Can surgical computer vision benefit from large-scale visual foundation models?

Navid Rabbani and Adrien Bartoli

DIA2M, DRCI, CHU Clermont-Ferrand, France.

*Corresponding author(s). E-mail(s): navid_rabbani@yahoo.com;
Contributing authors: adrien.bartoli@gmail.com;

Abstract

Purpose. We investigate if foundation models pretrained on diverse visual data could be beneficial to surgical computer vision. We use instrument and uterus segmentation in minimally-invasive procedures as benchmarks. We propose multiple supervised, unsupervised and few-shot supervised adaptations of foundation models, including two novel adaptation methods.
Methods. We use DINOv1, DINOv2, DINOv2 with registers and SAM backbones, with the ART-Net surgical instrument and the SurgAI3.8K uterus segmentation datasets. We investigate five approaches: DINO unsupervised, few-shot learning with a linear decoder, supervised learning with the proposed DINO-UNet adaptation, DPT with DINO encoder, and unsupervised learning with the proposed SAM adaptation.
Results. We evaluate 17 models for instrument segmentation and 7 models for uterus segmentation, and compare to existing ad hoc models for the tasks at hand. We show that the linear decoder can be learned with few shots. The unsupervised and linear decoder methods obtain slightly subpar results but could be considered useful in data scarcity settings. The unsupervised SAM model produces finer edges but has inconsistent outputs. However, DPT and DINO-UNet obtain strikingly good results, defining a new state-of-the-art by outperforming the previous-best by 5.6 and 4.1 pp for instrument and 4.4 and 1.5 pp for uterus segmentation. Both methods obtain semantic and spatial precision, accurately segmenting intricate details.
Conclusion. Our results show the huge potential of using DINO and SAM for surgical computer vision, indicating a promising role for visual foundation models in medical image analysis, particularly in scenarios with limited or complex data.

Keywords: Visual foundation models, minimally-invasive surgery, segmentation

1 Introduction

The dominant paradigm in natural language processing (NLP) is to adapt a foundation model to the task at hand [1]. A foundation model is a large machine learning feature extraction model pretrained on a vast quantity of data and generally used without fine-tuning for the downstream task. Such large-scale models were recently introduced for computer vision, including DINOv2 [2], SAM [3], SEEM [4] and CLIP [5]. These models distilled the information from hundreds of millions of general internet-gathered images and were shown to adapt to many downstream tasks. The strength of these models comes from applying self-supervised methods over enough curated images from diverse sources. In medical images, such data are generally not widely available. As the domain shift between medical images and the general images might be large, the performance may be limited in medical computer vision. A natural question is thus whether one should shift from a task-specific to a foundation model in medical computer vision. We address this question for surgical computer vision, specifically for the tasks of instrument and uterus segmentation in abdominal minimally-invasive procedures. We use seven foundation models and propose several adaptations, including the combination with a specifically trained UNet.

2 Methods

2.1 General Points

We use several variants of DINOv1 [6], DINOv2 [2], DINOv2 with registers [7] and SAM [3] backbones with the pretrained parameters as distributed and without fine-tuning. All the tested foundation models were left frozen, through all the experiments. We have kept them frozen as they are huge models with hundreds of millions of parameters. They cannot be fine-tuned on low-end computing hardware and the point of foundation models is indeed to avoid this step. DINO backbones use the ViT architecture, taking an image as input and producing patch tokens as output. Specifically, the image is divided in non-overlapping patches and a token, which is a spatially localised high-level descriptor, is returned for each patch. DINOv1 is distributed in two model sizes ViT-S and ViT-B and two patch sizes of 8×8 and 16×16 pixels. DINOv2 is distributed in four model sizes, namely ViT-S, ViT-B, ViT-L and ViT-g, with one patch size of 14×14 pixels and with optional registers [7]. The registers are additional tokens to the input sequence of the Vision Transformer. They prevent artefacts corresponding to high-norm tokens appearing during inference, primarily in low-informative background areas of images. We selected the 'base' and 'large' backbones ViT-B and ViT-L, with 85M and 300M parameters and with token dimensions of 768 and 1024 respectively. We skipped the smallest backbones ViT-S as they lack accuracy [2], and the largest backbones ViT-g as they barely fit into most GPU memories and do not bring a noticeable benefit [2]. We tested all patch size options. We eventually selected the pretrained ViT-B/8 and ViT-B/16 backbones of DINOv1 and the pretrained ViT-L/14 with and without registers of DINOv2. We use the output patch tokens as features for the downstream task of semantic segmentation. The SAM model can either produce high quality object masks from input prompts such as points or boxes or can be used to generate masks for all objects in an image but without semantic labels. We propose to use a coarse semantic segmentator to assign the labels to the SAM generated masks. The SAM model is published in three variants of ViT-B, ViT-L and ViT-H. We use the ART-Net [8] and SurgAI3.8K [9] datasets. The first dataset contains surgical images with segmentation masks and the second one contains surgery images with annotated gynaecological organs contours which we used to generate uterus masks. For the methods with training for the tasks of surgical instrument segmentation, we use the ART-Net training set of 662 images and evaluate over the test set of 154 images. For the task of uterus segmentation we train over 3436 training images and test over 382 images. We investigate the following five methods.

