Towards Effective Surgical Representation Learning with DINO Models

Ronald L.P.D. de Jong[1] [ID]   R.L.P.D.D.JONG@TUE.NL
Yiping Li[1] [ID]   Y.LI9@TUE.NL
Tim J.M. Jaspers[2] [ID]   T.J.M.JASPERS@TUE.NL
Romy C. van Jaarsveld[3] [ID]   R.C.VANJAARSVELD-3@UMCUTRECHT.NL
Gino M. Kuiper[3] [ID]   G.M.KUIPER-2@UMCUTRECHT.NL
Franco Badaloni[3] [ID]   F.BADALONI@UMCUTRECHT.NL
Richard van Hillegersberg[3] [ID]   R.VANHILLEGERSBERG@UMCUTRECHT.NL
Jelle P. Ruurda[3] [ID]   J.P.RUURDA@UMCUTRECHT.NL
Fons van der Sommen[2] [ID]   FVDSOMMEN@TUE.NL
Josien P.W. Pluim[1] [ID]   J.PLUIM@TUE.NL
Marcel Breeuwer[1,2] [ID]   M.BREEUWER@TUE.NL

[1] *Department of Biomedical Engineering, Eindhoven University of Technology, Groene Loper 3, Eindhoven, 5612 AE, Noord-Brabant, The Netherlands.*

[2] *Department of Electrical Engineering, Eindhoven University of Technology, Groene Loper 3, Eindhoven, 5612 AE, Noord-Brabant, The Netherlands.*

[3] *Department of Surgery, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, 3584 CX, Utrecht, The Netherlands.*

Abstract

Self-supervised learning (SSL) has emerged as a promising approach to address the limitations of annotated surgical datasets, which are often small, heterogeneous, and expensive to curate. Among SSL methods, self-distillation with no labels (DINO) has achieved state-of-the-art (SOTA) results in natural images, but its applicability to surgical data remains underexplored. In this work, we systematically investigate DINOv1, DINOv2, and DINOv3 for surgical representation learning. We pretrain these models on a large-scale surgical dataset of 4.7M video frames (SurgeNetXL) and evaluate their transferability on downstream tasks including semantic segmentation and surgical phase recognition. Our results demonstrate that in-domain pretraining consistently improves performance across all DINO variants, with DINOv2 and DINOv3 achieving SOTA performance. We further offer practical insights and visualizations highlighting the effectiveness of SSL. Finally, our study delivers ready-to-use DINO-based SSL models and pretraining protocols for surgical computer vision research, which are publicly available at: `github.com/rlpddejong/SurgeNetDINO`.

Keywords: Anatomy recognition, DINO, Self-supervised learning, Surgical phase recognition

1. Introduction

The success of deep learning in surgical data science has largely relied on supervised models trained on curated datasets. However, collecting and annotating surgical data is challenging: annotations require expert knowledge, surgical procedures are highly variable across centers,

and available datasets are often small compared to natural image benchmarks. These constraints limit the scalability of purely supervised approaches and motivate the use of self-supervised learning (SSL), where models are first pretrained on large amounts of unlabeled surgical data and then fine-tuned in a supervised manner on diverse downstream surgical tasks.

Self-distillation with no labels (DINO) has emerged as a leading SSL method in natural images, with successive versions (DINOv1, DINOv2, DINOv3) producing increasingly robust visual representations (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025). Despite their success in general computer vision, the utility of these methods for surgical applications remains largely unexplored. In particular, it is unclear how the different DINO variants compare to each other on surgical data science tasks, to what extent surgical pretraining improves their performance on downstream tasks, and which training strategies are most effective in this domain.

In this work, we investigate self-supervised pretraining of DINOv1, v2, and v3 on a large-scale surgical dataset of 4.7M frames, followed by supervised fine-tuning on a variety of downstream surgical computer vision tasks. This two-stage approach allows us to evaluate the transferability of pretrained representations and identify the design choices that maximize performance on each task. Beyond reporting comparative results, we examine key training parameters to offer practical recommendations for adapting SSL to surgical data. Our contributions can be summarized as follows:

- We present the first systematic study of DINOv1, v2, and v3 pretrained on a large-scale surgical dataset consisting of 4.7M frames, and release the model weights at: github.com/rlpddejong/SurgeNetDINO.

- We provide comparative insights into the transferability and strengths of each DINO variant for various downstream surgical computer vision applications, achieving state-of-the-art (SOTA) performance that surpasses prior methods.

- We perform analyses on architecture size, training duration, and fine-tuning strategies, providing practical recommendations for designing effective SSL pipelines in the surgical domain.

