Depth-Driven Geometric Prompt Learning for Laparoscopic Liver Landmark Detection

Jialun Pei*[1], Ruize Cui*[2], Yaoqian Li*[1], Weixin Si[3(✉)], Jing Qin[2], Pheng-Ann Heng[1]

[1]The Chinese University of Hong Kong, Hong Kong, China
[2]The Hong Kong Polytechnic University, Hong Kong, China
[3]Shenzhen Institute of Advanced Technology, CAS, Shenzhen, China
wx.si@siat.ac.cn

Abstract. Laparoscopic liver surgery poses a complex intraoperative dynamic environment for surgeons, where remains a significant challenge to distinguish critical or even hidden structures inside the liver. Liver anatomical landmarks, *e.g.*, ridge and ligament, serve as important markers for 2D-3D alignment, which can significantly enhance the spatial perception of surgeons for precise surgery. To facilitate the detection of laparoscopic liver landmarks, we collect a novel dataset called **L3D**, which comprises 1,152 frames with elaborated landmark annotations from surgical videos of 39 patients across two medical sites. For benchmarking purposes, 12 mainstream detection methods are selected and comprehensively evaluated on L3D. Further, we propose a depth-driven geometric prompt learning network, namely **D^{2}GPLand**. Specifically, we design a Depth-aware Prompt Embedding (DPE) module that is guided by self-supervised prompts and generates semantically relevant geometric information with the benefit of global depth cues extracted from SAM-based features. Additionally, a Semantic-specific Geometric Augmentation (SGA) scheme is introduced to efficiently merge RGB-D spatial and geometric information through reverse anatomic perception. The experimental results indicate that D^{2}GPLand obtains state-of-the-art performance on L3D, with 63.52% DICE and 48.68% IoU scores. Together with 2D-3D fusion technology, our method can directly provide the surgeon with intuitive guidance information in laparoscopic scenarios. Our code and dataset are available at https://github.com/PJLallen/D2GPLand.

Keywords: Anatomical landmark detection · Laparoscopic liver surgery · Landmark dataset · SAM · RGB-D prompt learning.

1 Introduction

Laparoscopic liver surgery allows surgeons to perform a variety of less invasive liver procedures through small incisions, enabling faster patient recovery and superior cosmetic outcomes [22]. However, it is difficult for surgeons to distinguish

* Equal contribution.

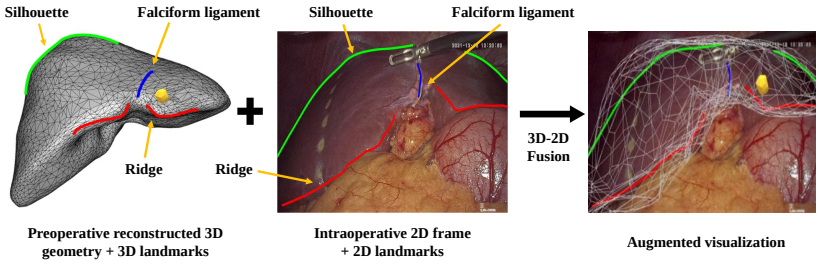

Fig. 1: Augmented visualization of liver tumor in the laparoscopic video via anatomic landmarks. With consistent anatomical landmarks on 2D frames (middle) and 3D geometry (left), the preoperative 3D anatomy can be overlaid on the intraoperative 2D image for augmented visualization guidance (right).

critical anatomical structures in the complex and variable laparoscopic surgical environment, making it heavily dependent on the experience of the surgeon. In this regard, augmented reality techniques tailored for laparoscopic liver surgery are urgently desired to provide surgeons with auxiliary information for precise resection and surgical risk reduction. The primary step in achieving augmented reality clues is to automatically identify guiding markers on key frames from intraoperative 2D videos and preoperative 3D anatomy samples, respectively, to assist in intraoperative decision-making. Liver anatomical landmarks, e.g., anterior ridge and falciform ligament, have been validated as effective consistent information for 2D-3D alignment[15,14]. As shown in Fig. 1, using 2D and 3D landmarks as references, internal liver structures are available for intraoperative fusion for enhanced visual guidance. However, accurate laparoscopic landmark detection remains challenging due to the lack of annotated datasets and how to comprehensively exploit the geometric information in video frames.