2.2 Unsupervised Learning with DINO Backbones

We first used PCA to perform dimension reduction on the feature vectors of all patches of all training images, empirically keeping the 5 first principal components. We then used K-means to cluster the patches into two clusters. We selected the largest cluster as background and the other as instruments. This method works at the patch level, producing a coarse segmentation result.

2.3 Supervised Learning with Linear Decoder

We trained a linear decoder mapping the feature vector of a patch to the classification result. The linear decoder has several thousands of parameters and is trained using a binary cross-entropy loss using few shots. Similarly to the unsupervised learning method, this method works at the patch level, producing a coarse segmentation result.

2.4 Supervised Learning with DINO-UNet

We propose the DINO-UNet architecture shown in figure 1. It is a modified UNet architecture using a foundation model as second encoder. The first encoder is the usual UNet encoder, which we chose as a ResNet34 with skip connections. The features from the two encoders are concatenated to form the input to the first standard UNet decoder. The UNet part of the architecture is trained with a binary cross-entropy loss. In contrast to the above two methods, DINO-UNet works at the pixel level, producing a fine segmentation result.

2.5 Supervised Learning with DPT

The dense prediction transformer (DPT) [10] can be used with ViT backbones. It assembles tokens from various stages of the vision transformer into image-like representations at various resolutions and progressively combine them into full-resolution predictions using a convolutional decoder. We adapt the DPT to use it with different DINO backbones. We kept the backbone parameters frozen and only updated the parameters in the DPT head during downstream task training.

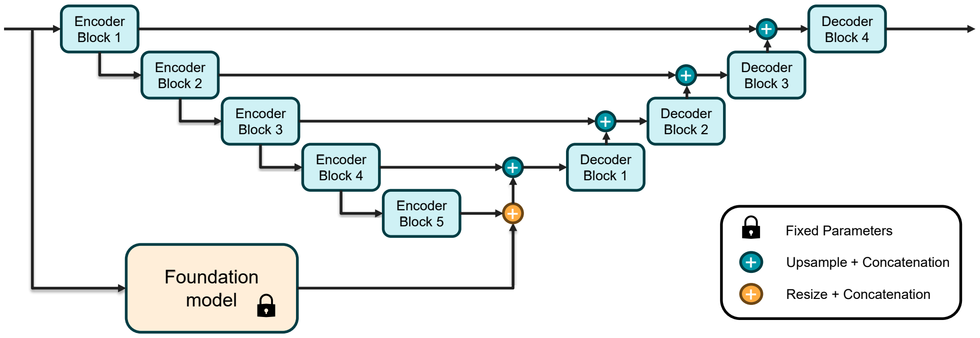

Fig. 1 Proposed DINO-UNet adaptation architecture, combining a UNet with a foundation model.

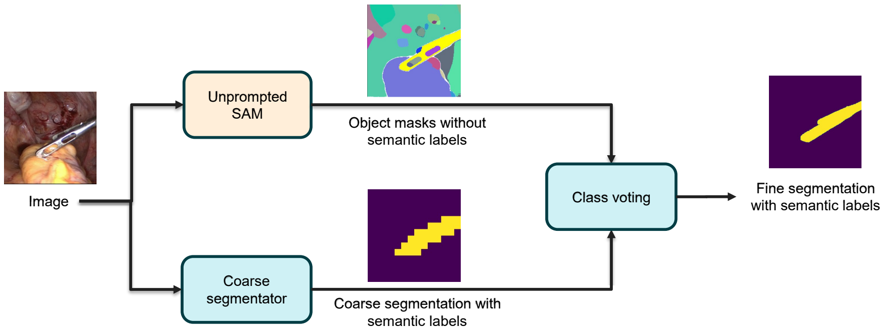
Image
Unprompted SAM
Object masks without semantic labels
Coarse segmentator
Coarse segmentation with semantic labels
Class voting
Fine segmentation with semantic labels

Fig. 2 Proposed adaptation architecture to assign class labels to SAM object masks using a coarse segmentator.

