Revisiting Surgical Instrument Segmentation Without Human Intervention: A Graph Partitioning View

Mingyu Sheng
The University of Sydney
Sydney, NSW, Australia
mshe0136@uni.sydney.edu.au

Jianan Fan
The University of Sydney
Sydney, NSW, Australia
jfan6480@uni.sydney.edu.au

Dongnan Liu
The University of Sydney
Sydney, NSW, Australia
dongnan.liu@sydney.edu.au

Ron Kikinis
Harvard Medical School
Boston, MA, USA
kikinis@bwh.harvard.edu

Weidong Cai
The University of Sydney
Sydney, NSW, Australia
tom.cai@sydney.edu.au

Abstract

Surgical instrument segmentation (SIS) on endoscopic images stands as a long-standing and essential task in the context of computer-assisted interventions for boosting minimally invasive surgery. Given the recent surge of deep learning methodologies and their data-hungry nature, training a neural predictive model based on massive expert-curated annotations has been dominating and served as an off-the-shelf approach in the field, which could, however, impose prohibitive burden to clinicians for preparing fine-grained pixel-wise labels corresponding to the collected surgical video frames. In this work, we propose an unsupervised method by reframing the video frame segmentation as a graph partitioning problem and regarding image pixels as graph nodes, which is significantly different from the previous efforts. A self-supervised pre-trained model is firstly leveraged as a feature extractor to capture high-level semantic features. Then, Laplacian matrixs are computed from the features and are eigendecomposed for graph partitioning. On the "deep" eigenvectors, a surgical video frame is meaningfully segmented into different modules such as tools and tissues, providing distinguishable semantic information like locations, classes, and relations. The segmentation problem can then be naturally tackled by applying clustering or threshold on the eigenvectors. Extensive experiments are conducted on various datasets (e.g., EndoVis2017, EndoVis2018, UCL, etc.) for different clinical endpoints. Across all the challenging scenarios, our method demonstrates outstanding performance and robustness higher than unsupervised state-of-the-art (SOTA) methods. The code is released at https://github.com/MingyuShengSMY/GraphClusteringSIS.git.

CCS Concepts

• **Computing methodologies** → **Video segmentation**.

Keywords

Surgical Instrument Segmentation, Unsupervised Learning, Graph Partitioning

ACM Reference Format:
Mingyu Sheng, Jianan Fan, Dongnan Liu, Ron Kikinis, and Weidong Cai. 2024. Revisiting Surgical Instrument Segmentation Without Human Intervention: A Graph Partitioning View. In *Proceedings of the 1st International Workshop on Multimedia Computing for Health and Medicine (MCHM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3688868.3689193

1 Introduction

Minimally invasive surgery (MIS) offers several advantages over standard open surgery, such as reduced pain, lower risk, and shorter recovery period [41]. Its extensive use of endoscopic cameras for probing human body allows surgeons to observe pathological tissues and manage surgical tools effectively. Despite the advantages of MIS, this technique still faces various challenges, including long surgical procedures, intricate tool operations, limited fields of view, and challenging hand-eye coordination [50].

Interest in addressing the limitations of MIS has grown significantly in recent years when various techniques have been developed to help surgeons overcome challenges, such as professional training and evaluation, procedure analysis and optimization, and visual-based frame processing [21, 61]. One effective approach is to utilize positional and movement data of surgical instruments during operations. It can be accomplished with infrared and electromagnetic tracking or external markers directly attached to the tools [7]. However, these methods require additional efforts and delicate preparations, which renders it troublesome to integrate those techniques into existing surgical workflows [60].

Due to these challenges, recent efforts have focused on visual-based endoscopic frame processing. With the explosive growth of machine learning techniques and the promoted accessibility of computational resources, a stream of visual-based approaches has been proposed. Hand-crafted-feature-based object detection appears as a classic method to localize surgical instruments and track them in a video stream [47]. While those approaches facilitate high-speed frame processing, the detection results are often unsatisfactory due to the varying positions, orientations, and shapes of surgical instruments. In an endoscopic video frame, surgical tools typically

MCHM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-1195-4/24/10
https://doi.org/10.1145/3688868.3689193

appear from the corners or edges towards the center, making traditional axis-aligned bounding box detection methods suboptimal [50]. In this context, surgical instrument segmentation (SIS) enables more accurate predictions of surgical instruments at the pixel level and therefore stands as a promising pathway for providing vital assistance to surgeons [62].

However, tremendous labeled ground-truth data are required at the price of training a high-performance supervised SIS model, and several obstacles could hinder the data acquisition procedure. Obtaining and collecting labeled data would rely on the deep involvement of domain experts to distinguish various specific surgical tools, pathological tissues, and organs, which could consume massive labour efforts. Accessing surgical image data might be further restrained due to patient privacy issues. In this regard, semi-supervised and unsupervised methods for SIS arise as worthwhile topics to explore. Compared to semi-supervised learning, unsupervised learning is fully annotation-free and is capable of prospecting more general and high-level patterns (e.g., semantic correlation) hiding in data distribution [38, 48, 63].

Given its intriguing potential, several unsupervised SIS models have been developed [51, 63], where the pseudo-label technique is widely used and plays an important role. The pseudo-label approach provides a straightforward solution to convert an unsupervised learning task into a pseudo-supervised one, thereby being easily implemented for training neural networks. Low-level features are the majority materials to derive pseudo-labels, such as color thresholds and edge detection [56]. Alternatively, pseudo-labels could also be generated by clustering the deep feature obtained from an encoder model [35].

On the other hand, the effectiveness of the above technique highly relies on the quality and accuracy of the pseudo-label. Therefore, it may suffer from certain limitations in the SIS field, especially when the pseudo-labels are generated from low-level features. For example, most endoscopic images are likely of low quality, containing significant noise and blur; the scenes of endoscopic images involve intermixed tissues with irregular shapes, colors, and fuzzy tissue connectives. As a result, the quality of pseudo-labels generated from low-level features may be compromised.

In a nutshell, the research gaps in the current field of SIS are summarized as: 1) the performance and robustness of supervised models are limited due to the scarcity of labeled data; 2) the pseudo-label technique may lead to erratic or unstable outcomes, especially with low-quality and intricate endoscopic images; and 3) there is a notable lack of studies on label-free methods for the SIS task.

To overcome the above problems, we devise a label-free unsupervised method. A self-supervised pre-trained model is leveraged as a feature extractor, based on Vision Transformer (ViT) to capture the global high-level context features from surgical video frames. We map the segmentation task into a graph-partitioning problem by eigendecomposing a "deep" Laplacian matrix calculated from the deep dense features. The segmentation mask can be predicted by conducting clustering or thresholding on top of the eigenvectors. Our method outperforms unsupervised SOTA methods on both binary and multi-class segmentation tasks and demonstrates significant robustness across different datasets, as evidenced by the experimental results.

Our key contributions are summarized as below:

- We propose an unsupervised method that can handle all categories of surgical instrument segmentation (SIS) tasks, including binary, part, type, and semantic segmentation tasks.
- We develop a novel framework that solves the SIS task via graph theory in a graph-cutting manner, that the high-level context features are regarded as a graph and its pixels as nodes.
- We introduce unsupervised graph clustering applied on eigenvectors decomposed from the Laplacian matrix of deep features, and unsupervised salient detection based on the Fiedler Vector, where the most apparent object is effectively detected in surgical video frames.
- We test our method on various datasets, conducting extensive and comprehensive experiments, to demonstrate our method's SOTA performance and robustness.

In Section 2, we reviewed related studies, including recent unsupervised segmentation methods in both the natural image field and the SIS field. Our approach is then introduced in Section 3. The experiment details and results are reported and analyzed in Section 4, followed by the conclusion and future expectations in Section 5.