Traditionally, landmarks in laparoscopic augmented reality are defined as points or contours [16,8]. In intricate surgical environments, however, the performance of existing structure-based methods suffers from the instability of detection accuracy due to susceptibility to interruptions and tissue deformation together with the lack of global geometric information [18]. Additionally, traditional landmarks fail to provide semantic information for precise correspondence between 2D and 3D medical images, which has great importance for estimating cross-dimensional spatial relationships in laparoscopic liver surgery. To address these challenges, we adapt silhouettes, ridges, and ligaments from laparoscopic video frames as landmarks, which are continuous anatomies with clear semantic features in the preoperative 3D anatomy, facilitating efficient 2D-3D alignment.

However, existing laparoscopic liver landmark datasets lack sufficient annotations for training deep learning-based landmark models [1,20,17]. To address the limited sample of liver landmarks, we build the current largest-scale laparoscopic liver landmark dataset, named L3D. Specifically, we invite four senior surgeons to select 1152 critical frames from surgical videos of 39 patients at two medical sites, while labeling each frame with three types of semantic landmarks.

Table 1: Statistics of the L3D dataset. l, r, s are ligament, ridge, and silhouette.

Annotations	Frames	Tumor locations	Cases	Tumor sizes (mm)	Cases
$[l, r, s]$	1,056	Quadrate lobe	7	10-19	3
$[l, r]$	3	Left lobe	11	20-29	10
$[r, s]$	74	Right lobe	21	30-39	16
$[l, s]$	19	Caudate lobe	0	40-49	10

Based on the proposed L3D dataset, we contribute a systematic study on 12 mainstream baselines [21,15,26,5,29,23,4,2,24,28,11,3]. We observe that existing detection methods concentrate more on semantic feature capture and edge detection while ignoring global geometric features of the liver region, especially the depth information [13,15,14]. Hence, we delve into a straightforward and effective framework that leverages depth maps and pre-trained large vision models to enhance the accuracy of detecting laparoscopic liver landmarks.

In this work, we introduce a depth-driven geometric prompt learning network called D^{2}GPLand. Specifically, we first employ an off-the-shelf depth estimation model to generate depth maps that provide inherent anatomic information. Considering that Segment Anything Model (SAM)-based approaches [12,25] have shown superior performance in extracting global high-level features in surgical scenes, we adopt a pre-trained SAM encoder combined with the CNN encoder to respectively extract RGB multi-level features and depth geometric information. Then, a *Bi-modal Feature Unification (BFU)* module is designed to integrate RGB and depth features. To distinguish highly similar landmark characteristics in laparoscopic liver surgery, we propose a *Depth-aware Prompt Embedding (DPE)* operation to highlight geometric attributes guided by prompt contrastive learning and produce class-aware geometric features. Moreover, we propose a *Semantic-specific Geometric Augmentation (SGA)* scheme to effectively fuse class-aware geometric features with RGB-D spatial features, where a reverse anatomic attention mechanism is embedded to focus on the perception of anatomical structures and overcome the difficulty of capturing ambiguous landmarks. Extensive experimental results on the L3D benchmark show that D^{2}GPLand achieves a promising performance. Our method has great potential to be applied in augmented reality-based intra-operative guidance for laparoscopic liver surgery.

2 L3D Dataset

To facilitate the detection of laparoscopic liver landmarks, we establish a landmark detection dataset, termed L3D. Relevant information about patients and annotation is shown in Table 1. To provide enhanced visualization guidance efficiently during the ever-changing surgical environment, we extract key frames from laparoscopic liver surgery videos to annotate liver landmarks according to the suggestions of surgeons. To this end, four surgeons are invited to select key frames and label them, two of whom perform the labeling and the other two check the labels. The selection criterion for the keyframes is to allow the surgeon