2. Related work

2.1. Self-supervised learning (SSL)

SSL has emerged as a powerful paradigm for learning useful representations from unlabeled data, reducing reliance on costly manual annotations. Early approaches in SSL relied on pretext tasks such as image inpainting, colorization, or predicting rotations (Pathak et al., 2016; Larsson et al., 2016; Gidaris et al., 2018). More recent contrastive learning methods, such as SimCLR (Chen et al., 2020) and MoCo (He et al., 2020), learn representations by encouraging similarity between augmented views of the same image while pushing apart representations of different images. These methods have shown remarkable success on natural image benchmarks.

2.2. Self-distillation with no labels (DINO)

DINO is an SSL method that avoids explicit contrastive objectives by leveraging knowledge distillation between student and teacher networks (Caron et al., 2021). Subsequent versions, DINOv2 and DINOv3, have improved representation quality and generalization by employing stronger training protocols, larger datasets, and better architectural choices (Oquab et al., 2024; Siméoni et al., 2025). DINO has demonstrated SOTA performance in a variety of downstream tasks in natural image domains, including classification, detection, and segmentation. However, most evaluations of these methods, especially the latest versions, focus primarily on general computer vision datasets, such as ImageNet.

2.3. SSL in surgical computer vision

Recent work has examined the potential of DINO in the surgical computer vision domain. Ramesh et al. (2023) systematically analyzed self-supervised methods, including DINO, for surgical computer vision tasks and highlighted their ability to improve representation learning. Cui et al. (2024) proposed Surgical-DINO, adapting foundation models with adapters for depth estimation in endoscopic surgery. More recently, Jaspers et al. (2026) showed that scaling self-supervised pretraining on large surgical datasets can substantially enhance the performance of foundation models for a variety of surgical computer vision tasks. Additionally, the LEMON dataset (Che et al., 2025) highlights the growing importance of large-scale data curation efforts in this field, with a particular focus on organizing and refining extensive unlabeled endoscopic video collections for foundation model training. Despite these advances, important gaps remain: DINOv3 has not yet been explored in surgical applications, and no direct comparisons between different DINO variants have been performed. These limitations leave open questions about the most effective pretraining strategies and model choices for this specialized domain.

3. Methods

3.1. Pretraining datasets

The original DINO models were pretrained on large-scale natural image datasets, such as ImageNet (Russakovsky et al., 2014), LVD-142M (Oquab et al., 2024), and LVD-1689M (Siméoni et al., 2025), which provide strong visual priors for general computer vision tasks. However, these datasets are limited in their relevance to surgical domains, as they lack the anatomical structures, tools, and visual characteristics present in operative scenes.

For self-supervised pretraining, we used the SurgeNetXL dataset introduced by Jaspers et al. (2026). SurgeNetXL is a large-scale collection of 4.7M surgical video frames sampled at 1 fps from a combination of public surgical datasets, curated YouTube surgical footage, and two institutional datasets. All videos were anonymized where required, and quality-filtered to exclude non-surgical or irrelevant content. By integrating diverse procedures and video sources, SurgeNetXL provides both scale and heterogeneity, making it well-suited for pretraining large vision models in surgical domains.

3.2. Pretraining implementation

We pretrained DINOv1, DINOv2, and DINOv3 following their original frameworks on the small, base, and large variants of the Vision Transformer (ViT-s, ViT-b, and ViT-l, respectively), models originally proposed proposed by Dosovitskiy et al. (2021). Notably, DINOv1 was not pretrained on ViT-l, as this variant was not included in the original implementation. The trained model weights, along with the exact pretraining details, are publicly available at github.com/rlpddejong/SurgeNetDINO. Pretraining was initialized from ImageNet, LVD-142M, and LVD-1689M weights for DINOv1, DINOv2, and DINOv3, respectively. All pretraining experiments were carried out on four NVIDIA H100 GPUs (NVIDIA Corp., CA, USA) using the largest feasible batch size for each model during 120 hours, with a maximum of 50 epochs.

3.3. Fine-tuning datasets

In this study, we evaluated our pretrained models on four downstream datasets covering two key tasks: semantic segmentation and surgical phase recognition. These datasets include a range of surgical procedures such as laparoscopic cholecystectomy, robot-assisted minimally invasive esophagectomy (RAMIE), and hysterectomy. For semantic segmentation, we use the CholecSeg8k dataset (Hong et al., 2020), and the RAMIE-seg dataset (de Jong et al., 2025), which contain pixel-level annotations of anatomical structures and surgical tools. These datasets represent both laparoscopic and robotic procedures, with varying numbers of patients, frames, and classes. Surgical phase recognition was performed on the Auto-Laparo dataset (Wang et al., 2022) and an extended version of the RAMIE-phase dataset proposed by Li et al. (2025). These datasets consist of full-length surgical recordings annotated with temporal phases, including seven phases for hysterectomy and thirteen phases for esophagectomy. The combination of these datasets offers a diverse and representative benchmark for evaluating the effectiveness of self-supervised pretraining in multiple clinically relevant downstream tasks, supporting potential applications such as real-time surgical guidance, workflow monitoring, and postoperative analysis.