2 Related Work

2.1 Unsupervised Image Segmentation

Compared with image classification, image segmentation is a pixel-wise classification task. There are three main categories for image segmentation: salient segmentation, semantic segmentation, and instance segmentation. Salient segmentation (also called salient detection) aims to segment the most salient object or a specific object as foreground against the background. Semantic segmentation is a multi-class task that segments and classifies all labeled objects according to their semantic features. Instance segmentation is also a multi-class task, but every object is considered a unique instance or class. Unsupervised segmentation indicates that no annotated ground-truth is provided for training.

For the earliest works before 2018, unsupervised image segmentation was typically addressed by using traditional machine learning and computer vision methods, such as threshold, watershed method, and Markov Random Field (MRF) [10, 22, 65]. Nowadays, deep learning techniques are widely leveraged in this field [14, 23, 33, 37, 40]. In our work, unsupervised image segmentation methods are basically classified into two types: *Pseudo-Label* based and *Label-Free* based approaches.

Pseudo-Label. An alternative approach for unsupervised image segmentation is using pseudo-label as supervision. The pseudo-label is usually generated from low-level features, predicted from pre-trained models, or clustered from the output feature vectors [33, 35], such as PiCIE [29] via clustering and MaskContrast [56] via a pre-trained model. When the pseudo-labels are generated from clustering, the generating and training processes proceed by loop until convergence. Regarding the pseudo-labels as hints and supervision, contrastive learning is a popular training strategy adopted in pseudo-label-based methods to increase and decrease the similarity between two features according to their correlation [29, 33, 56]. However, the pseudo-label technique can be unreliable and inaccurate for tasks containing complex scenes and low-quality

images (e.g., blur, light reflection, and narrow perspective in endoscopic image frames), and thus uncertainly trains the model. Weak robustness is another problem, for which a model trained on a dataset with its pseudo-labels as supervision cannot be effortlessly generalized to other situations/scenes.

Label-Free. Extracting high-level dense features via an image encoder and then clustering (usually by K-Means) is a general framework for label-free methods. A promising approach to further enhance the model performance is by extending the encoder with a segmentation head and optimizing/training the head by maximizing the consistency among output feature maps extracted from an image and its random augmentations or nearest neighbors [28, 31], such as STEGO [23]. *Graph method* is recently a new topic for unsupervised learning because of its outstanding accuracy and robustness over different data distributions. It transforms image segmentation tasks into graph-partitioning tasks by treating pixels as nodes and their similarities as edges [15, 42, 59]. In addition, another promising approach is directly generating segmentation masks by training a *Generative Adversarial Network (GAN)*. To date, most GAN-based methods are proposed mainly focusing on salient segmentation, such as [1, 5, 6, 12, 40]. Nevertheless, how to efficiently and stably train a GAN model is an inevitable challenge that would be more severe due to insufficient data.

2.2 Surgical Instrument Segmentation

Surgical instrument segmentation (SIS) is to classify every single pixel of a medical endoscopic image into a specific class including background, instrument, and pathological tissue. There are four categories of SIS tasks: *Binary*, *Part*, *Type*, and *Semantic Segmentation*. Binary segmentation is to segment the instrument and background; part segmentation extends it to classify different parts of instruments, like shaft, wrist, and clasper; type segmentation aims to identify different instrument types, like clamps, suturing needles, and threads; semantic segmentation categorizes all objects in an endoscopic image frame, including instruments, tissues, organs, and even bleeding areas.

Supervised SIS. [8] firstly proposed a detector based on Support Vector Machine (SVM), designed for instrument segmentation and attitude estimation; their work marked a significant beginning in machine-learning-based surgical tool segmentation. Following their work, an increasing number of studies [4, 17, 18, 34] are proposed, leveraging deep learning techniques (e.g., LSTM, CNN, etc.). Various classic backbones were utilized in these works, such as FCN [39], U-Net [49], and ResNet [26]. To further improve performance, multi-scale spatial features were widely employed to capture more features including low-level and high-level semantic features [30, 32, 43–45, 52, 66]. In addition, attempts to extract temporal features from surgical videos have been made for fully leveraging the semantic correlation among continuous frames [19, 58]. However, the performance and generalization ability of supervised methods are constrained by data scarcity.

Unsupervised SIS. [36] firstly proposed an unsupervised method called AGSD for the binary SIS task by utilizing pseudo-labels generated from low-level features (e.g., color and lightness). They assume that most surgical instruments are displayed with higher lightness and plainer colors than background tissues. This strategy may be effective for simple scenes but can be challenged in other more complex scenes, and cannot tackle multi-class segmentation. Similarly, [11] adopted Region Adjacency Graph (RAG) to generate pseudo-label and trained a model based on Masked Autoencoders (MAE) [25], whereas their main target is to segment blood vessels instead of instruments. [51] leveraged optical flow for supervision and shape-prior as a hint to train a model based on the Teacher-Student structure. Nevertheless, the recent pseudo-label-based methods in the SIS field are restrained to binary segmentation, and robustness is limited due to the drawbacks and limitations of low-level pseudo-labels. In this study, recognizing the above limitations and inspired by the graph theory, we proposed a label-free method based on graph cutting, demonstrating the SOTA performance and robustness on various SIS datasets.

3 Method

Our proposed approach is illustrated in Figure 1. It takes a single frame from a surgical video stream as input whose deep context feature map is extracted from a self-supervised pre-trained model. Then, the Laplacian matrix is calculated from the feature map for the subsequent partitioning progress (i.e., graph clustering and salient detection).

3.1 Backbone

We employ a self-supervised pre-trained model named DINO (developed by [9] based on ViT) as a feature extractor, noted as Θ. The DINO is trained in a distillation approach and can effectively capture global high-level context information. Let $I \in \mathbb{R}^{H\times W\times 3}$ do an RGB frame from a surgical video, where H and W are the frame height and width, respectively. Let $M \in \mathbb{R}^{h\times w\times d}$ represent the extracted feature map, where d indicates feature channels, h and w are the height and width, respectively. The feature extraction is denoted by $M = \Theta(I)$. Then, the feature map M is reshaped into a feature matrix $F \in \mathbb{R}^{s\times d}$ for the subsequent procedures, where $s = h \times w$ and its row $f_i \in \mathbb{R}^d$ represents a feature vector of pixel i.

3.2 Affinity Matrix Computation

Assuming the feature matrix F as a graph $G = (P, E)$ where P is the set of nodes/pixels, and E is the set of edges/similarities. A positive semi-definite affinity matrix $W \in \mathbb{R}^{s\times s}$ is computed to represent the similarities among pixels, whose element is denoted by $w_(i, j)$ measuring the affinity between pixels i and j. The computation of the affinity matrix W follows:

$$w_{i,j} = \begin{cases} \cos(f_i, f_j) & i \neq j \text{ and } \cos(f_i, f_j) > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

where $\cos(\cdot, \cdot)$ indicates cosine similarity between two vectors. Apart from normalizing the cosine similarity into [0, 1] to get the positive semi-definite affinity matrix, we particularly threshold the similarity by 0, because in our preliminary experiments, eigendecomposing a dense matrix is extremely time-consuming, and sometimes even fails due to ill-conditioning.