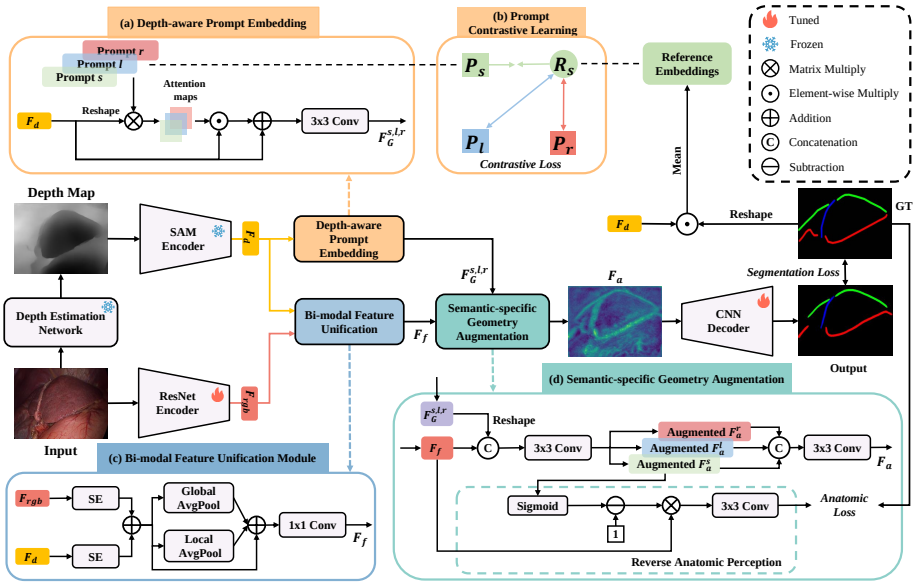

Fig. 2: Overview of the proposed D^{2}GPLand. s, l, r denote the three types of landmarks, silhouette, ligament, and ridge, to be detected.

to observe the global view of the liver, which can greatly reduce anatomical misperception during complex laparoscopic liver surgery. In our dataset, the ridge landmark is defined as the lower anterior ridge of the liver, and the ligament landmark is defined as the junction between the falciform ligament with the liver. In addition, the visible silhouette is also considered as a landmark category.

Our dataset is collected from two medical sites, and all surgeries are liver resections for hepatocellular carcinoma (HCC). The annotators screen 1,500 initial frames from 39 patient surgery videos with an original resolution of 1920*1080, and retain 1,152 key frames after checking. We divide all samples in L3D into three sets, where 921 images are used as the training set, 122 images as the validation set, and 109 images as the test set. To ensure the fairness of the experiment, images from the same patient are not shared across these sets.

3 Methodology

Fig. 2 outlines the architecture of the proposed D^{2}GPLand. Our model first takes key frame images from laparoscopic liver surgery as inputs and further generates depth maps using an off-the-shelf depth estimation network (AdelaiDepth [27]) as auxiliary inputs to supplement the geometric information. Then, we employ a ResNet-34 encoder [9] for RGB spatial feature extraction together with a frozen SAM encoder [12] for depth geometric cue acquisition. Notably, the original RGB frames are encoded through a CNN encoder to capture lower-level fea-

tures for anatomical structure identification, while depth maps mainly provide global shape attributes and geometric insights. Thanks to the transformer-based structure and pre-training with large amounts of natural images, the SAM encoder exhibits heightened sensitivity towards global geometric features from the depth modality. We conduct ablation studies for different encoder combinations in Sec. 4.3. Subsequently, depth feature F_d is passed into the proposed Depth-aware Prompt Embedding (DPE) module to highlight geometric attributes under the guidance of semantic prompts and then output the class-aware geometric features $F_G^{s,l,r}$. In parallel, the Bi-modal Feature Unification (BFU) module is applied to incorporate RGB feature F_{rgb} and F_d, producing integrated features F_f. Then, we interact geometric features $F_G^{s,l,r}$ focusing on different landmark categories with the fused RGB-D features through our Semantic-specific Geometric Augmentation (SGA) scheme to obtain augmented unified features F_a. Finally, a CNN decoder is used to produce the detection maps. The following subsections will elaborate on the key components of D^{2}GPLand.