3.4. Fine-tuning implementation

3.4.1. Semantic segmentation

For semantic segmentation, we adopt the recently introduced Encoder-only Mask Transformer (EoMT) (Kerssies et al., 2025). EoMT uses a ViT backbone with learnable object queries and a lightweight mask–class head, avoiding the need for task-specific decoders. We use EoMT because it enables efficient fine-tuning on small surgical datasets, reduces architectural complexity, and aligns well with our goal of evaluating the quality of DINO-based representations without confounding effects from heavy decoder designs.

For CholecSeg8k, we used 6,800 frames for training and 1,280 frames for testing. Consistent with previous studies (Grammatikopoulou et al., 2024; Zhang et al., 2024), we excluded low-prevalence classes (blood, cystic duct, hepatic vein, and liver ligament) to ensure a robust evaluation. For RAMIE, 749 frames were allocated for training and 120 frames for testing. All splits were performed on a per-patient basis.

4

The model was trained using five-fold cross-validation at the patient level, with approximately 80% of the data used for training and 20% for validation. All frames were resized to match the resolution used during pretraining, using bicubic interpolation. We used the loss function described by Kerssies et al. (2025) and the AdamW optimizer with a learning rate of 1×10^{-5}, which was reduced by half after 10 epochs without an improvement in validation loss. All models were trained on a single NVIDIA H100 GPU (NVIDIA Corp., CA, USA) with a batch size of 8 and an early-stopping criterion of 15 epochs. Data augmentation was limited to horizontal and vertical flips and rotations, each applied with a 50% probability.

3.4.2. Phase recognition model

For fine-tuning on the phase recognition datasets, we adopted a two-stage training procedure following Jaspers et al. (2026). In the first stage, the backbone model was trained to predict surgical phases from individual frames. In the second stage, a causal Multi-Stage Temporal Convolutional Network (MS-TCN) was trained on the extracted features.

For the AutoLaparo dataset (Wang et al., 2022), we followed the original train, validation, and test splits. For the extended RAMIE-phase dataset, we used 18 videos for training, 7 for validation, and 16 for testing. Frames were extracted at 1 fps and resized using bicubic interpolation to match the pre-training resolution of each backbone.

Both training stages used the AdamW optimizer and cross-entropy loss. In the first stage, the backbone was trained on individual frames with a learning rate of 1×10^{-5}; the batch size was 32 for ViT-l and 64 for ViT-s and ViT-b. In the second stage, the causal MS-TCN was trained on full-video feature sequences with a batch size of 1, a learning rate of 7×10^{-4}, and 200 epochs. Data augmentation was applied only in the first stage and included random scaling and rotation, RGBShift, and RandomBrightnessContrast. All models were trained on a single NVIDIA A100 GPU (NVIDIA Corp., CA, USA).

3.5. Evaluation

For semantic segmentation, model performance was assessed using the Dice score, which measures the overlap between predicted and annotated masks, and the 95th percentile Hausdorff distance (HD95), which quantifies boundary alignment. Metrics were first averaged across classes for each patient, and then across patients to obtain overall scores. For surgical phase recognition, we report both frame-wise accuracy and macro-averaged F1 score, meaning that the F1 score is computed per phase and then averaged across phases. Reported results are expressed as mean ± standard deviation.

To benchmark our models against prior literature, we also fine-tuned the following SSL-based SOTA models: CAFormer (Jaspers et al., 2026), EndoViT (Batić et al., 2023), GSViT (Schmidgall et al., 2024), and EndoFM (Wang et al., 2023). To ensure a fair comparison, all SOTA models were fine-tuned using the same training configuration. For GSViT, we implemented an FPN decoder, as the original model did not include a decoder for semantic segmentation. We use a bootstrapping approach (Wiesenfarth et al., 2021; Varoquaux and Cheplygina, 2022) to evaluate the stability of our models with the SOTA models. The predictions are pooled to generate 1,000 bootstrap samples with replacement, maintaining a consistent random seed across experiments. This procedure accounts for sampling variability, which is especially important for small datasets or limited patient numbers.