2.6 Unsupervised SAM Training with Semantic Labels

We use the SAM model in unprompted mode, in which the model gives the masks for all of the objects it detects in the image but without semantic labels. As shown in figure 2, we propose to use a coarse segmentator to resolve the semantic labels. The coarse segmentator can be either an unsupervised segmentator or a simple segmentator trained supervisedly to give coarse segmentation. We use the above-described unsupervised coarse segmentator. To determine the fine segmentation, we assign the most frequent label from the coarse segmentation pixels inside the object to all pixels of that object.

3 Experimental Results

We report results for the tasks of instrument then uterus segmentation. We use mean intersection-over-union (mIoU) as evaluation metric.

3.1 Instrument Segmentation

For the task of surgical instrument segmentation, we evaluated 17 models formed by combining the 5 methods of section 2 and the 7 selected foundation models. We also trained the baseline UNet network with ResNet34 encoders, which serves as an ablation of the foundation model from the proposed DINO-UNet architecture. ART-Net represents the state-of-the-art. We observe in table 1 that unsupervised learning with DINO and SAM and supervised learning with linear decoding are subpar UNet and ART-Net, yet give reasonably good performance. They could be useful for cases with an absence or a strong shortage of labelled data. The unsupervised learning with SAM shows outstanding results for an unsupervised method, especially with the ViT-H backbone. We observe that DINO-UNet and DPT substantially outperform both UNet and ART-Net. While all the four tested DINO backbones show good performance, DINOv2 with registers is overall the one with the best performance.

Table 1 Evaluation of instrument segmentation, expressed as mIoU in %.

	Foundation model		Num. of parameters		
Adaptation	Model	Variant	Fixed	Trainable	mIoU (%)
ART-Net (mIoU from [8])	-	-	0	17M	88.2
UNet	-	-	0	38M	86.5
Unsupervised with DINO	DINOv1	ViT-B/8	85M	0	74.7
	DINOv1	ViT-B/16	85M	0	64.9
	DINOv2	ViT-L/14	300M	0	68.9
	DINOv2+reg	ViT-L/14	300M	0	72.4
DINO+Linear decoder	DINOv1	ViT-B/8	85M	3K	79.4
	DINOv1	ViT-B/16	85M	3K	69.2
	DINOv2	ViT-L/14	300M	4K	82.0
	DINOv2+reg	ViT-L/14	300M	4K	82.6
DINO-UNet	DINOv1	ViT-B/8	85M	38M	86.7
	DINOv1	ViT-B/16	85M	38M	86.7
	DINOv2	ViT-L/14	300M	38M	87.3
	DINOv2+reg	ViT-L/14	300M	38M	<u>92.3</u>
DPT	DINOv2	ViT-L/14	300M	40M	90.6
	DINOv2+reg	ViT-L/14	300M	40M	**93.8**
Unsupervised with SAM	SAM	ViT-B	93M	0	52.3
	SAM	ViT-L	312M	0	76.0
	SAM	ViT-H	641M	0	80.6

As the number of trainable parameters in models with DINO encoder and linear decoder are limited to a few thousands, we expect few shot learning to perform well. We therefore conducted experiments of training with 1, 5 and 50 training samples along with the full 662 training samples. Figure 3 shows the performance of the models in these experiments, which shows that the performance downgrade for few-shot learning is almost negligible.

We show in figure 4 a qualitative comparison between the methods and ground-truth, as well as the first three PCA components of the DINO features using the red, green and blue channels. We observe that all four DINO backbones produce discriminative features and that DINOv2 with registers has the features with the highest quality. Comparing the feature maps in figure 4 confirms the presence of artefacts in low-informative regions when the registers are not used. Hence, registers are expected

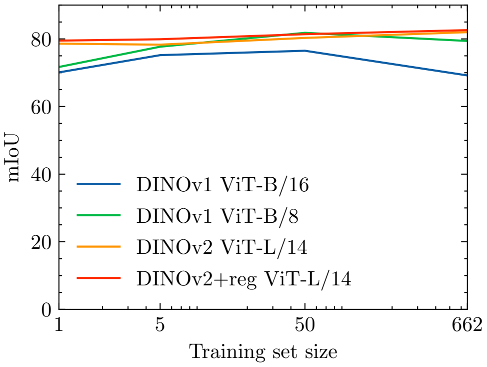

Fig. 3 Performance of models with DINO encoder and linear decoder trained on few shots in instrument segmentation.

to improve performance in all adaptations. Although unsupervised learning and supervised learning with linear decoder are loyal to semantics, they, as expected, do not have a precise spatial resolution and do not segment the objects with precise boundaries. In contrast, DINO-UNet, DPT and unsupervised learning using SAM are faithful both to semantics and spatial precision, segmenting small details such as the fine structures in the instrument heads, which are particularly challenging.