3.1 Depth-aware Prompt Embedding

To capitalize on the advantages of pre-trained foundation models while reducing the computational costs for fine-tuning, we maintain the SAM encoder frozen in our model. Nonetheless, it still requires further guidance for extracting semantic geometry features related to landmark anatomy. To address this challenge, we propose three randomly initialized efficient class-specific geometric prompts and the DPE module to guide the extraction of geometric information related to different classes from the features derived from the SAM encoder. As shown in Fig. 2(a), we initially execute matrix multiplication between the input F_d and the geometric prompts, generating spatial attention maps to highlight regions associated with specific classes. Moreover, for each attention map, an element-wise multiplication is applied to depth features with a residual operation to obtain class-activated geometric features $F_G^{s,l,r}$.

In addition, the proposed DPE module relies on discriminative prompts to guide the class-specific geometry feature extraction. However, it is challenging to learn precise class-specific prompts due to the highly similar landmark characteristics of the liver. To enhance prompt discriminativeness for better guidance, we apply the contrastive learning technique as illustrated in Fig. 2(b). Here we take the silhouette prompt P_s as an example. Given the ground truth of the silhouette landmark and F_d, a dot product is conducted on them, followed by taking the channel-wise mean values to obtain the reference embeddings R_s. Upon obtaining all reference embeddings of the three landmark classes, we modify the NT-Xent Loss [7] as the contrastive loss, formulated as follows:

$$\mathcal{L}_{cl} = \frac{1}{N}\sum_{l\in L} log\frac{exp(P_l \cdot R_l/\tau)}{\sum_{k\in L} exp(P_l \cdot R_k/\tau)}, \tag{1}$$

where $N = 3$ is the number of classes, $L = \{s, l, r\}$ denotes the set of all classes, and τ refers to the temperature-scaled parameter. This contrastive learning strategy enhances the distinctiveness of the class-specific prompt representations.

3.2 Geometry-enhanced Cross-modal Fusion

Bi-modal Feature Unification. To capture holistic landmark features, we propose a BFU module to merge CNN-based lower-level structural features and SAM-based global geometric features. As depicted in Fig. 2(c), we first adaptively adjust the channel weights of F_{rgb} and F_d with Squeeze and Excitation (SE) blocks [10] and add them together. Afterward, we embrace the local and global average pooling modules to unify F_{rgb} and F_d at different scales and output the fused feature F_f.

Semantic-specific Geometry Augmentation. To further inject the class-activated geometric information from feature $F_G^{s,l,r}$ into the fused feature F_f, we present the SGA scheme shown in Fig. 2(d). We concatenate each class-specific feature in $F_G^{s,l,r}$ with the fused feature F_f respectively, and then obtain the corresponding augmented feature $F_a^{s,l,r}$ by 3×3 convolutional block. Subsequently, we concatenate all three semantic geometric features and generate the final augmented feature F_a. Considering the high similarity between anatomical structure and surrounding tissue features, we also embed a reverse anatomical perception module in the SGA to improve the sensitivity to ambiguous anatomical structures. Inspired by reverse attention[6,19], we apply a sigmoid function and reverse the attention weights to yield the anatomic attention maps. Afterward, we interplay the attention map with F_f via element-wise multiplication to predict anatomical features. Here, we use the dice loss as the anatomic Loss $\mathcal{L}_{ana}$ to supervise the anatomic learning.

3.3 Loss Function

In addition to the above-mentioned contrast loss and anatomic loss, we also add the segmentation loss $\mathcal{L}_{seg}$ to the overall loss function to supervise the final landmark detection map. In summary, the total loss function can be defined as:

$$\mathcal{L}_{total} = \lambda_{seg}\mathcal{L}_{seg} + \lambda_{cl}\mathcal{L}_{cl} + \lambda_{ana}\mathcal{L}_{ana}, \tag{2}$$

$$\mathcal{L}_{seg} = \frac{1}{N}\sum_{l\in L}(\mathcal{L}_{dice}^{(l)} + \mathcal{L}_{bce}^{(l)}), \tag{3}$$

where $\mathcal{L}_{dice}^{(l)}$ denotes the Dice Loss, $\mathcal{L}_{bce}^{(l)}$ denotes the binary cross-entropy (BCE) loss. λ_{seg}, λ_{cl}, and λ_{ana} are the balancing parameters for $\mathcal{L}_{seg}$, $\mathcal{L}_{cl}$, and $\mathcal{L}_{ana}$, respectively. All balancing parameters are set to 1 for optimal performance.