4. Results

Table 1 presents the performance of DINOv1–v3 variants with various ViT architectures and pretraining strategies on the CholecSeg8k and RAMIE-seg datasets. In general, SurgeNetXL in-domain pretraining improves performance across all DINO variants compared to ImageNet or LVD, with an average improvement of 9.1%. It is worth noting that DINOv2 and DINOv3 models pretrained on the LVD datasets achieve performance comparable to DINOv1 models pretrained on SurgeNetXL, highlighting the advancements introduced in the newer versions of DINO. Larger backbones in DINOv2 and DINOv3 (ViT-b and ViT-l) further enhance segmentation performance. Specifically, DINOv2 with a ViT-l backbone reaches Dice scores of up to 0.77 and 0.79 on CholecSeg8k and RAMIE-seg, respectively, with substantially reduced HD95. DINOv3 reinforces these trends, with the ViT-l model achieving the highest Dice scores (0.78 and 0.74) and the lowest HD95 (29 and 46) compared to ViT-s and ViT-b. These results demonstrate that both DINOv2 and DINOv3 are strong segmentation models, particularly when paired with larger backbones and domain-specific pretraining.

Table 1: Performance comparison of DINOv1–v3 variants with different ViT architectures and pretraining strategies on the CholecSeg8k and RAMIE-seg datasets.

Model		Pretraining	CholecSeg8k dataset		RAMIE-seg dataset	
			Dice score ↑	HD95 ↓	Dice score ↑	HD95 ↓
v1	ViT-s	ImageNet	0.66 ± 0.09	65 ± 25	0.57 ± 0.04	73 ± 11
		SurgeNetXL	$\mathbf{0.73 \pm 0.09}$	$\mathbf{56 \pm 19}$	0.62 ± 0.03	63 ± 6
	ViT-b	ImageNet	0.68 ± 0.07	61 ± 23	0.61 ± 0.02	69 ± 9
		SurgeNetXL	$\mathbf{0.73 \pm 0.10}$	57 ± 19	$\mathbf{0.67 \pm 0.03}$	$\mathbf{58 \pm 14}$
v2	ViT-s	LVD-142M	0.70 ± 0.07	54 ± 19	0.60 ± 0.04	70 ± 7
		SurgeNetXL	0.66 ± 0.06	55 ± 18	0.59 ± 0.03	67 ± 5
	ViT-b	LVD-142M	0.71 ± 0.06	50 ± 18	0.67 ± 0.05	52 ± 6
		SurgeNetXL	0.75 ± 0.10	44 ± 22	0.73 ± 0.04	46 ± 11
	ViT-l	LVD-142M	0.63 ± 0.09	58 ± 13	0.70 ± 0.05	49 ± 6
		SurgeNetXL	$\mathbf{0.77 \pm 0.09}$	$\mathbf{38 \pm 20}$	$\mathbf{0.79 \pm 0.04}$	$\mathbf{40 \pm 7}$
v3	ViT-s	LVD-1689M	0.73 ± 0.06	52 ± 21	0.63 ± 0.06	59 ± 6
		SurgeNetXL	0.74 ± 0.07	48 ± 17	0.69 ± 0.05	54 ± 8
	ViT-b	LVD-1689M	0.71 ± 0.07	55 ± 24	0.67 ± 0.04	53 ± 10
		SurgeNetXL	0.75 ± 0.09	42 ± 22	0.73 ± 0.05	55 ± 14
	ViT-l	LVD-1689M	0.76 ± 0.09	34 ± 21	0.70 ± 0.04	$\mathbf{46 \pm 3}$
		SurgeNetXL	$\mathbf{0.78 \pm 0.11}$	$\mathbf{29 \pm 15}$	$\mathbf{0.74 \pm 0.04}$	46 ± 12

Table 2 presents the performance of DINOv1–v3 variants with various ViT architectures and pretraining strategies on the phase recognition datasets. Similarly to segmentation, in-domain pretraining generally improves performance over general pretraining, with a few exceptions for certain model variants. Interestingly, for phase recognition, the benefit of

larger architectures is less pronounced, with DINOv1 even performing slightly better on ViT-s. Furthermore, although DINOv2 and DINOv3 achieve the highest scores, the differences between versions are smaller compared to segmentation.

Figure 1 illustrates the ranking stability of our best models in comparison with top-performing SOTA models reported in the literature. From the figure, it is evident that the majority of our ViT-l models achieve higher performance rankings than the SOTA baselines. Notably, the models that were pretrained on SurgeNetXL demonstrate the best overall performance, indicating the effectiveness of this pretraining strategy in enhancing model robustness and stability across different evaluation scenarios.

