., Jolesz, F.: Neuronavigation in the surgical management of brain tumors: current and future trends. Expert review of medical devices **9**(5), 491–500 (9 2012)
3. Iversen, D.H., Wein, W., Lindseth, F., Unsgård, G., Reinertsen, I.: Automatic Intraoperative Correction of Brain Shift for Accurate Neuronavigation. World neurosurgery **120**, e1071–e1078 (12 2018)
4. Berkmann, S., Schlaffer, S., Nimsky, C., Fahlbusch, R., Buchfelder, M.: Intraoperative high-field MRI for transsphenoidal reoperations of nonfunctioning pituitary adenoma. Journal of neurosurgery **121**(5), 1166–1175 (11 2014)
5. Staartjes, V.E., Volokitin, A., Regli, L., Konukoglu, E., Serra, C.: Machine Vision for Real-Time Intraoperative Anatomic Guidance: A Proof-of-Concept Study in Endoscopic Pituitary Surgery. Operative neurosurgery (Hagerstown, Md.) **21**(4), 242–247 (10 2021)

6. Burkhardt, J.K., Serra, C., Neidert, M.C., Woernle, C.M., Fierstra, J., Regli, L., Bozinov, O.: High-frequency intra-operative ultrasound-guided surgery of superficial intra-cerebral lesions via a single-burr-hole approach. Ultrasound in medicine & biology **40**(7), 1469–1475 (2014)

7. Ulrich, N.H., Burkhardt, J.K., Serra, C., Bernays, R.L., Bozinov, O.: Resection of pediatric intracerebral tumors with the aid of intraoperative real-time 3-D ultrasound. Child's nervous system : ChNS : official journal of the International Society for Pediatric Neurosurgery **28**(1), 101–109 (1 2012)

8. Stummer, W., Stepp, H., Wiestler, O.D., Pichlmeier, U.: Randomized, Prospective Double-Blinded Study Comparing 3 Different Doses of 5-Aminolevulinic Acid for Fluorescence-Guided Resections of Malignant Gliomas. Neurosurgery **81**(2), 230–239 (8 2017)

9. Hadjipanayis, C.G., Widhalm, G., Stummer, W.: What is the Surgical Benefit of Utilizing 5-Aminolevulinic Acid for Fluorescence-Guided Surgery of Malignant Gliomas? Neurosurgery **77**(5), 663–673 (8 2015)

10. Hervey-Jumper, S.L., Li, J., Lau, D., Molinaro, A.M., Perry, D.W., Meng, L., Berger, M.S.: Awake craniotomy to maximize glioma resection: methods and technical nuances over a 27-year period. Journal of neurosurgery **123**(2), 325–339 (8 2015)

11. De Witt Hamer, P.C., Robles, S.G., Zwinderman, A.H., Duffau, H., Berger, M.S.: Impact of intraoperative stimulation brain mapping on glioma surgery outcome: a meta-analysis. Journal of clinical oncology : official journal of the American Society of Clinical Oncology **30**(20), 2559–2565 (7 2012)

12. Sanai, N., Mirzadeh, Z., Berger, M.S.: Functional outcome after language mapping for glioma resection. The New England journal of medicine **358**(1), 18–27 (1 2008)

13. Staartjes, V.E., Stumpo, V., Kernbach, J.M., Klukowska, A.M., Gadjradj, P.S., Schröder, M.L., Veeravagu, A., Stienen, M.N., van Niftrik, C.H., Serra, C., Regli, L.: Machine learning in neurosurgery: a global survey. Acta Neurochirurgica **162**(12), 3081–3091 (12 2020)

14. Grasa, O.G., Civera, J., Montiel, J.M.: EKF monocular SLAM with relocalization for laparoscopic sequences. Proceedings - IEEE International Conference on Robotics and Automation pp. 4816–4821 (2011)

15. Ozyoruk, K.B., Gokceler, G.I., Bobrow, T.L., Coskun, G., Incetan, K., Almalioglu, Y., Mahmood, F., Curto, E., Perdigoto, L., Oliveira, M., Sahin, H., Araujo, H., Alexandrino, H., Durr, N.J., Gilbert, H.B., Turan, M.: EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Medical Image Analysis **71**, 102058 (7 2021)

16. Mahmoud, N., Cirauqui, I., Hostettler, A., Doignon, C., Soler, L., Marescaux, J., Montiel, J.M.: ORBSLAM-based Endoscope Tracking and 3D Reconstruction. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **10170 LNCS**, 72–83 (8 2016)

17. Leonard, S., Sinha, A., Reiter, A., Ishii, M., Gallia, G.L., Taylor, R.H., Hager, G.D.: Evaluation and Stability Analysis of Video-Based Navigation System for Functional Endoscopic Sinus Surgery on In Vivo Clinical Data. IEEE Transactions on Medical Imaging **37**(10), 2185–2195 (10 2018)

18. Sarwin, G., Carretta, A., Staartjes, V., Zoli, M., Mazzatenta, D., Regli, L., Serra, C., Konukoglu, E.: Live image-based neurosurgical guidance and roadmap generation using unsupervised embedding. Information Processing in Medical Imaging. IPMI 2023. Lecture Notes in Computer Science, **13939**, 107–118 (06 2023)

19. Das, A., Khan, D.Z., Williams, S.C., Hanrahan, J.G., Borg, A., Dorward, N.L., Bano, S., Marcus, H.J. and Stoyanov, D.: A Multi-task Network for Anatomy Identification in Endoscopic Pituitary Surgery. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. MICCAI 2023. Lecture Notes in Computer Science, **14228**, 472–482 (10 2023)
20. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors (7 2022)
21. Community, B. O. (2018). Blender - a 3D modelling and rendering package. Stichting Blender Foundation, Amsterdam. Retrieved from http://www.blender.org
22. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. 7th International Conference on Learning Representations, ICLR 2019 (11 2017)