4 Experiments

4.1 Implementation Details

The proposed D^{2}GPLand is developed with PyTorch, and the training and testing processes are executed on a single RTX A6000 GPU. We run 60 epochs for training with a batch size of 4. A frozen pre-trained SAM-B [12] is implemented

Table 2: Comparison with state-of-the-art methods on L3D test set.

	Models	Model Params	DSC ↑	IoU ↑	Assd ↓
Non-SAM-based	UNet [21]	**7.84M**	51.39	36.35	84.94
	COSNet [15]	81.24M	56.24	40.98	69.22
	ResUNet [26]	69.73M	55.47	40.68	70.66
	UNet++ [29]	9.16M	57.09	41.92	74.31
	HRNet [23]	9.64M	58.36	43.50	70.02
	DeepLabv3+ [5]	43.90M	59.74	44.92	60.86
	TransUNet [4]	71.01M	56.81	41.44	76.16
	SwinUNet [2]	27.17M	57.35	42.09	72.80
SAM-based	SAM-Adapter [24]	93.93M	57.57	42.88	74.31
	SAMed [28]	183.55M	62.03	47.17	61.55
	SAM-LST [3]	183.12M	60.51	45.03	68.87
	AutoSAM [11]	101.43M	59.12	44.21	62.49
	D^{2}GPLand	139.03M	**63.52**	**48.68**	**59.38**

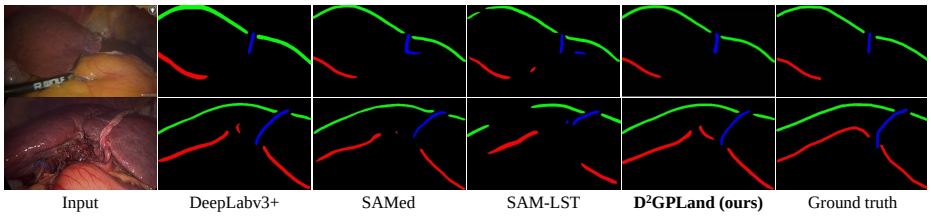

Fig. 3: Visualizations of our D^{2}GPLand and competitors on L3D test set.

in the depth encoder. We resize all the images to 1024×1024 and apply random flip, rotation, and crop for data augmentation. The Adam optimizer is used with the initial learning rate of 1e-4 and weight decay factor of 3e-5. In addition, the CosineAnnealingLR scheduler is applied to adjust the learning rate to 1e-6. For evaluation, we utilize the Intersection over Union (IoU), Dice Score Coefficient (DSC), and Average Symmetric Surface Distance (Assd) as evaluation metrics.

4.2 Comparison with State-of-the-Art Methods

We compare the proposed D^{2}GPLand with 12 cutting-edge methods on the L3D test set. For a fair comparison, these methods are divided into two types: (1) Non-SAM-based models, including UNet [21], COSNet [15], ResUNet [26], DeepLabV3+ [5], UNet++ [29], HRNet [23], TranUNet [4], and SwinUNet [2], and (2) SAM-based models, including SAM-Adapter [24], SAMed [28], AutoSAM [11], and SAM-LST [3]. All compared models were trained to converge with their official implementations. As shown in Table. 2, D^{2}GPLand outperforms competitors on all evaluation metrics. Compared to the top-ranked model SAMed, our method improves 1.51% on DSC, 1.49% on DSC, and 2.17 pixels on Assd metrics with 44.52M fewer parameters, demonstrating the effectiveness of utilizing depth-aware prompt and semantic-specific geometric augmentation for landmark detection. Besides, we observe that non-SAM-based methods ex-

Table 3: Ablations for key Designs.