Table 2: Performance comparison of DINOv1–v3 variants with different ViT architectures and pretraining strategies on the Autolaparo and RAMIE-phase datasets.

Model		Pretraining	Autolaparo dataset		RAMIE-phase dataset	
			Accuracy ↑	F1 score ↑	Accuracy ↑	F1 score ↑
v1	ViT-s	ImageNet	81.8 ± 9.4	70.4 ± 6.8	74.1 ± 9.3	65.3 ± 10.0
		SurgeNetXL	$\mathbf{85.3 \pm 5.6}$	$\mathbf{75.1 \pm 5.2}$	$\mathbf{77.4 \pm 9.8}$	$\mathbf{70.5 \pm 11.8}$
	ViT-b	ImageNet	81.1 ± 9.0	68.7 ± 8.4	75.4 ± 10.0	66.4 ± 10.9
		SurgeNetXL	85.0 ± 9.1	70.6 ± 7.6	77.2 ± 10.6	70.1 ± 11.5
v2	ViT-s	LVD-142M	83.0 ± 9.9	73.0 ± 11.0	76.6 ± 9.5	67.9 ± 9.9
		SurgeNetXL	84.5 ± 8.6	69.6 ± 6.4	76.0 ± 11.0	66.3 ± 12.2
	ViT-b	LVD-142M	85.3 ± 7.4	74.2 ± 8.2	77.4 ± 9.9	67.8 ± 10.4
		SurgeNetXL	85.9 ± 8.3	74.0 ± 5.9	78.2 ± 10.2	71.1 ± 11.8
	ViT-l	LVD-142M	$\mathbf{86.1 \pm 7.6}$	$\mathbf{77.8 \pm 8.9}$	77.9 ± 10.3	70.2 ± 11.4
		SurgeNetXL	84.1 ± 8.4	70.5 ± 8.0	$\mathbf{80.8 \pm 9.1}$	$\mathbf{73.7 \pm 11.5}$
v3	ViT-s	LVD-1689M	81.1 ± 10.0	69.7 ± 8.6	75.4 ± 10.8	63.3 ± 11.8
		SurgeNetXL	82.2 ± 8.5	72.1 ± 8.0	74.4 ± 10.9	62.9 ± 11.7
	ViT-b	LVD-1689M	83.4 ± 9.4	70.8 ± 10.1	75.8 ± 10.7	66.8 ± 10.5
		SurgeNetXL	83.3 ± 8.5	73.0 ± 4.7	76.3 ± 10.5	66.7 ± 12.1
	ViT-l	LVD-1689M	86.0 ± 8.6	75.0 ± 9.6	77.8 ± 9.2	69.8 ± 10.3
		SurgeNetXL	$\mathbf{86.4 \pm 6.6}$	$\mathbf{78.5 \pm 6.7}$	$\mathbf{78.0 \pm 10.5}$	$\mathbf{70.1 \pm 11.8}$

