POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Evaluation of SAM2 and EdgeSAM zero-shot capabilities for robotic tool segmentation in minimally invasive surgery

Author: Irene Caridi, Francesco Leni, Matteo Missana

Advisor: Elena De Momi

Co-advisor: Mattia Magro

Academic year: 2024-2025

1. Introduction

Minimally invasive surgery is a highly demanding task, where extreme precision and dexterity are needed to ensure the optimal outcome of the procedure. For that reason, many efforts are being made to create valuable assessing and teaching tools for the novel surgeon, as well as to enable robotics to become the standard for many new surgical fields [5]. Microsurgery is one area where robots can be introduced into current standard practices, leading to better performance in extremely tight and intricate environments, reducing the surgeon's mental load. Moreover, these robotics solutions can effectively enhance surgeon capabilities by scaling his/her movements, removing tremors and, when used in conjunction with Artificial Intelligence (AI), can also allow integration of visual data to improve robot control strategy or to provide meaningful information about the ongoing procedure to the operating surgeon [13]. Hence, there is a strong need to develop robust and automated software tools to efficiently detect and segment surgical tools [24]. In recent years, Deep Learning (DL) has emerged as a promising solution for achieving such a goal, as it allows to learn meaningful semantic information by processing a vast amount of visual data. Despite its remarkable results, two major drawbacks are still affecting DL: the nontrivial trade-off between efficiency and performances; and the extreme dependence on huge amounts of training data that, especially in the medical field, are often very hard to acquire. Whereas Vision Transformers (ViT) [7] are becoming the standard in many imaging scenarios, their large dimensions make them hardly suitable to real-time constraints, as surgical practice would require. However, large pre-trained transformer-based foundation models [2] are generating a lot of interest due to their claimed ability to reach state-of-the-art performances even in zero-shot deployment. In the tentative of solving the efficiency-performance trade-off, Knowledge Distillation (KD) emerged as a promising tool [10]. In this process, a lightweight model, called student, is trained based on the knowledge of a large pre-trained model, called teacher.

Therefore, this work will focus on understanding the real potentialities of zero-shot transfer for Vision foundation models, while also investigating whether KD could really offer a valuable solution to reduce models size. Therefore, the contribution of this work are twofold: first a zero-shot capabilities analysis of Segment Anything Model 2 (SAM2) [19], one of the major vision foundation model currently available, will be conducted, then its results will be compared to that of EdgeSAM [26], a novel model generated by distilling a CNN backbone from the original pre-trained SAM.

2. Related Works

Nowadays, the task of segmentation is mainly tackled with Neural Networks, especially classical Convolutional Neural Networks (CNN), which are based on convolutions as their main operation [15], but also with Vision Transformers (ViT) [7], that are, instead, based on attention [22]. Moreover, in the very last couple of years, transformer-based foundation models are becoming a reality also for Vision related tasks and, thanks to their zero-shot deployment capabilities, they are setting a new frontier in the field of segmentation.

2.1. CNN & ViT for segmentation

This class of models is strictly task-based, which means that the task they are trained on defines themselves. Moreover, the whole training process is centred on the particular class of problems these models are expected to solve. As we are addressing the segmentation problem, the training procedure is typically supervised and, hence, requires a lot of labelled data to be available [25]. For semantic segmnentation, i.e. categorization of each pixel in an image into a class or object, Unet-like [12, 20] architectures are still very popular and remain extremely valuable when the difficulty of the problem remains constrained. Instead, for instance segmentation, i.e. instances identification and classification before being segmented, mask-RCNN [9] is still a preferred choice, especially when fast inference time is required. ViTs, instead, have shown superior capabilities in handling vast and broad problems, where the attention mechanism efficiently allows them to learn very general features. They are typically built in an encoder-decoder structure, regardless of the addressed task being semantic or instance segmentation [3, 4].