Methods	BFU	DPE	$\mathcal{L}_{cl}$	SGA	DSC	IoU
M.1					59.34	44.90
M.2	✓				61.13	46.22
M.3	✓	✓			61.98	47.12
M.4	✓			✓	62.20	47.20
M.5	✓	✓	✓		62.41	47.34
M.6		✓	✓	✓	62.95	47.81
Ours	✓	✓	✓	✓	**63.52**	**48.68**

Table 4: Ablations for backbones.

Methods	Backbones	DSC	IoU
Dual CNN	ResNet-34	62.54	46.91
Dual SAM	SAM	62.97	47.63
SAM+CNN	ResNet-34(Depth) + SAM(RGB)	62.83	47.39
CNN+SAM	ResNet-34(RGB) + SAM(Depth)	**63.52**	**48.68**

hibit inferior performance compared to most SAM-based methods. It illustrates that the global geometric information extracted by the pre-trained SAM encoder can enhance the perception of landmark features. Fig. 3 also exhibits the visual results of D^{2}GPLand and other well-performed methods. We can see that our method provides more accurate detection of liver landmarks while mitigating the impact of occlusion by other tissues and surgical tools.

4.3 Ablation Study

Ablations for Key Designs. Table 3 shows the contribution of each key design in D^{2}GPLand on the L3D test set. Notably, all variants are trained with the same settings as mentioned in Sec. 4.1. The baseline (M.1) comprises a ResNet-34 encoder and frozen SAM-B encoder, and we directly concatenate RGB and depth features before feeding them into the decoder. Overall, each component contributes to the performance of our model in varying degrees. Specifically, M.2 and M.6 show the effectiveness of our BFU module in merging RGB and depth features. Based on M.2, M.3 and M.5 sequentially integrate our DPE and contrastive loss $\mathcal{L}_{cl}$ to further enhance the model performance. Further, M.4 adds the SGA scheme to M.2, resulting in 1.07% and 0.89% improvements in DSC and IoU, respectively, indicating the advantages of geometric cues.
Backbone Selections. To explore the effect of different backbones in feature extraction across RGB and depth modalities, we conduct additional ablation experiments on L3D with the CNN-based encoder and the SAM-based encoder. As shown in Table 4, D^{2}GPLand achieves the optimal performance when leveraging the ResNet-34 encoder for RGB inputs and the SAM encoder for depth modality. This experiment further validates the description in Sec. 3 that the ResNet-34 encoder is more effective in capturing lower-level anatomical structural features while SAM excels in extracting global geometric features.

5 Conclusion

This paper proposes a novel geometric prompt learning framework, D^{2}GPLand, for liver landmark detection on key frames of laparoscopic videos. Our method utilizes depth-aware prompt embeddings and semantic-specific geometric augmentation to explore the intrinsic geometric and spatial information, improving

the accuracy of landmark detection. Moreover, we release a new laparoscopic liver landmark detection dataset, L3D, to advance the landmark detection community. Experimental results indicate that D^{2}GPLand outperforms cutting-edge approaches on L3D, demonstrating the effectiveness of our method in capturing anatomical information in various surgeries. We hope this work can pave the way for extracting consistent anatomical information from 2D video frames and 3D reconstructed geometries, thereby directly promoting 2D-3D fusion and providing surgeons with intuitive guidance information in laparoscopic scenarios.