Figure 2 presents a principal component analysis (PCA) of features extracted by the unsupervised DINOv2 ViT-l model. For visualization, the three principal components were mapped to the RGB color channels. To better resemble the similarity of the PCA to the reference annotations, the colors of the annotations have been manually matched retrospectively. The PCA reveals that models pretrained on SurgeNetXL better capture the semantic meaning of the surgical images: most annotated structures are represented as distinct, coherent colors, which is remarkable given that the model was trained entirely with self-supervision. In contrast, the LVD-1689M features produce noisier results, with most structures less clearly separated.
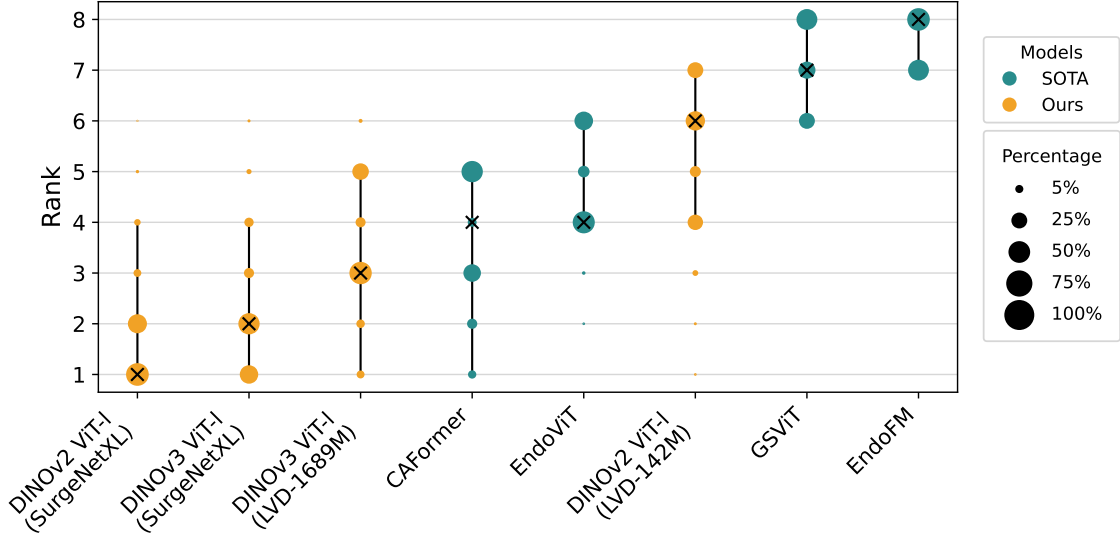
Figure 1: Ranking stability of our models vs. SOTA models: CAFormer (Jaspers et al., 2026), EndoViT (Batić et al., 2023), GSViT (Schmidgall et al., 2024), and EndoFM (Wang et al., 2023). Results are based on the Dice score and HD95 for the segmentation datasets. Blob size indicates how often an architecture attains a given rank. Black crosses mark the median ranks rounded to nearest integer, and black lines show 95% bootstrap intervals (2.5–97.5th percentiles). Models are ordered left to right by mean bootstrap rank, best to worst.
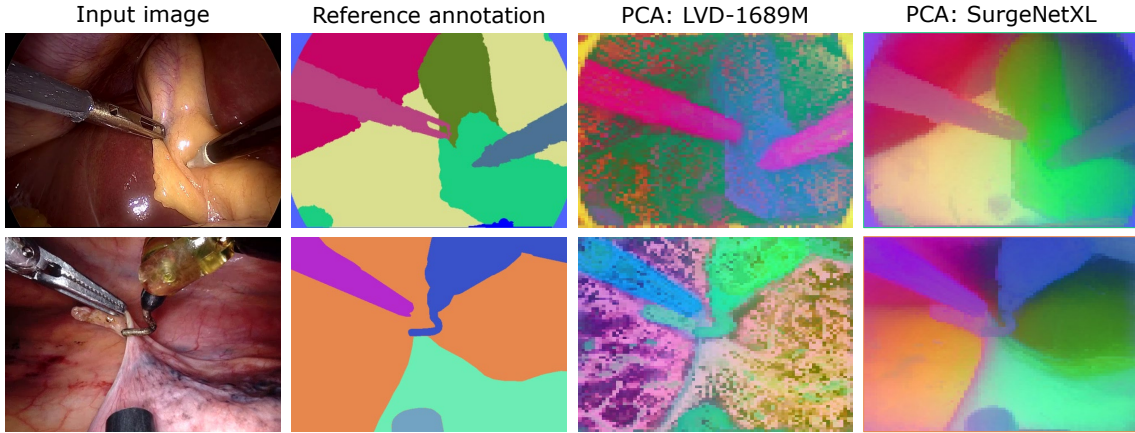

Figure 2: PCA of unsupervised DINOv2 ViT-l features on CholecSeg8k (top row) and RAMIE-seg (bottom row). Columns show: input image, reference annotation, PCA of LVD-1689M features, and PCA of SurgeNetXL features.

The left plot in Figure 3 shows the Dice score improvements for each of the DINO variants over training time. All models improve as training progresses, highlighting the importance of pretraining duration. Interestingly, DINOv3 exhibits slightly lower final score improvements, which may be due to its stronger pretrained features compared to the other variants, reducing the apparent gain from additional training. A similar, albeit smaller, effect is also observed for DINOv1 and DINOv2, which could indicate slight overfitting.