2.2. Foundation models for segmentation

Foundation models [2] are transformer-based architecture pre-trained on broad data that, thanks to their promptable nature, can be adapted to a wide range of downstream tasks. Originally, they were born for Natural Language Processing (NLP) [6, 16], but their are now promisingly emerging, also, for visual tasks.
Dealing with vision, a pioneering work was CLIP [17], which managed to align visual and textual information using contrastive learning and paved the way for all the subsequent development in the field. It is worth noting, that, even if foundation models are in principle fine-tunable to better accomplish more specific tasks, due to their extreme complexity, this is usually not easily feasible as it would require a great amount of hardware and time resources. moreover, fine-tuning a model that is designed to be general purpose may, indeed, spoil its original generalizability by bounding it within a too narrow data space, so occurring into the same issues of traditional task-based models.

2.2.1 SAM & SAM2

Dealing with segmentation, SAM [14] is widely considered the first attempt to build a foundation model specifically designed for that task. It introduces a strong paradigm shift, by being the first model capable of performing general-purpose, promptable segmentation without the need for any task-specific training. It can accept either points, text, bounding boxes or masks as prompts and it is essentially composed by three main components: a heavy image encoder, a prompt encoder and finally a decoder to retrieve the final prediction. The image encoder is a ViT [7] pre-trained using Mask Auto Encoding (MAE) strategy [8], a self-supervised training method that consists of training a transformer to reconstruct a given input image while masking a gradually increasing number of patches throughout the training process. It runs once per image and is responsible for the generation of the most general and meaningful image embedding prior to prompting the model. The prompt encoder is responsible for the embedding of the different prompt inputs. When dealing with points and bounding boxes, the embedding is composed of positional encodings summed with learned embeddings for each prompt type. Text is embedded using an off-the-shelf text encoder from CLIP [17] and masks are embedded using convolutions and summed element-wise to the image embedding. The mask decoder is deputed to map the image embedding into a mask, guided by the prompt embeddings. This is achieved using prompt self-attention and cross-attention to update all embeddings. Finally, a Multi Layer perceptron (MLP) maps the output token to a dynamic linear classifier, which then computes

the mask foreground probability at each image location. SAM2 [19] was introduced with the same architecture of SAM, adding memory bank and memory encoder to exploit the relationships between subsequent frames to enhance its prediction capability in videos. The memory encoder is a convolutional module that downsamples the final predicted mask and sums it element-wise with the unconditioned frame embedding from the image-encoder to generate a so called memory. Memories of past frames are stored in the memory bank along with learned object pointers to track instances throughout frames and are used to condition the current frame via cross-attention. Thanks to this modifications, the user can prompt only the first frame, and the model automatically propagates the information across all the frames of the video.

2.3. Knowledge Distillation

KD is a training strategy in which a simpler model, called student, is guided to mimic the behavior of a more complex, typically pre-trained network, called teacher, and it is nowadays becoming standard practice to try bringing the gap between the efficiency of CNN and the generalizability of ViTs. It was originally designated for classification problems [10] where the canonical Cross-Entropy loss is added a further element to account for the teacher output. This additional term is referred as the distillation term and, in the classical formulation, is the Kullback-Leibler (KL) divergence between the softmaxed student's logits (Zs) and the softmaxed teacher's ones (Zt). The formula below shows the analytical formulation of such loss, where ϕ stands for softmax and τ is the temperature coefficient to smooth the comparison.

$$\mathcal{L}_{\text{distillation}} = \mathcal{L}_{\text{CE}} + \tau^2 \cdot \text{KL}\left(\phi\left(\frac{\mathbf{z}_s}{\tau}\right), \phi\left(\frac{\mathbf{z}_t}{\tau}\right)\right)$$

2.3.1 EdgeSAM

EdgeSAM [26] is a super light-weighted version of SAM especially designed to be deployed on edge devices such as smartphones. To do so, the ViT backbone of SAM has been distilled into a very efficient and light CNN called RepViT [23], minimizing the Mean Squared Error (MSE) respect SAM image embedding. The obtained backbone is then merged with the existing prompt and mask encoders and finally refined doing a novel prompt-in-the-loop distillation to allow really capturing the essence of SAM. This technique involves actively aligning the student model with the output masks of SAM by iteratively introducing new prompts in regions where the student model exhibits inaccuracies with respect to the teacher. Specifically, a positive point is drawn in regions marked as false negatives, or a negative point in areas identified as false positives.