Acknowledgments. The work was supported in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No.: T45-401/22-N), in part by a grant from Hong Kong Innovation and Technology Fund (Project No.: MHP/085/21), in part by a General Research Fund of Hong Kong Research Grants Council (project No.: 15218521), in part by grants from National Natural Science Foundation of China (62372441, U22A2034), in part by Guangdong Basic and Applied Basic Research Foundation (2023A1515030268), in part by Shenzhen Science and Technology Program (Grant No.: RCYX20231211090127030), and in part by Guangzhou Municipal Key R&D Program (2024B03J0947).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Ali, S., Espinel, Y., Jin, Y., Liu, P., Güttner, B., Zhang, X., Zhang, L., Dowrick, T., Clarkson, M.J., Xiao, S., et al.: An objective comparison of methods for augmented reality in laparoscopic liver resection by preoperative-to-intraoperative image fusion. arXiv preprint arXiv:2401.15753 (2024)
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: ECCV (2022)
3. Chai, S., Jain, R.K., Teng, S., Liu, J., Li, Y., Tateyama, T., Chen, Y.w.: Ladder fine-tuning approach for sam integrating complementary network. arXiv preprint arXiv:2306.12737 (2023)
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
6. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. In: ECCV (2018)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
8. Collins, T., Pizarro, D., Gasparini, S., Bourdel, N., Chauvet, P., Canis, M., Calvet, L., Bartoli, A.: Augmented reality guided laparoscopic surgery of the uterus. IEEE TMI **40**(1), 371–380 (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2015)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE CVPR (2018)
11. Hu, X., Xu, X., Shi, Y.: How to efficiently adapt large segmentation model (sam) to medical images. arXiv preprint arXiv:2306.13731 (2023)

12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: IEEE ICCV (2023)
13. Koo, B., Robu, M.R., Allam, M., Pfeiffer, M., Thompson, S., Gurusamy, K., Davidson, B., Speidel, S., Hawkes, D., Stoyanov, D., et al.: Automatic, global registration in laparoscopic liver surgery. Int. J. Comput. Assist. Radiol. Surg. pp. 1–10 (2022)
14. Labrunie, M., Ribeiro, M., Mourthadhoi, F., Tilmant, C., Le Roy, B., Buc, E., Bartoli, A.: Automatic preoperative 3d model registration in laparoscopic liver resection. Int. J. Comput. Assist. Radiol. Surg. **17**(8), 1429–1436 (2022)
15. Labrunie, M., Pizarro, D., Tilmant, C., Bartoli, A.: Automatic 3d/2d deformable registration in minimally invasive liver resection using a mesh recovery network. In: MIDL (2023)
16. Li, W., Lu, Y., Zheng, K., Liao, H., Lin, C., Luo, J., Cheng, C.T., Xiao, J., Lu, L., Kuo, C.F., et al.: Structured landmark detection via topology-adapting deep graph learning. In: ECCV (2020)
17. Modrzejewski, R., Collins, T., Seeliger, B., Bartoli, A., Hostettler, A., Marescaux, J.: An in vivo porcine dataset and evaluation methodology to measure soft-body laparoscopic liver registration accuracy with an extended algorithm that handles collisions. Int. J. Comput. Assist. Radiol. Surg. **14**, 1237–1245 (2019)
18. Özgür, E., Koo, B., Le Roy, B., Buc, E., Bartoli, A.: Preoperative liver registration for augmented monocular laparoscopy using backward–forward biomechanical simulation. Int. J. Comput. Assist. Radiol. Surg. **13**, 1629–1640 (2018)
19. Pei, J., Cheng, T., Fan, D.P., Tang, H., Chen, C., Van Gool, L.: Osformer: One-stage camouflaged instance segmentation with transformers. In: ECCV (2022)
20. Rabbani, N., Calvet, L., Espinel, Y., Le Roy, B., Ribeiro, M., Buc, E., Bartoli, A.: A methodology and clinical dataset with ground-truth to evaluate registration accuracy quantitatively in computer-assisted laparoscopic liver resection. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **10**(4), 441–450 (2022)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
22. Schneider, C., Allam, M., Stoyanov, D., Hawkes, D., Gurusamy, K., Davidson, B.: Performance of image guided navigation in laparoscopic liver surgery–a systematic review. Surgical Oncology **38**, 101637 (2021)
23. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE TPAMI (2020)
24. Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
25. Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
26. Xiao, X., Lian, S., Luo, Z., Li, S.: Weighted res-unet for high-quality retina vessel segmentation. In: ITME. pp. 327–331. IEEE (2018)
27. Yin, W., Zhang, J., Wang, O., Niklaus, S., Chen, S., Liu, Y., Shen, C.: Towards accurate reconstruction of 3d scene shape from a single monocular image. IEEE TPAMI (2022)
28. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)

29. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE TMI **39**(6), 1856–1867 (2019)