The right plot in Figure 3 shows the Dice scores for DINO models pretrained on ImageNet, LVD variants, and SurgeNetXL, comparing frozen and trainable encoders during fine-tuning. While median scores are slightly higher with a trainable encoder, the differences are minimal, and using a frozen encoder significantly reduces training complexity due to fewer trainable parameters. This also demonstrates the high quality of DINO features, for which extensive fine-tuning is not necessary.
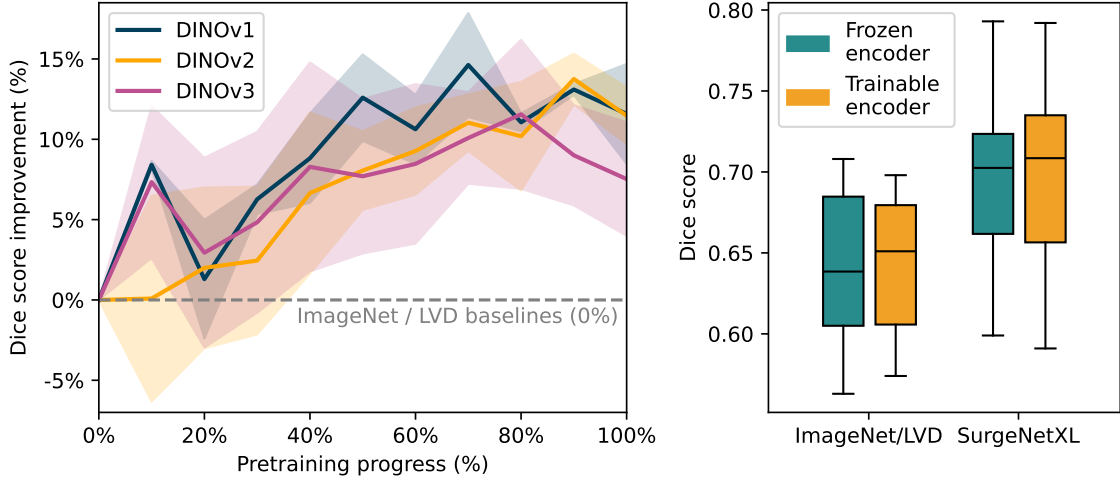

Figure 3: Left: Dice score improvements of DINOv1–v3 relative to the ImageNet and LVD baselines over training time. Lines represent the mean, and shaded areas indicate the standard deviation. Right: Boxplots showing the Dice scores for all models based on a frozen or trainable encoder during fine-tuning.

5. Discussion

Our experiments provide several insights into the application of DINO models for surgical representation learning. First, in-domain pretraining on SurgeNetXL consistently improves performance across both semantic segmentation and surgical phase recognition tasks. This highlights the importance of domain-specific visual priors, which better capture the unique appearance of surgical scenes.

The improvements observed for segmentation and phase recognition differ in magnitude and behavior. Segmentation tasks benefit substantially from larger ViT backbones (ViT-b and ViT-l), as they require fine-grained spatial understanding and precise delineation

of anatomical structures and surgical instruments. In contrast, phase recognition shows smaller gains from increased model capacity; even smaller ViT-s models achieve competitive accuracy. This difference likely reflects the inherent nature of the tasks: segmentation demands high-resolution spatial reasoning, whereas phase recognition relies primarily on temporal patterns and higher-level contextual cues, which smaller architectures seem to capture sufficiently.

Visualization of unsupervised features via PCA further supports our findings. Models pretrained on SurgeNetXL produce coherent feature clusters corresponding to anatomical structures, whereas models pretrained on natural images alone produce noisier and less semantically meaningful representations. This suggests that DINO can capture relevant surgical semantics without supervision, provided the pretraining data is representative of the domain.

Generally, DINOv2 and DINOv3 outperform DINOv1, with only minor differences between DINOv2 and DINOv3. One possible explanation is that for DINOv3 we only applied high-resolution fine-tuning. Further improvements could be achieved by training even larger ViT models, such as a ViT-7B, followed by knowledge distillation to smaller models like in the original implementation. However, it would require substantially more GPU resources to be effective.

Finally, our ranking stability analysis demonstrates that scaling SSL methods and model architecture leads to robust SOTA performance. This indicates that carefully designed self-supervised pipelines can produce generalizable surgical representations, suitable for a range of downstream applications, including real-time guidance, workflow analysis, and post-operative review.

A limitation of our study is that SurgeNetXL is considerably smaller than the LVD datasets used for general-domain DINO pretraining, suggesting that future work should explore substantially larger surgical video collections to further improve SSL performance.

In summary, our study highlights the effectiveness of SSL in surgical computer vision, emphasizing the roles of data domain, model size, and training parameters. These insights provide practical guidance for future work seeking to leverage large self-supervised models in surgical settings.

6. Conclusion

In this work, we conducted a systematic study of DINOv1–v3 models for surgical representation learning. Our results show that in-domain pretraining on surgical data enhances performance across semantic segmentation and phase recognition tasks. DINOv2 and DINOv3, particularly with larger architectures, achieve SOTA performance. These findings highlight the value of domain-specific SSL for surgical computer vision and offer practical guidance for future applications.

Acknowledgments

References

Dominik Batić, Felix Holm, Ege Özsoy, Tobias Czempiel, and Nassir Navab. Whether and when does endoscopy domain pretraining make sense?, 2023. URL https://arxiv.org/abs/2303.17636.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, October 2021.