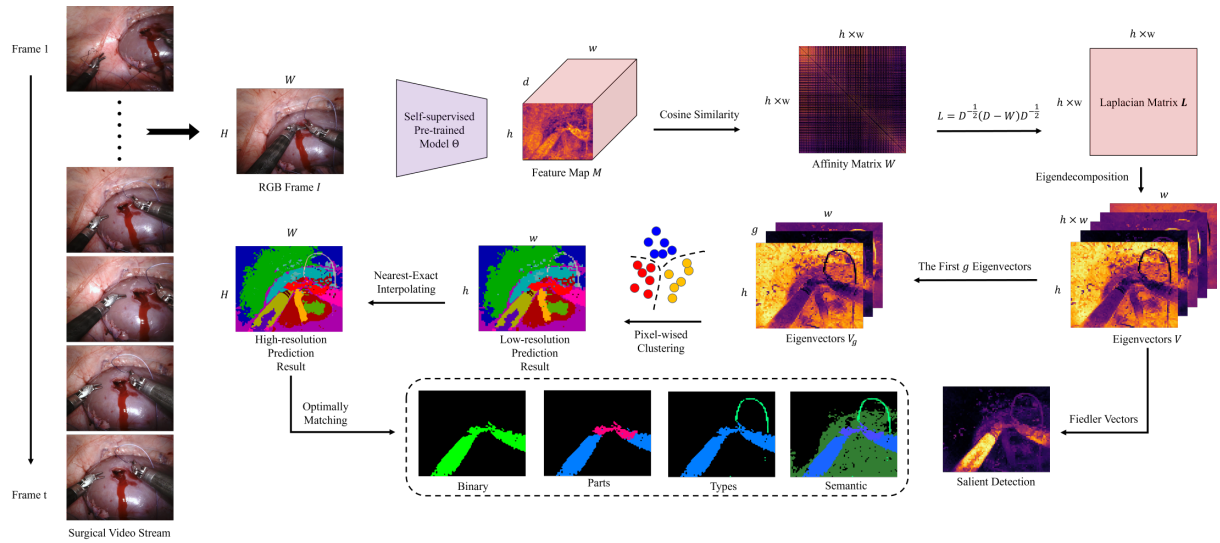

Figure 1: Overview of Our Method. Every surgical video frame is fed into a ViT-based feature extractor to generate high-level dense features. Then, an affinity matrix W is computed and its Laplacian matrix L is calculated for the subsequent eigendecomposition, from which eigenvectors provide distinct features to distinguish different modules in a frame, where the first g eigenvectors are stacked together for clustering and the second eigenvector (the Fiedler Vector) is leveraged for salient detection.

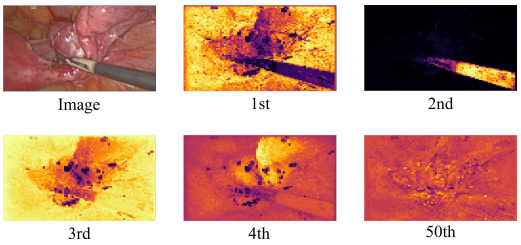

Figure 2: Sample Result of Eigendecomposition. The top-left is the origin image frame, and "i th" represents a visualized eigenvector with the i-th smallest eigenvalue.

3.3 Eigendecomposition of Laplacian Matrix

Background. Normalized Cut [53] introduces that the eigenvectors decomposed from the Laplacian matrix display different modules/partitions in a graph, which conclusion is derived from minimizing a normalized graph cut cost denoted by:

$$Ncut(A,B) = \frac{\sum_{i\in A, j\in B} w_{i,j}}{\sum_{i\in A, j\in P} w_{i,j}} + \frac{\sum_{i\in A, j\in B} w_{i,j}}{\sum_{i\in B, j\in P} w_{i,j}}. \tag{2}$$

Eigenvectors with smaller eigenvalues indicate lower cut cost, thus presenting more accurate cutting for a graph. Given the matrix W, its Laplacian matrix is computed by:

$$L = D^{-1/2}(D-W)D^{-1/2}, \tag{3}$$

where $D \in \mathbb{R}^{s\times s}$ is a diagonal matrix whose entries are the row-summation of W.

In this study, we apply eigendecomposition on the "deep" Laplacian matrix. Let $V \in \mathbb{R}^{s\times s}$ note column-stacked eigenvectors sorted in ascent according to their eigenvalues, where the i-th smallest eigenvector is represented by $v_i \in \mathbb{R}^s$. The eigenvectors effectively express informative and distinctive modules in the surgical image. Some of them for a surgical video frame are presented in Figure 2, in which the eigenvectors with lower eigenvalues such as "1st" and "2nd" meaningfully segment the endoscopic image frame, where the background and surgical tool are feasibly distinguished. The eigenvectors with comparatively higher eigenvalues like "3rd" and "4th" also demonstrate informative modules in the image, where the background is further finely segmented into different semantic modules (e.g., light reflection and operated target tissues). However, as the eigenvalue increases, the eigenvector becomes chaotic and noise-like, such as "50th". With the eigenvectors, the two different strategies for graph cutting are introduced in the following sections.

3.4 Salient Detection

The eigenvector with the second-smallest eigenvalue is called Fiedler Vector [16] denoted by v_2, significantly detecting the most salient object in an image. For example, the "2nd" in Figure 2 effectively detects the surgical tool with high magnitudes on its shaft and head. In this case, a binary segmentation mask can be obtained by thresholding the Fiedler Vector with a certain value often with 0. We follow the common threshold method, setting the elements greater than 0 as foreground, and background otherwise. This binary-partitioning method can only deal with the binary segmentation task. Therefore, another more general method is introduced in the next section to deal with all four segmentation tasks.

3.5 Graph Clustering

In this section, the unsupervised segmentation task is solved via clustering. Firstly, informative eigenvectors are selected according

to their eigenvalues. The first g smallest eigenvectors are selected because lower eigenvalues indicate more meaningful feature representation possessed. Let $V_g \in \mathbb{R}^{s\times g}$ note the stacked selected eigenvectors. By reshaping the V_g into $\hat{V}_g \in \mathbb{R}^{h\times w\times g}$, an embedded feature map is obtained, consisting of the meaningful eigenvectors. The K-Means algorithm with the given cluster number k is implemented on the new feature map to cluster the pixels into different partitions. The clustering result is denoted by $\hat{Y} \in \mathbb{R}^{h\times w}$.

3.6 Interpolation

The preliminary prediction is a low-resolution label mask. The final high-resolution prediction noted by $Y \in \mathbb{R}^{H\times W}$ is calculated by Nearest-Exact Interpolation (NEI). Compared to the normal Nearest-Neighbor Interpolation (NNI), the NEI searches for a more accurate value by considering more pixels around the position instead of only one. For example, at a given position, the NNI only finds one pixel closest to the position and directly assigns the pixel value to it; in contrast, the NEI searches for the majority value around the interpolating position and provides precise interpolation.

3.7 Optimal Matching

Our method is fully unsupervised without any access to the ground-truth labels. Therefore, the prediction labels are optimally matched to the ground-truth labels for visualization and quantitative analysis. We employed two matching strategies: Hungarian matching and Majority-Vote matching. Hungarian matching is used to uniquely match each ground-truth label with every prediction if the numbers of clusters and classes are equal, otherwise, Majority-Vote matching is leveraged, which means a ground-truth label may be repeatedly matched to multiple prediction labels.

4 Experiment and Results

4.1 Evaluation Metric and Datasets

Evaluation Metric. The mean Intersection over Union (mIoU) is calculated to analyze the method performance. mIoU is an averaged IoU over all classes, in which IoU is a pixel-wised measure that comprehensively reflects the segmentation accuracy.

Datasets. Our method is tested on five benchmark datasets for four segmentation tasks. The complexity, difficulty, and object classes are various across the datasets, significantly challenging the method's performance and generalization ability. *EndoVis2017*, a dataset released in 2017 MICCAI EndoVis Robotic Instrument Segmentation Challenge [3]. Binary, part, and type segmentation are available in this dataset. Similar to EndoVis2017, *EndoVis2018* was released in 2018 [2]. Binary, part, type, and semantic segmentation tasks are available for this dataset. *ARTNetDataset* is an image-based dataset proposed and annotated by [24]. Only binary labels are officially provided. *CholecSeg8k* [27] is a subset of the endoscopic dataset Cholec80 [55]. Only the semantic segmentation is officially available in this dataset. *UCL*. An ex-vivo synthetic dataset consists of 20 videos [13]. Scenes in UCL are distinctly divergent from the above in-vivo datasets. Binary ground-truth labels are officially provided.