3.2 Uterus Segmentation

For the task of uterus segmentation, we assessed the performance of 7 models generated by combining 3 methods with 5 selected foundation models. We also trained a baseline UNet with ResNet34 encoders to serve as a ablation of the foundation model within the proposed DINO-UNet architecture. Our analysis, following the results given in table 2, is that unsupervised learning with SAM models underperforms compared to UNet. We observe consistent superior performance from DINO-UNet and DPT compared to UNet. DINOv2 with registers is, again, the foundation model with the best performance.

Table 2 Evaluation of uterus segmentation, expressed as mIoU in %.

	Foundation model		Num. of parameters		
Adaptation	Model	Variant	Fixed	Trainable	mIoU (%)
UNet	-	-	0	38M	84.9
DINO-UNet	DINOv2	ViT-L/14	300M	38M	85.5
	DINOv2+reg	ViT-L/14	300M	38M	<u>86.4</u>
DPT	DINOv2	ViT-L/14	300M	40M	86.0
	DINOv2+reg	ViT-L/14	300M	40M	**89.3**
Unsupervised with SAM	SAM	ViT-B	93M	0	41.3
	SAM	ViT-L	312M	0	70.5
	SAM	ViT-H	641M	0	71.7

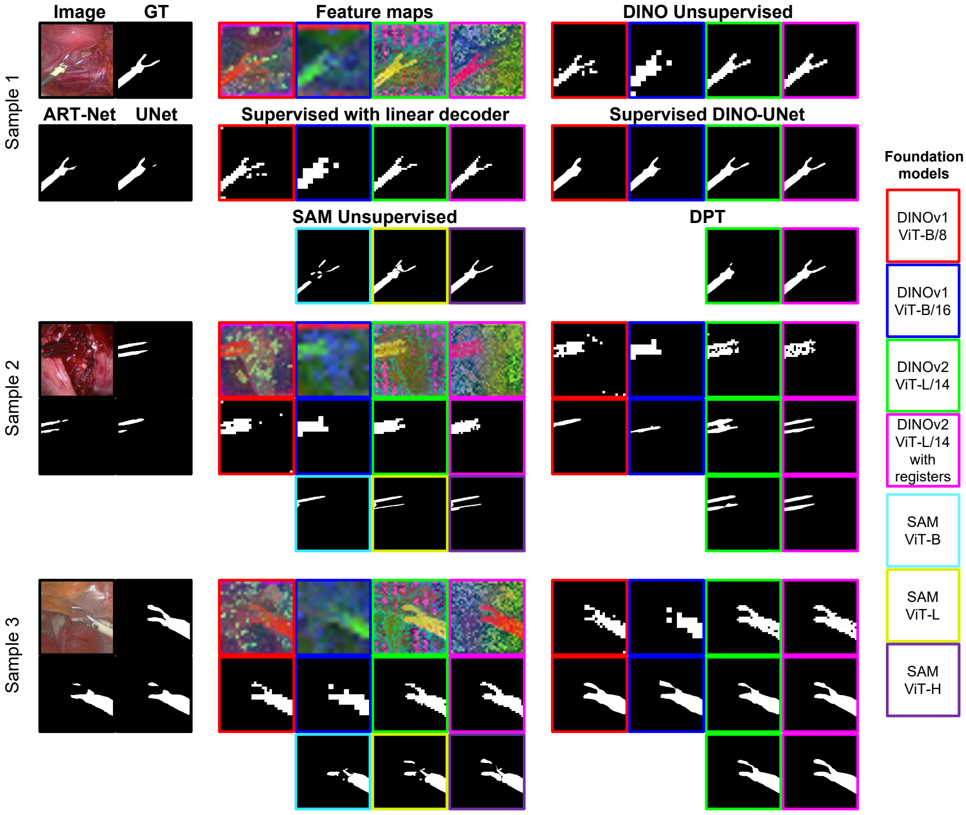

Fig. 4 Sample images from the ART-Net test set, with ground-truth (GT) segmentation, feature maps and segmentation outputs for different models.