3. Materials & Methods

3.1. Dataset

Instrument	Count
N° images	876
Grasper	896
Hook	346
Bipolar	33
Irrigator	25
Clipper	21
Scissors	9

Table 1: *Counts of surgical instruments in images dataset.*

The dataset used in this work is generated from the existing CholecInstanceSeg dataset [1], which includes frames extracted from Cholecystectomy videos collected in the CholectSeg8K [11], CholecT50 and CholecT80 datasets [21], annotated with semantic masks of the present surgical instruments. Specifically, it comprises 6 different classes: grasper, hook, bipolar, irrigator, clipper, scissor. When evaluating SAM2 and EdgeSAM on single images, we decided to under-sample the majority of those videos, as they were extracted at a very high frame rate, retaining only 1 frame every 15. This to try increasing the variability of data. More details about the obtained dataset are shown in Table 1, the unbalanced of the dataset is shown, which is due to the intrinsic nature of the addressed surgical procedure. For the video segmentation evaluation, we used a subset of the original CholecInstanceSeg dataset.

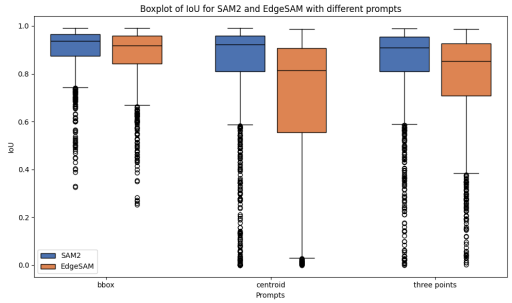


Figure 1: *Comparison of SAM2 and EdgeSAM along different prompting strategy.*

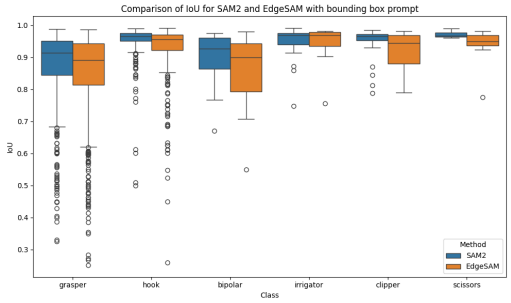


Figure 2: *Comparison of SAM2 and EdgeSAM with bounding box prompt along different surgical tools.*

This subset is divided into two groups. The first group consists of videos from CholecSeg8k, recorded at a frame rate of 25 fps, but containing only two of the six original classes: grasper and hook. Although SAM2 can segment videos, the process is very slow. To address this limitation, we used a second group of videos with a lower frame rate to test the feasibility of segmenting just one frame per second. This approach reduces computational cost and processing time, making it more suitable for near-real-time applications. The second group includes videos from CholecT50, which run at only 1 fps and contain all six classes. Results are analyzed separately for these two sub-datasets to better understand the method's performance on each type of video. Specifically, videos 12, 25, 35, 48, and 52 from CholecSeg8k and videos 01, 12, and 23 from CholecT50 were used for this evaluation.

3.2. Prompting strategy

To better study zero-shot capabilities of SAM2 and EdgeSAM, three different types of prompts have been examined: bounding boxes, centroid point and three points. All prompts are derived from the ground truth mask annotations and, to standardize the selection of the three points, the centerline of the instrument is identified and the centroid plus two points at 10% of the length from it are retrieved. Considering image segmentation, all the 3 prompting strategies are evaluated, then a per-instrument analysis is provided only for the prompt that proved to be the most effective, i.e. bounding boxes. For video evaluation, only bounding boxes are considered and the prompt is given just at the first frame of each video or whenever a new surgical tool appears in the scene. In contrast, objects disappearing from the scene are treated as routine occurrences that SAM2 is expected to handle.