4.2 Implementation Details

The backbone model DINO is loaded as a feature extractor via Pytorch from https://github.com/facebookresearch/dino. Following the existing works [42, 54], the deep features in the key vector at the last attention layer are extracted for computing the affinity matrix because of its better localization performance. The number of clusters is set as $k = 15$. This is because we refer to the number of labeled classes in EndoVis2018 (12 classes) and CHolecSeg8k (13 classes), where the "background" is a class in the datasets, which can be further finely segmented into more classes (e.g., bleeding blood, tissue connectives, etc.), thereby we determine a slightly larger number for clusters. We set $g = k$ because each eigenvector represents a particular module in a frame, and we assume clustering with the number of modules can yield relatively superior performance, which is partially verified in our ablation study in Section 4.5.

4.3 Comparison Results with SOTA Methods

Binary Segmentation Task. Tables 1 and 2 report comparisons between our method and SOTA methods, in two aspects of performance and robustness. The three kinds of supervision are noted as "Sup.", "Semi." and "Unsup." for supervised, semi-supervised and unsupervised methods respectively. The underline indicates the best results for all methods, and the best results for unsupervised methods are highlighted in bold. "SAL" and "CLU" represent "salient detection" and "graph clustering" corresponding to the Sections 3.4 and 3.5. AGSD is reproduced in our study but cannot be trained on the non-video dataset (i.e., ARTNetDataset), so there is no corresponding result in Table 1 for AGSD on ARTNetDataset.

Table 1: Binary Segmentation Performance (mIoU [%])**.** Methods noted with "$*$" are reproduced. "CLU" indicate our method based on graph clustering. **Bold** marks the best result across unsupervised methods. Underline marks the best result across all methods.

	Supervision	EndoVis2017	EndoVis2018	ARTNetDataset
ART-Net [24]	Sup.	81.00	-	88.20
DRLIS [46]		89.60	-	-
MF-TAPNet [32]		87.56	-	-
AOMA [63]	Semi.	77.10	68.40	-
Duel-MF [64]		84.05	-	-
FUN-SIS [51]	Unsup.	76.25	-	-
AGSD* [36]		81.08	63.83	-
Ours (CLU)		**81.12**	**79.46**	84.58

On the EndoVis2017 dataset, Table 1 illustrates that our "CLU" method slightly outperforms the AGSD, and distinctly outperforms FUN-SIS by about 5%. Regarding the EndoVis2018 dataset, our method shows dramatically high performance at 79.46% higher than the AGSD by about 17%. The severe performance degeneration for AGSD from 81.08% on EndoVis2017 to 63.83% on EndoVis2018 is because the AGSD method is a pseudo-label-based method trained with pseudo-labels generated from low-level features such as color and lightness, whose effectiveness is severely deducted due to the complex surgical scene in the EndoVis2018 dataset. Compared with some supervised and semi-supervised methods like ART-Net and

Table 2: Binary Segmentation Robustness (mIoU [%]). "SAL" indicates our method based on salient detection.

	EndoVis2017	EndoVis2018	UCL	Avg.	Std.
AGSD* [36]	81.08	63.83	78.94	74.62	± 7.68
AGSD* (o-1)	72.52	78.70	58.71	69.98	± 8.36
Ours (SAL)	65.09	64.04	65.51	64.88	± **0.62**
Ours (CLU)	**81.12**	**79.46**	**79.75**	**80.11**	± 0.72

AOMA, our method demonstrates higher performance, but slightly lower than most supervised methods like Duel-MF and DRLIS.

For the generalization ability comparison across unsupervised methods, as reported in Table 2, the SOTA method AGSD shows comparatively weaker robustness indicated by the high standard deviation of ± 7.68% across the three datasets, higher than ours by 7.06% and 6.96% for "SAL" (± 0.62%) and "CLU" (± 0.72%) respectively. Performance degeneration of AGSD is exposed when testing on unseen datasets, except for the EndoVis2018 dataset, where mIoU increases, because the quality and reliability of pseudo-labels for EndoVis2018 are severely disrupted due to its high complexity, thus the "AGSD" is under-fitted on EndoVis2018 compared to "AGSD (o-1)". On the other hand, when AGSD is trained on the two EndoVis datasets and tested on the UCL dataset, the mIoU severely declines to 58.71% due to the distinctly different scenes in the UCL dataset, such as tissue textures, lightness, and angles of view. In contrast, our method demonstrates outstanding robustness across different datasets justified with a small standard deviation at only ± 0.62% and ± 0.72% for "SAL" and "CLU" respectively. In particular, our "SAL" method shows inferior performance than others because only one eigenvector (the Fiedler Vector) is used, and its dramatic stability across the three datasets indicates its capability for stable detection on a wider range of scenes.

Multi-class Segmentation Tasks. The recent SOTA unsupervised methods cannot tackle multi-class segmentation tasks (i.e., part, type, and semantics) because of the limitations of pseudo-labels that can only reflect pixel labels in binary. In contrast, our CLU method is capable of both binary and multi-class segmentation tasks. We report the experimental results in Tables 3, 4, and 5 for part, type, and semantic segmentation respectively.

Table 3: Part Segmentation Performance (mIoU [%]).

	Supervision	EndoVis2017	EndoVis2018
DRLIS [46]	Sup.	<u>76.40</u>	-
DMNet [58]		-	<u>67.50</u>
MF-TAPNet [32]		67.92	-
Duel-MF [64]	Semi.	62.51	-
Ours (CLU)	Unsup.	59.23	57.52

In Table 3, we compare our method with supervised and semi-supervised methods on the part segmentation task. For the EndoVis2017 dataset, our method performance is close to the semi-supervised method Duel-MF with a small gap of 3%, whereas there is still a large gap of 17% compared to the fully supervised method

Table 4: Type Segmentation Performance (mIoU [%]).

	Supervision	EndoVis2017	EndoVis2018
BAANet [52]	Sup.	61.59	42.67
DMNet [58]		53.89	-
SurgNet [45]		66.30	-
LWANet [43]		58.30	-
ISINet [19]		38.08	45.29
MF-TAPNet [32]		36.62	-
SurgicalSAM [62]		<u>67.03</u>	<u>58.87</u>
Duel-MF [64]	Semi.	52.80	-
Ours (CLU)	Unsup.	58.86	44.68

Table 5: Semantic Segmentation Performance (mIoU [%]).

	Supervision	EndoVis2018	CholecSeg8k
SwinSP-TCN [20]	Sup.	-	<u>69.38</u>
Noisy-LSTM [57]		<u>62.30</u>	-
Ours (CLU)	Unsup.	46.46	46.31

DRLIS. Our method achieves a mIoU score of 57.52% on the EndoVis2018 dataset, which is slightly lower than that of EndoVis2017 due to the difficulty and complexity of the EndoVis2018 dataset. The experimental results on the type segmentation task are reported in Table 4, where our method shows medium-level performance at 58.86% on EndoVis2017. However, relatively low performance is exposed on the EndoVis2018 dataset at 44.68%, close to some supervised methods (e.g., BAANet and ISINet) but remaining a distinct gap (about 15%) to the SOTA supervised method SurgicalSAM at 58.87%. The experiment for the semantic segmentation task uses the EndoVis2018 and CholecSeg8k datasets, as demonstrated in Table 5. Although, we are the first unsupervised method dealing with the semantic segmentation task where the mIoU scores are 46.46% (for EndoVis2018) and 46.31% (for CholecSeg8k), supervised methods outperform our unsupervised method by 15.84% and 23.07% on EndoVis2018 and CholecSeg8k respectively.