A qualitative comparison between the methods and ground-truth is shown in figure 5. DPT and DINO-UNet segment the uterus correctly and with precise outlines. DINOv2 with registers outperforms DINOv2 without registers. Unsupervised learning with the SAM foundation model has an inconsistent performance over all the images.

4 Conclusion

The segmentation results obtained from both unsupervised and supervised training with a linear decoder, even if coarse, are accurate enough for some applications such as segmentation, as their mIoU is comparable to the UNet's, while not needing or needing much fewer training data. The unsupervised method with the SAM foundation model shows good performance in instrument segmentation but not in uterus segmentation. The DPT and proposed DINO-UNet architectures substantially outperform the previous-best by 5.6 and 4.1 pp (percentage points) over the ART-Net dataset and 4.4 and 1.5 pp over the SurgAI3.8K dataset. They both effectively segment instruments and the uterus, even in the most challenging conditions, maintaining nearly perfect boundaries and capturing complex details. The analysis of the complete experiment

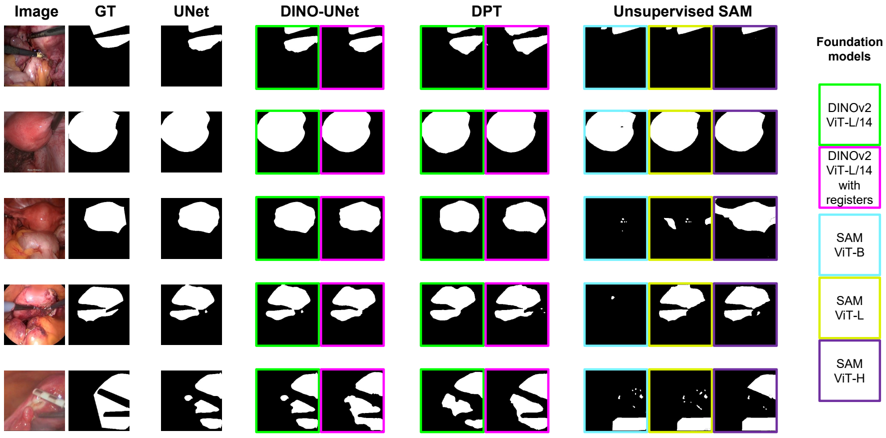

Fig. 5 Sample images from the SurgAI3.8K test set, with ground-truth (GT) segmentation and segmentation outputs for different models.

set suggests that visual foundation models, even trained on general images, could play an important role in surgical computer vision in unsupervised, supervised and few-shot learning, as medical data is often scarce and complex to collect. The proposed DINO-UNet could represent an initial method to exploit these generic features for cases where some annotated data are available. The proposed architecture combining SAM with coarse segmentation could represent an initial method for cases without labels. Obtaining a definitive answer will however require additional investigations. We plan to provide more solid evidence in future work by experimenting with more foundation models and downstream tasks, and fine-tuning the DINO backbones over a wide dataset of surgical images to address the possibility of training a surgical foundation model. The proposed framework could serve as a benchmark to continuously test the new foundation models from computer vision or medical image analysis such as [11] in the surgical context.

Declarations

The authors declare that they have no conflict of interest. All procedures involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors. Informed consent was obtained from the patients included in the study.

References

[1] Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)

[2] Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., et al.: DINOv2: Learning robust visual features without supervision. arXiv:2304.07193 (2023)

[3] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al.: Segment anything. ICCV (2023)

[4] Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. In: NeurIPS (2023)

[5] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., et al.: Learning transferable visual models from natural language supervision. ICML (2021)

[6] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. ICCV (2021)

[7] Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv:2309.16588 (2023)

[8] Hasan, M.K., Calvet, L., Rabbani, N., Bartoli, A.: Detection, segmentation, and 3D pose estimation of surgical tools using convolutional neural networks and algebraic geometry. Medical Image Analysis (2021)

[9] Zadeh, S.M., François, T., Comptour, A., Canis, M., Bourdel, N., Bartoli, A.: Surgai3. 8k: A labeled dataset of gynecologic organs in laparoscopy with application to automatic augmented reality surgical guidance. Journal of Minimally Invasive Gynecology **30**(5), 397–405 (2023)

[10] Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ArXiv preprint (2021)

[11] Ramesh, S., Srivastav, V., Alapatt, D., Yu, T., Murali, A., Sestini, L., Nwoye, C.I., Hamoud, I., Sharma, S., Fleurentin, A., *et al.*: Dissecting self-supervised learning methods for surgical computer vision. Medical Image Analysis **88**, 102844 (2023)