4. Results

4.1. Image segmentation

Image segmentation evaluation of SAM2 and EdgeSAM is performed first by considering all surgical tools together and then analyzing each type of tool separately. Both models show remarkable performance in all the three different prompting strategies (fig. 1), with SAM2 achieving a best mean IoU of 90.21% when using bounding boxes and EdgeSAM achieving a mean IoU of 87.60% exploiting the same prompt type. While SAM2 consistently performs well in all three prompts, EdgeSAM's performance varies depending on the type of prompt, falling to mean IoU of 69.28% when using just the single centroid point.

Model	Tool	Metrics (%)		
		Mean	Median	IQR
SAM2	Grasper	88.00	91.41	10.71
	Hook	95.22	96.49	2.52
	Bipolar	90.19	92.65	9.66
	Irrigator	94.68	96.85	3.61
	Clipper	94.08	96.54	1.93
	Scissors	97.10	96.73	1.21
EdgeSAM	Grasper	85.37	89.11	12.90
	Hook	92.55	95.54	4.94
	Bipolar	86.13	89.87	15.08
	Irrigator	95.00	96.83	4.25
	Clipper	92.15	94.46	8.93
	Scissors	93.42	94.88	3.15

Table 2: *SAM2 and EdgeSAM across different tools using bounding boxs.*

Model	Prompt	Metrics (%)		
		Mean	Median	IQR
SAM2	Bounding box	90.21	93.73	8.94
	Centroid	82.57	92.16	14.95
	Three points	84.07	90.98	14.66
EdgeSAM	Bounding box	87.60	91.73	11.68
	Centroid	69.28	81.49	35.16
	Three points	78.06	85.37	21.72

Table 3: *Statistics of SAM2 and EdgeSAM across different prompts.*

The bounding box prompting strategy has the smallest IQR for both methods: 8.94% for SAM2 and 11.68% for EdgeSAM; in contrast, for points prompting, the IQR is bigger (table 3). These results underscore the stability of image segmentation using bounding boxes as prompts, as it demonstrates higher performance among the different methods. In addition, the smaller IQR for the prompt of the bounding boxes suggests that both methods show more consistent and reliable performance in instrument segmentation. Considering the bounding box prompting strategy as the best one, proficiency is evaluated along different types of surgical tools. Both models perform well across all six types of tools, demonstrating their robustness. Examining the various tools, while SAM2 performs the best overall, performance is still comparable due to high IoU scores, with averages generally above 0.9 for most classes, indicating strong segmentation performance. The performance across these tools can be further explored in fig. 2 and table 2.

4.2. Video segmentation

SAM2 predictions are analyzed separately for the two subdatasets used. The performance on CholecSeg8k is lower than for the single image segmentation discussed above, but still extremely solid, with a mean IoU of 68.68% for graspers and of 91.30% for hooks. The IQR is 22.10% for graspers and 3.80% for hooks, with the relative boxplots that can be seen in figure 4. Regarding the CholecT50 videos, with a frame rate of 1 fps, the results are not as good as for the former. Mean IoUs are reported in table 4 and boxplots emphasize poorer performance in this type of videos, as can be seen in figure 5.

Tool	Metrics (%)		
	Mean	Median	IQR
Grasper	53.04	72.89	84.45
Hook	84.14	90.63	7.60
Irrigator	66.63	91.46	94.01
Bipolar	65.38	79.51	26.19
Clipper	75.14	90.15	7.92
Scissors	67.90	88.48	33.48