4.4 Results Visualization

Eigenvectors Visualization. Samples of eigenvectors are visualized in Figure 3, where every eigenvector reflects a particular module in its corresponding image. For example, in Figure 3(a), the surgical tool is entirely detected in the first and second eigenvectors, suggesting sufficient information for the binary segmentation task, and in the third eigenvector, the part of the tool head is detected with the highest significance while the lowest magnitude for the tool shaft, thereby the part segmentation task can be easily solved. In terms of Figure 3(b), combining the second and third eigenvector, the surgical tools and suturing thread are practically distinguished in the purpose of the type segmentation task. For the challenging semantic segmentation task, in Figures 3(c) and 3(d), the organs and tissues are detected in the fourth and third eigenvectors respectively. The detection becomes more fine-grained, as eigenvalues increase, such as the "10th" eigenvector where more

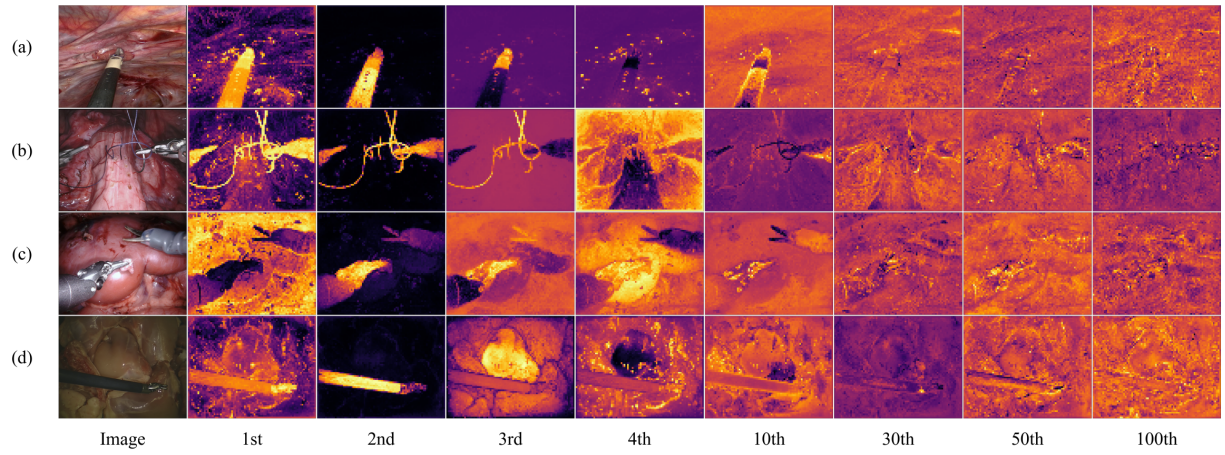

Figure 3: Eigenvectors Visualization. Each row demonstrates the input image and its corresponding eigenvectors. (a) - (d) are from ARTNetDataset, EndoVis2017, EndoVis2018, and UCL respectively. "i th" indicates the eigenvector with the i-th smallest eigenvalue. The Fiedler vector is denoted by "2nd".

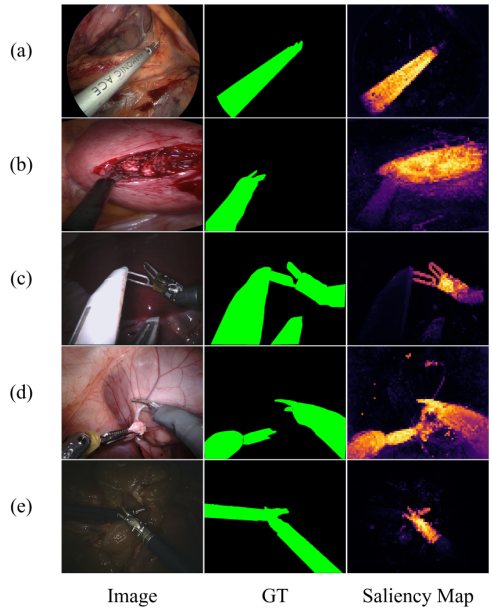

Figure 4: Salient Detection Visualization. Each row illustrates an input image (Image), its binary ground-truth (GT), and the corresponding saliency map generated from the Fiedler vector. (a) and (b) are from ARTNetDataset; (c), (d), and (e) are from EndoVis2017, EndoVis2018 and UCL, respectively.

details like textures are detected. However, for the eigenvectors with overlarge eigenvalues (e.g., "30th", "50th" and "100th"), the detection becomes unintuitive and noise-like, which may negatively influence the method performance.

Salient Detection Visualization. Samples of salient detection results on the binary SIS task are visualized in Figure 4. The outstanding effectiveness of salient detection is illustrated in Figure 4(a) where the tool is detected as a salient object with high significance in the saliency map normalized from the corresponding Fiedler vector. However, the limitations and disadvantages of salient detection are displayed in Figures 4(b) - 4(e). The most common shortcoming is partial detection, such as in Figures 4(c) - 4(e) where the instrument heads are detected as salient objects, whereas shafts are omitted or assigned low significance. Neglecting detection is another drawback as shown in Figures 4(c) and 4(e) where only one instrument is partially detected, even though two or three tools appearing in the images. Another infrequent but severe disadvantage is misdiagnosis as shown in Figure 4(b) where the tool is barely detected with very low significance, while the surgical incision is emphasized with high significance.

Binary Segmentation Visualization. Qualitative visualizations of our CLU method are illustrated in Figure 5. Our method is more capable of detecting the entire surgical tools, such as in Figures 5(b) and 5(e). More fine-grained objects are practically segmented via our method, such as tool clamps in Figures 5(c) and 5(d). For Figures 5(b), 5(d) and 5(e), comparatively strong robustness of our method on lightless and over-lighted frames is demonstrated, while the severe light-reflection and the extreme darkness in the image frames significantly impact the AGSD method. However, some drawbacks are exposed in our method. For instance, the threads in Figure 5(a) are wrongly detected when they are connected with tools; and minor mis-segmentation on the light-reflection in Figure 5(d) that have a similar color to surgical instruments. Rough detection edges and low segmentation quality/resolution are inevitable disadvantages because the final results are obtained by interpolating the low-resolution prediction mask.

4.5 Ablation Studies

We conduct ablation experiments on the EndoVis2018 dataset because it supports all four segmentation tasks. As shown in Table 6, the hyper-parameter g influences performance promotion and degeneration. Increasing k always leads to higher performance because of over-clustering. For few-class segmentation tasks (i.e.,

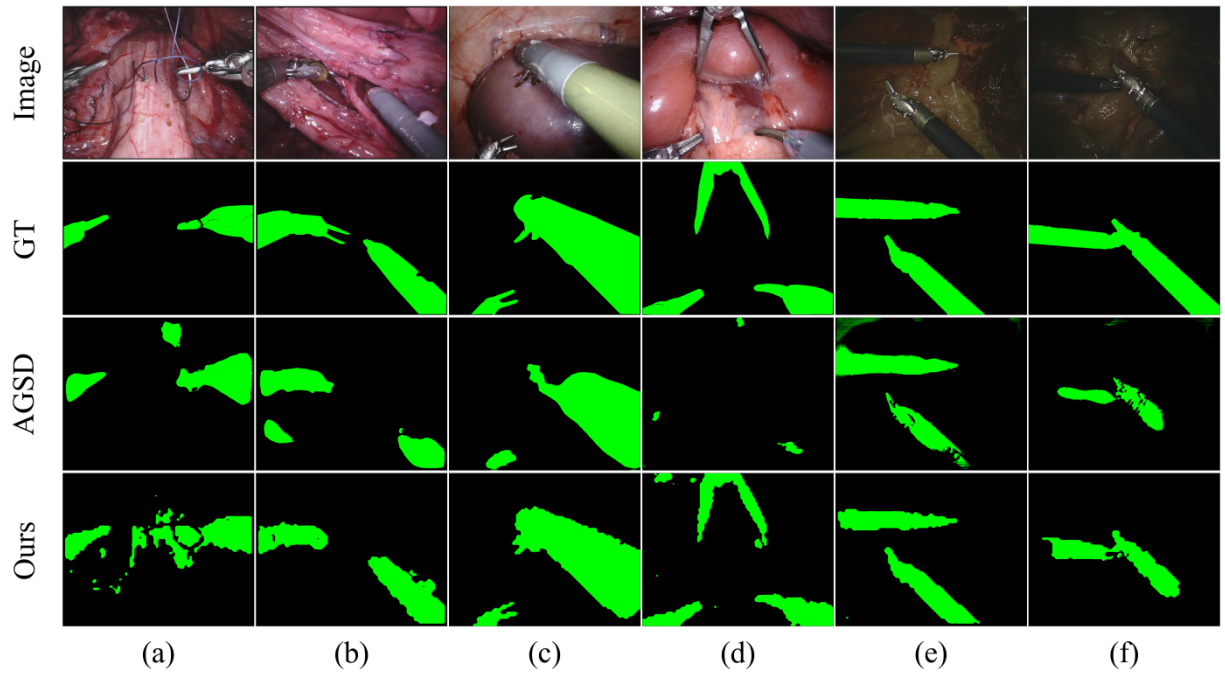

Figure 5: Binary Segmentation Visualization. Each column demonstrates the input image (Image), binary ground-truth (GT), and prediction masks of AGSD and our CLU method. (a) and (b) are from EndoVis2017, (c) and (d) are from EndoVis2018, (e) and (f) are from UCL, where (a) shows a fail case of our CLU method.