Chengan Che, Chao Wang, Tom Vercauteren, Sophia Tsoka, and Luis C. Garcia-Peraza-Herrera. Lemon: A large endoscopic monocular dataset and foundation model for perception in surgical settings, 2025. URL https://arxiv.org/abs/2503.19740.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Beilei Cui, Mobarakol Islam, Long Bai, and Hongliang Ren. Surgical-dino: adapter learning of foundation models for depth estimation in endoscopic surgery. *International Journal of Computer Assisted Radiology and Surgery*, 19(6):1013–1020, 2024.

Ronald L. P. D. de Jong, Yasmina al Khalil, Tim J. M. Jaspers, Romy C. van Jaarsveld, Gino M. Kuiper, Yiping Li, Richard van Hillegersberg, Jelle P. Ruurda, Marcel Breeuwer, and Fons van der Sommen. Benchmarking pretrained attention-based models for real-time recognition in robot-assisted esophagectomy. In *Medical Imaging 2025: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 13408, page 1340810. International Society for Optics and Photonics, SPIE, 2025. doi: 10.1117/12.3045187. URL https://doi.org/10.1117/12.3045187.

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. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018. URL https://arxiv.org/abs/1803.07728.

Maria Grammatikopoulou, Ricardo Sanchez-Matilla, Felix Bragman, David Owen, Lucy Culshaw, Karen Kerr, Danail Stoyanov, and Imanol Luengo. A spatio-temporal network for video semantic segmentation in surgical videos. *International Journal of Computer Assisted Radiology and Surgery*, 19(2):375–382, 2024. doi: 10.1007/s11548-023-02971-6.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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, 2020. URL https://arxiv.org/abs/2012.12453.

Tim J.M. Jaspers, Ronald L.P.D. de Jong, Yiping Li, Carolus H.J. Kusters, Franciscus H.A. Bakker, Romy C. van Jaarsveld, Gino M. Kuiper, Richard van Hillegersberg, Jelle P. Ruurda, Willem M. Brinkman, Josien P.W. Pluim, Peter H.N. de With, Marcel Breeuwer, Yasmina Al Khalil, and Fons van der Sommen. Scaling up self-supervised learning for improved surgical foundation models. *Medical Image Analysis*, 108:103873, 2026. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103873. URL https://www.sciencedirect.com/science/article/pii/S1361841525004190.

Tommie Kerssies, Niccolò Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your ViT is Secretly an Image Segmentation Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 577–593, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

Yiping Li, Romy van Jaarsveld, Ronald de Jong, Jasper Bongers, Gino Kuiper, Richard van Hillegersberg, Jelle Ruurda, Marcel Breeuwer, and Yasmina Al Khalil. Benchmarking and enhancing surgical phase recognition models for robot-assisted esophagectomy. In *Medical Imaging 2025: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 13408, page 134080T. International Society for Optics and Photonics, SPIE, 2025. doi: 10.1117/12.3040084. URL https://doi.org/10.1117/12.3040084.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Sanat Ramesh, Vinkle Srivastav, Deepak Alapatt, Tong Yu, Aditya Murali, Luca Sestini, Chinedu Innocent Nwoye, Idris Hamoud, Saurav Sharma, Antoine Fleurentin, Georgios Exarchakis, Alexandros Karargyris, and Nicolas Padoy. Dissecting self-supervised learning methods for surgical computer vision. *Medical Image Analysis*, 88:102844, 2023. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102844. URL https://www.sciencedirect.com/science/article/pii/S1361841523001044.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 09 2014. doi: 10.1007/s11263-015-0816-y.

Samuel Schmidgall, Ji Woong Kim, Jeffrey Jopling, and Axel Krieger. General surgery vision transformer: A video pre-trained foundation model for general surgery, 2024. URL https://arxiv.org/abs/2403.05949.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine*, 5(1):48, 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00592-y.

Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 101–111, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43996-4.

Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 486–496, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16449-1.

Manuel Wiesenfarth, Annika Reinke, Bennett A. Landman, Matthias Eisenmann, Laura Aguilera Saiz, M. Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. Methods and open-source toolkit for analyzing and visualizing challenge results. *Scientific Reports*, 11(1):2369, 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-82017-6.

Luyang Zhang, Yuichiro Hayashi, Masahiro Oda, and Kensaku Mori. Towards better laparoscopic video segmentation: A class-wise contrastive learning approach with multi-scale feature extraction. *Healthcare Technology Letters*, 11(2-3):126–136, 2024. doi: 10.1049/htl2.12069.