Table 4: *Statistics of SAM2 across different tools on the CholecT50.*

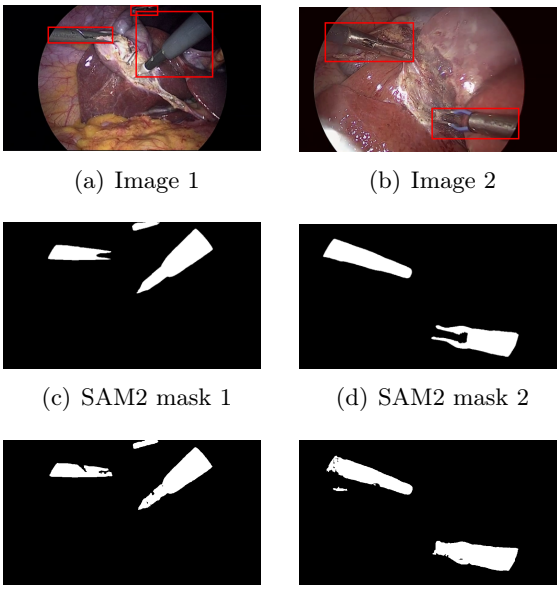
5. Discussion

5.1. Image segmentation

As the evaluation of SAM2 and EdgeSAM was conducted without any fine-tuning, thus in a perfect zero-shot fashion, this underlines the great generalization capabilities of these models, highlighting their potential for application in specific scenarios, such as minimally invasive surgery. Adaptability to different prompt strategies, in zero-shot deployment, is a key performance point of SAM2, in fact, the average IoU is 85.61% across different prompt strategies. Additionally, it is also very robust against different types of tools. On the other hand, EdgeSAM achieved comparable performance to SAM2 using bounding boxes, with a IoU of 87.60%, showing that a distilled version of SAM can achieve good segmentation results. As shown in Figure 3 both models, with the same prompt bounding box, extract the same more or less detailed masks. Another important aspect to consider is the computation time of both methods. EdgeSAM is significantly faster than SAM2, with a processing speed of 0.63 s/it compared to 2.32 s/it on a standard laptop CPU. EdgeSAM performance and computational time highlight the potential of knowledge distillation to create lighter and more efficient models without decreasing performance.

5.2. Video segmentation

Performance of SAM2 in videos is generally lower than in images, reflecting the lower prompt information provided to the model. Regarding the results on CholectSeg8k videos, the high frame rate allows the model to still retain good performance, even though the grasper class has



(a) Image 1 (b) Image 2

(c) SAM2 mask 1 (d) SAM2 mask 2

(e) EdgeSAM mask 1 (f) EdgeSAM mask 2

Figure 3: *Example of dataset's images with bounding box prompt and predicted masks for SAM2 abd EdgeSAM*

a high IQR and a lot of outliers that fall near 0. A limitation of this evaluation is the presence in the videos of only 2 classes of the 6 total present in the dataset. To overcome this, and to evaluate performance on lower frame rates, we tested the model on the CholecT50 subdataset. The lower frame rate of these videos is reflected in very poor performance, especially for grasper and irrigator classes. As seen from the boxplots, these classes have a very wide distribution, that reflects a great amount of IoUs near 0%. That's due to the model loosing track of an object at a certain frame in the video, with all the subsequent frames having an IoU of 0% or almost 0%. The lower frame rate could be the most important factor for the model to not loose track of the objects, because it means that they move visibly from one frame to the other. Another important thing to keep in mind is that SAM2 is not intrinsically built for real-time applications, and has an inference time per frame that is the same as for the image segmentation task. This means that a full frame video segmentation is not possible at the moment and, considering the low performance on the lower frame rate dataset, even segmenting a small amount of the total frames to save time, is not feasible.