binary and part), the suitable g is approximately located in fixed numbers such as $g = 10$ for "Binary" and $g = 15$ for "Part". In terms of complex segmentation tasks (i.e., type and semantic), the optimal g depends on the number of clusters k, and $g = k$ often yields superior performance, as shown in Table 6 for "Type" and "Semantic". The main reason for the performance degeneration caused by overlarge g is owed to the unintuitive and noise-like eigenvectors possessing large eigenvalues, as illustrated in Figure 3.

5 Conclusion and Future Work

This work proposes an unsupervised SIS method based on graph theory, where a ViT-based model is used as a backbone to capture the global context information, and the image is segmented in a graph partitioning manner by regarding the pixels in the feature map as nodes in a graph. Then, the Laplacian matrix is calculated from the feature map and decomposed into "deep" eigenvectors representing particular semantic modules in an image (e.g., tools, threads, and organs), providing distinguishable information. The performance and robustness of our method are verified for the different SIS tasks (i.e., binary, part, type, and semantic) by conducting comprehensive experiments on various datasets such as CholedSeg8k, UCL, and EndoVis datasets. Nevertheless, due to the time-consuming eigendecomposition, low-resolution feature map, and unascertainable hyper-parameter g, computational inefficiency, relatively poor prediction quality, and extra effort for locating optimal g are inevitable limitations of our methods. Therefore, future work should focus on developing a real-time, high-resolution, and adaptive unsupervised method that can generate high-quality prediction masks promptly and adaptively determine the hyper-parameter g, which may involve extra neural network structures, and design a strategy for locating the optimal g that critically influences the method performance.

Table 6: Ablation Study for Different g and k. "g" is the number of used eigenvectors. "k" is the number of clusters. **Bold** indicates the best result for each k.

	mIoU (%) k \ g	2	3	5	10	15	20	30
Binary	5	73.54	**76.05**	73.80	69.28	65.14	64.02	61.75
	10	74.07	78.05	78.67	**79.77**	77.39	75.48	72.51
	15	74.17	78.53	79.71	**80.20**	79.46	78.81	77.07
	20	74.69	78.87	80.79	**80.85**	80.36	79.91	79.33
	30	75.01	79.21	81.82	**82.24**	81.82	81.46	81.20
Part	5	44.69	**48.77**	48.58	45.27	41.54	40.13	37.80
	10	45.61	51.28	53.28	**56.79**	54.61	51.85	48.85
	15	45.82	51.68	54.13	57.13	**57.52**	56.85	55.10
	20	46.27	51.85	55.23	58.41	**58.75**	58.60	57.83
	30	46.52	52.11	56.24	59.50	**59.94**	59.86	59.79
Type	5	30.49	32.81	**34.35**	31.54	28.64	26.70	25.06
	10	32.55	36.24	39.46	**42.64**	40.41	39.81	38.12
	15	33.69	36.81	40.81	43.49	**44.68**	43.99	42.36
	20	34.15	37.58	41.62	43.93	46.03	**46.80**	45.48
	30	34.99	38.01	42.34	44.70	47.51	48.29	**49.70**
Semantic	5	27.45	30.87	**34.19**	31.91	29.28	27.83	26.22
	10	29.85	34.61	39.56	**43.06**	42.14	41.31	39.72
	15	30.90	36.09	41.45	45.54	**46.46**	46.45	45.70
	20	31.57	36.94	42.85	47.15	48.57	49.16	**49.70**
	30	32.43	37.74	44.28	49.10	51.36	51.99	**53.35**

References

[1] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. 2021. Labels4Free: Unsupervised Segmentation using StyleGAN. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 13950–13959. https://doi.org/10.1109/ICCV48922.2021.01371

[2] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, Avinash Kori, Varghese Alex, Ganapathy Krishnamurthi, David Rauber, Robert Mendel, Christoph Palm, Sophia Bano, Guinther Saibro, Chi-Sheng Shih, Hsun-An Chiang, Juntang Zhuang, Junlin Yang, Vladimir Iglovikov, Anton Dobrenkii, Madhu Reddiboina, Anubhav Reddy, Xingtong Liu, Cong Gao, Mathias Unberath, Myeonghyeon Kim, Chanho Kim, Chaewon Kim, Hyejin Kim, Gyeongmin Lee, Ihsan Ullah, Miguel Luna, Sang Hyun Park, Mahdi Azizian, Danail Stoyanov, Lena Maier-Hein, and Stefanie Speidel. 2020. 2018 Robotic Scene Segmentation Challenge. https://doi.org/10.48550/arXiv.2001.11190 arXiv:2001.11190 [cs.CV]

[3] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, Luis Herrera, Wenqi Li, Vladimir Iglovikov, Huoling Luo, Jian Yang, Danail Stoyanov, Lena Maier-Hein, Stefanie Speidel, and Mahdi Azizian. 2019. 2017 Robotic Instrument Segmentation Challenge. https://doi.org/10.48550/arXiv.1902.06426 arXiv:1902.06426 [cs.CV]

[4] Mohamed Attia, Mohammed Hossny, Saeid Nahavandi, and Hamed Asadi. 2017. Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 3373–3378. https://doi.org/10.1109/SMC.2017.8123151

[5] Yaniv Benny and Lior Wolf. 2020. OneGAN: Simultaneous Unsupervised Learning of Conditional Image Generation, Foreground Segmentation, and Fine-Grained Clustering. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 514–530. https://doi.org/10.1007/978-3-030-58574-7_31

[6] Adam Bielski and Paolo Favaro. 2019. Emergence of Object Segmentation in Perturbed Generative Models. *Advances in Neural Information Processing Systems* 32 (2019). https://proceedings.neurips.cc/paper_files/paper/2019/file/af8d9c4e238c63fb074b44eb6aed80ae-Paper.pdf

[7] Loubna Bouarfa, Oytun Akman, Armin Schneider, Pieter P. Jonker, and Jenny Dankelman. 2012. In-vivo real-time tracking of surgical instruments in endoscopic video. *Minimally Invasive Therapy & Allied Technologies* 21, 3 (2012), 129–134. https://doi.org/10.3109/13645706.2011.580764 PMID: 21574828.

[8] David Bouget, Rodrigo Benenson, Mohamed Omran, Laurent Riffaud, Bernt Schiele, and Pierre Jannin. 2015. Detecting Surgical Tools by Modelling Local Appearance and Global Shape. *IEEE Transactions on Medical Imaging* 34, 12 (2015), 2603–2617. https://doi.org/10.1109/TMI.2015.2450831

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9630–9640. https://doi.org/10.1109/ICCV48922.2021.00951

[10] Kai-Yueh Chang, Tyng-Luh Liu, and Shang-Hong Lai. 2011. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In *Computer Vision and Pattern Recognition (CVPR)*. 2129–2136. https://doi.org/10.1109/CVPR.2011.5995415

[11] Jiachen Chen, Mengyang Li, Hu Han, Zhiming Zhao, and Xilin Chen. 2024. SurgNet: Self-Supervised Pretraining With Semantic Consistency for Vessel and Instrument Segmentation in Surgical Images. *IEEE Transactions on Medical Imaging* 43, 4 (2024), 1513–1525. https://doi.org/10.1109/TMI.2023.3341948

[12] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. 2019. Unsupervised Object Segmentation by Redrawing. *Advances in Neural Information Processing Systems* 32 (2019). https://proceedings.neurips.cc/paper_files/paper/2019/file/32bbf7b2bc4ed14eb1e9c2580056a989-Paper.pdf

[13] Emanuele Colleoni, Philip Edwards, and Danail Stoyanov. 2020. Synthetic and Real Inputs for Tool Segmentation in Robotic Surgery. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz (Eds.). Springer International Publishing, Cham, 700–710. https://doi.org/10.1007/978-3-030-59716-0_67

[14] Jianan Fan, Dongnan Liu, Hang Chang, Heng Huang, Mei Chen, and Weidong Cai. 2023. Taxonomy Adaptive Cross-Domain Adaptation in Medical Imaging via Optimization Trajectory Distillation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 21117–21127. https://doi.org/10.1109/ICCV51070.2023.01936

[15] Jianan Fan, Dongnan Liu, Hang Chang, Heng Huang, Mei Chen, and Weidong Cai. 2024. Seeing Unseen: Discover Novel Biomedical Concepts via Geometry-Constrained Probabilistic Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11524–11534.

[16] Miroslav Fiedler. 1973. Algebraic connectivity of graphs. *Czechoslovak mathematical journal* 23, 2 (1973), 298–305. https://dml.cz/handle/10338.dmlcz/101168

[17] Luis C. García-Peraza-Herrera, Wenqi Li, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, and Sébastien Ourselin. 2017. Real-Time Segmentation of Non-rigid Surgical Tools Based on Deep Learning and Tracking. In *Computer Assisted and Robotic Endoscopy*, Terry Peters, Guang-Zhong Yang, Nassir Navab, Kensaku Mori, Xiongbiao Luo, Tobias Reichl, and Jonathan McLeod (Eds.). Springer International Publishing, Cham, 84–95. https://doi.org/10.1007/978-3-319-54057-3_8

[18] Luis C. García-Peraza-Herrera, Wenqi Li, Lucas Fidon, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, and Sébastien Ourselin. 2017. ToolNet: Holistically-nested real-time segmentation of robotic surgical tools. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 5717–5722. https://doi.org/10.1109/IROS.2017.8206462

[19] Cristina González, Laura Bravo-Sánchez, and Pablo Arbelaez. 2020. ISINet: An Instance-Based Approach for Surgical Instrument Segmentation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020*, Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz (Eds.). Springer International Publishing, Cham, 595–605. https://doi.org/10.1007/978-3-030-59716-0_57

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

[21] Tamás Haidegger, Stefanie Speidel, Danail Stoyanov, and Richard M. Satava. 2022. Robot-Assisted Minimally Invasive Surgery—Surgical Robotics in the Data Age. *Proc. IEEE* 110, 7 (2022), 835–846. https://doi.org/10.1109/JPROC.2022.3180350

[22] Emanuela Haller and Marius Leordeanu. 2017. Unsupervised Object Segmentation in Video by Efficient Selection of Highly Probable Positive Features. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 5095–5103. https://doi.org/10.1109/ICCV.2017.544

[23] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. 2022. Unsupervised Semantic Segmentation by Distilling Feature Correspondences. https://doi.org/10.48550/arXiv.2203.08414 arXiv:2203.08414 [cs.CV]

[24] Md. Kamrul Hasan, Lilian Calvet, Navid Rabbani, and Adrien Bartoli. 2021. Detection, segmentation, and 3D pose estimation of surgical tools using convolutional neural networks and algebraic geometry. *Medical Image Analysis* 70 (2021), 101994. https://doi.org/10.1016/j.media.2021.101994

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15979–15988. https://doi.org/10.1109/CVPR52688.2022.01553

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[27] W. Y. Hong, C. L. Kao, Y. H. Kuo, J. R. Wang, W. L. Chang, and C. S. Shih. 2020. CholecSeg8k: A Semantic Segmentation Dataset for Laparoscopic Cholecystectomy Based on Cholec80. https://doi.org/10.48550/arXiv.2012.12453 arXiv:2012.12453 [cs.CV]

[28] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. 2019. SCOPS: Self-Supervised Co-Part Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 869–878. https://doi.org/10.1109/CVPR.2019.00096

[29] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. 2021. PiCIE: Unsupervised Semantic Segmentation using Invariance and Equivariance in Clustering. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16789–16799. https://doi.org/10.1109/CVPR46437.2021.01652

[30] Mobarakol Islam, Daniel Anojan Atputharuban, Ravikiran Ramesh, and Hongliang Ren. 2019. Real-Time Instrument Segmentation in Robotic Surgery Using Auxiliary Supervised Deep Adversarial Learning. *IEEE Robotics and Automation Letters* 4, 2 (2019), 2188–2195. https://doi.org/10.1109/LRA.2019.2900854

[31] Xu Ji, Andrea Vedaldi, and Joao Henriques. 2019. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9864–9873. https://doi.org/10.1109/ICCV.2019.00996

[32] Yueming Jin, Keyun Cheng, Qi Dou, and Pheng-Ann Heng. 2019. Incorporating Temporal Prior from Motion Flow for Instrument Segmentation in Minimally Invasive Surgery Video. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019*, Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan (Eds.). Springer International Publishing, Cham, 440–448. https://doi.org/10.1007/978-3-030-32254-0_49

[33] Tsung-Wei Ke, Jyh-Jing Hwang, Yunhui Guo, Xudong Wang, and Stella X. Yu. 2022. Unsupervised Hierarchical Semantic Segmentation with Multiview Cosegmentation and Clustering Transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2561–2571. https://doi.org/10.1109/CVPR52688.2022.00260

[34] Iro Laina, Nicola Rieke, Christian Rupprecht, Josué Page Vizcaíno, Abouzar Eslami, Federico Tombari, and Nassir Navab. 2017. Concurrent Segmentation and Localization for Tracking of Surgical Instruments. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne (Eds.). Springer International Publishing, Cham, 664–672. https://doi.org/10.1007/978-3-319-66185-8_75

[35] Måns Larsson, Erik Stenborg, Carl Toft, Lars Hammarstrand, Torsten Sattler, and Fredrik Kahl. 2019. Fine-Grained Segmentation Networks: Self-Supervised Segmentation for Improved Long-Term Visual Localization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 31–41. https://doi.org/10.1109/ICCV.2019.00012

[36] Daochang Liu, Yuhui Wei, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. 2020. Unsupervised Surgical Instrument Segmentation via Anchor Generation and Semantic Diffusion. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020*, Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz (Eds.). Springer International Publishing, Cham, 657–667. https://doi.org/10.1007/978-3-030-59716-0_63

[37] Dongnan Liu, Donghao Zhang, Yang Song, Fan Zhang, Lauren O'Donnell, Heng Huang, Mei Chen, and Weidong Cai. 2020. Unsupervised Instance Segmentation in Microscopy Images via Panoptic Domain Adaptation and Task Re-Weighting. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4242–4251. https://doi.org/10.1109/CVPR42600.2020.00430

[38] Jie Liu, Xiaoqing Guo, and Yixuan Yuan. 2022. Graph-Based Surgical Instrument Adaptive Segmentation via Domain-Common Knowledge. *IEEE Transactions on Medical Imaging* 41, 3 (2022), 715–726. https://doi.org/10.1109/TMI.2021.3121138

[39] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

[40] Yuxin Ma, Yang Hua, Hanming Deng, Tao Song, Hao Wang, Zhengui Xue, Heng Cao, Ruhui Ma, and Haibing Guan. 2021. Self-Supervised Vessel Segmentation via Adversarial Learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 7516–7525. https://doi.org/10.1109/ICCV48922.2021.00744