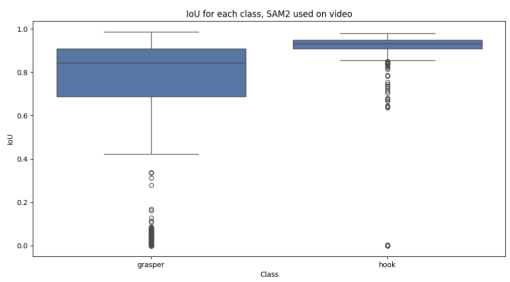


Figure 4: *IoU results of SAM2 on the CholecSeg8k videos.*

6. Conclusion

This study evaluates the capabilities of SAM2 [18] and EdgeSAM [26] for zero-shot surgical instrument segmentation, demonstrating foundation models abilities in segmenting surgical tools for minimally invasive surgery. SAM2 exhibits strong generalization abilities, achieving a high average IoU across different prompting strategies, showing its potential for applications in surgical environments. EdgeSAM, while being a distilled and computationally efficient version, performed comparably to SAM2 in the specific context, highlighting the effectiveness of knowledge distillation for creating a lighter model without significant performance degradation. The results showed that both models excelled in segmenting surgical instruments in images, particularly using bounding box prompts. SAM2 was also evaluated on video segmentation, showing lower performance, especially on lower frame rate videos, revealing challenges with object tracking and continuity under this condition, and suggesting the need for videos at a quite high frame rate to be effective. The computation time is substantially lower for EdgeSAM due to its lighter model design; however, it remains quite high for real-time segmentation applications.

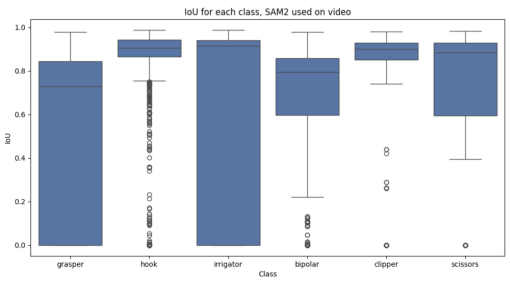


Figure 5: *IoU results of SAM2 on the CholecT50.*

References

[1] Oluwatosin Alabi, Ko Ko Zayar Toe, Zijian Zhou, Charlie Budd, Nicholas Raison, Miaojing Shi, and Tom Vercauteren. Cholecinstanceseg: A tool instance segmentation dataset for laparoscopic surgery. *arXiv preprint arXiv:2408.00714*, 6 2024.

[2] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. 5 2020.

[4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 12 2021.

[5] Giulio Dagnino and Dennis Kundrat. Robot-assistive minimally invasive surgery: trends and future directions, 12 2024.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. Bert: Pre-training of deep bidirectional transformers for language understanding.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 10 2020.

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2408.00714*, 11 2021.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. 3 2017.

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:2408.00714*, 3 2015.

[11] W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020.

[12] Lina Huang, Alina Miron, Kate Hone, and Yongmin Li. Segmenting medical images: From unet to res-unet and nnunet. 7 2024.

[13] Muhammad Iftikhar, Muhammad Saqib, Muhammad Zareen, and Hassan Mumtaz. Artificial intelligence: revolutionizing robotic surgery: review. *Annals of Medicine Surgery*, 86:5401–5409, 9 2024.

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2408.00714*, 4 2023.

[15] Yann Lecun and Yoshua Bengio. *Convolutional Networks for Images, Speech and Time Series*, pages 255–258. The MIT Press, 1995.

[16] Alec Radford Openai, Karthik Narasimhan Openai, Tim Salimans Openai, and Ilya Sutskever Openai. Improving language understanding by generative pre-training.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2408.00714*, 2 2021.

[18] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. Version 2.

[19] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 8 2024.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. 5 2015.

[21] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[23] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. *arXiv preprint arXiv:2408.00714*, 7 2023.

[24] Yan Wang, Qiyuan Sun, Zhenzhong Liu, and Lin Gu. Visual detection and tracking algorithms for minimally invasive surgical instruments: A comprehensive review of the state-of-the-art. *Robotics and Autonomous Systems*, 149, 3 2022.

[25] Wenjian Yao, Jiajun Bai, Wei Liao, Yuheng Chen, Mengjuan Liu, and Yao Xie. From cnn to transformer: A review of medical image segmentation models. 8 2023.

[26] Chong Zhou, Xiangtai Li, Chen Change Loy, and Bo Dai. Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam. *arXiv preprint arXiv:2408.00714*, 12 2023.