[41] Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. 2017. Surgical data science for next-generation interventions. *Nature Biomedical Engineering* 1, 9 (2017), 691–696. https://doi.org/10.1038/s41551-017-0132-7

[42] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. 2022. Deep Spectral Methods: A Surprisingly Strong Baseline for Unsupervised Semantic Segmentation and Localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8354–8365. https://doi.org/10.1109/CVPR52688.2022.00818

[43] Zhen-Liang Ni, Gui-Bin Bian, Zeng-Guang Hou, Xiao-Hu Zhou, Xiao-Liang Xie, and Zhen Li. 2020. Attention-Guided Lightweight Network for Real-Time Segmentation of Robotic Surgical Instruments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 9939–9945. https://doi.org/10.1109/ICRA40945.2020.9197425

[44] Zhen-Liang Ni, Gui-Bin Bian, Xiao-Liang Xie, Zeng-Guang Hou, Xiao-Hu Zhou, and Yan-Jie Zhou. 2019. RASNet: Segmentation for Tracking Surgical Instruments in Surgical Videos Using Refined Attention Segmentation Network. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 5735–5738. https://doi.org/10.1109/EMBC.2019.8856495

[45] Zhen-Liang Ni, Xiao-Hu Zhou, Guan-An Wang, Wen-Qian Yue, Zhen Li, Gui-Bin Bian, and Zeng-Guang Hou. 2022. SurgiNet: Pyramid Attention Aggregation and Class-wise Self-Distillation for Surgical Instrument Segmentation. *Medical Image Analysis* 76 (2022), 102310. https://doi.org/10.1016/j.media.2021.102310

[46] Daniil Pakhomov, Vittal Premachandran, Max Allan, Mahdi Azizian, and Nassir Navab. 2019. Deep Residual Learning for Instrument Segmentation in Robotic Surgery. In *Machine Learning in Medical Imaging*, Heung-Il Suk, Mingxia Liu, Pingkun Yan, and Chunfeng Lian (Eds.). Springer International Publishing, Cham, 566–573. https://doi.org/10.1007/978-3-030-32692-0_65

[47] Liang Qiu, Changsheng Li, and Hongliang Ren. 2019. Real-time surgical instrument tracking in robot-assisted surgery using multi-domain convolutional neural network. *Healthcare Technology Letters* 6, 6 (2019), 159–164. https://doi.org/10.1049%2Fhtl.2019.0068

[48] Cristian da Costa Rocha, Nicolas Padoy, and Benoit Rosa. 2019. Self-Supervised Surgical Tool Segmentation using Kinematic Information. In *2019 International Conference on Robotics and Automation (ICRA)*. 8720–8726. https://doi.org/10.1109/ICRA.2019.8794334

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2015*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (Eds.). Springer International Publishing, Cham, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

[50] Tobias Rueckert, Daniel Rueckert, and Christoph Palm. 2024. Corrigendum to "Methods and datasets for segmentation of minimally invasive surgical instruments in endoscopic images and videos: A review of the state of the art" [Comput. Biol. Med. 169 (2024) 107929]. *Computers in Biology and Medicine* 170 (2024), 108027. https://doi.org/10.1016/j.compbiomed.2024.108027

[51] Luca Sestini, Benoit Rosa, Elena De Momi, Giancarlo Ferrigno, and Nicolas Padoy. 2023. FUN-SIS: A Fully Unsupervised approach for Surgical Instrument Segmentation. *Medical Image Analysis* 85 (2023), 102751. https://doi.org/10.1016/j.media.2023.102751

[52] Wenting Shen, Yaonan Wang, Min Liu, Jiazheng Wang, Renjie Ding, Zhe Zhang, and Erik Meijering. 2023. Branch Aggregation Attention Network for Robotic Surgical Instrument Segmentation. *IEEE Transactions on Medical Imaging* 42, 11 (2023), 3408–3419. https://doi.org/10.1109/TMI.2023.3288127

[53] Jianbo Shi and J. Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 8 (2000), 888–905. https://doi.org/10.1109/34.868688

[54] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. 2021. Localizing Objects with Self-Supervised Transformers and no Labels. *Proceedings of the British Machine Vision Conference (BMVC)* (2021). https://doi.org/10.48550/arXiv.2109.14279 arXiv:2109.14279 [cs.CV]

[55] Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. 2017. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Transactions on Medical Imaging* 36, 1 (2017), 86–97. https://doi.org/10.1109/TMI.2016.2593957

[56] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. 2021. Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 10032–10042. https://doi.org/10.1109/ICCV48922.2021.00990

[57] Bowen Wang, Liangzhi Li, Yuta Nakashima, Ryo Kawasaki, Hajime Nagahara, and Yasushi Yagi. 2021. Noisy-LSTM: Improving Temporal Awareness for Video Semantic Segmentation. *IEEE Access* 9 (2021), 46810–46820. https://doi.org/10.1109/ACCESS.2021.3067928

[58] Jiacheng Wang, Yueming Jin, Liansheng Wang, Shuntian Cai, Pheng-Ann Heng, and Jing Qin. 2021. Efficient Global-Local Memory for Real-Time Instrument Segmentation of Robotic Surgical Video. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (Eds.). Springer International Publishing, Cham, 341–351. https://doi.org/10.1007/978-3-030-87202-1_33

[59] Xudong Wang, Rohit Girdhar, Stella X. Yu, and Ishan Misra. 2023. Cut and Learn for Unsupervised Object Detection and Instance Segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3124–3134. https://doi.org/10.1109/CVPR52729.2023.00305

[60] Yan Wang, Qiyuan Sun, Zhenzhong Liu, and Lin Gu. 2022. Visual detection and tracking algorithms for minimally invasive surgical instruments: A comprehensive review of the state-of-the-art. *Robotics and Autonomous Systems* 149 (2022), 103945. https://doi.org/10.1016/j.robot.2021.103945

[61] Wenxi Yue, Hongen Liao, Yong Xia, Vincent Lam, Jiebo Luo, and Zhiyong Wang. 2023. Cascade Multi-Level Transformer Network for Surgical Workflow Analysis. *IEEE Transactions on Medical Imaging* 42, 10 (Oct 2023), 2817–2831. https://doi.org/10.1109/TMI.2023.3265354

[62] Wenxi Yue, Jing Zhang, Kun Hu, Yong Xia, Jiebo Luo, and Zhiyong Wang. 2024. SurgicalSAM: Efficient Class Promptable Surgical Instrument Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 7 (Mar. 2024), 6890–6898. https://doi.org/10.1609/aaai.v38i7.28514

[63] Zixu Zhao, Yueming Jin, Junming Chen, Bo Lu, Chi-Fai Ng, Yun-Hui Liu, Qi Dou, and Pheng-Ann Heng. 2021. Anchor-guided online meta adaptation for fast one-Shot instrument segmentation from robotic surgical videos. *Medical Image Analysis* 74 (2021), 102240. https://doi.org/10.1016/j.media.2021.102240

[64] Zixu Zhao, Yueming Jin, Xiaojie Gao, Qi Dou, and Pheng-Ann Heng. 2020. Learning Motion Flows for Semi-supervised Instrument Segmentation from Robotic Surgical Video. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz (Eds.). Springer International Publishing, Cham, 679–689. https://doi.org/10.1007/978-3-030-59716-0_65

[65] Hongbo Zhou and Qiang Cheng. 2011. O(N) implicit subspace embedding for unsupervised multi-scale image segmentation. In *Computer Vision and Pattern Recognition (CVPR)*. 2209–2215. https://doi.org/10.1109/CVPR.2011.5995606

[66] Juan Carlos Ángeles Cerón, Gilberto Ochoa Ruiz, Leonardo Chang, and Sharib Ali. 2022. Real-time instance segmentation of surgical instruments using attention and multi-scale feature fusion. *Medical Image Analysis* 81 (2022), 102569. https://doi.org/10.1016/j.media.2022.102569