Motion-Boundary-Driven Unsupervised Surgical Instrument Segmentation in Low-Quality Optical Flow

Yang Liu[1], Peiran Wu[2], Jiayu Huo[1], Gongyu Zhang[1], Zhen Yuan[1], Christos Bergeles[1], Rachel Sparks[1], Prokar Dasgupta[1], Alejandro Granados[1], and Sebastien Ourselin[1]

[1] King's College London
[2] University of Bristol
yang.9.liu@kcl.ac.uk

Abstract. Unsupervised video-based surgical instrument segmentation has the potential to accelerate the adoption of robot-assisted procedures by reducing the reliance on manual annotations. However, the generally low quality of optical flow in endoscopic footage poses a great challenge for unsupervised methods that rely heavily on motion cues. To overcome this limitation, we propose a novel approach that pinpoints motion boundaries, regions with abrupt flow changes, while selectively discarding frames with globally low-quality flow and adapting to varying motion patterns. Experiments on the EndoVis2017 VOS and EndoVis2017 Challenge datasets show that our method achieves mean Intersection-over-Union (mIoU) scores of 0.75 and 0.72, respectively, effectively alleviating the constraints imposed by suboptimal optical flow. This enables a more scalable and robust surgical instrument segmentation solution in clinical settings. The code is publicly available at https://github.com/wpr1018001/Rethinking-Low-quality-Optical-Flow.git.

Keywords: Surgical instrument segmentation · Unsupervised learning · Low-quality optical flow · Motion boundary

1 Introduction

Instrument segmentation is a key component of robotic-assisted surgeries, providing improved guidance and supporting decision-making. It is also essential for some other AI-driven surgical tasks, such as workflow recognition [11,10,17], action identification [15], and tracking [2].

Although deep learning has made significant progress in fully- [23,14,13,18,6] and semi-supervised [21,22] segmentation, these approaches rely heavily on manual annotations, which increases the workload and limits their application. In contrast, we focus on unsupervised segmentation to enhance surgery understanding without any additional manual effort. Early unsupervised approaches in surgical tasks mostly focuesd on workflow analysis [1] and motion prediction [3], while instrument segmentation has received less attention. Liu et. al. [9] used
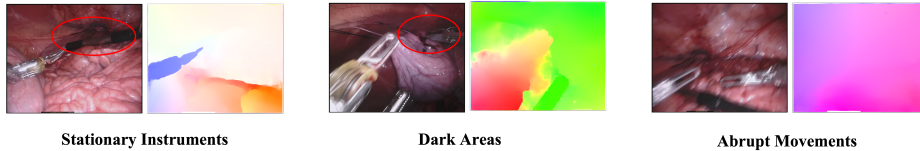
Stationary Instruments   **Dark Areas**   **Abrupt Movements**

Fig. 1. Example of some low-quality optical flow frames, including stationary instruments, dark areas and abrupt movements, which greatly limit the model performance.

unsupervised techniques and basic handcrafted cues for segmentation. However, their approach's reliance on specific signals limits its adaptability across different surgeries.

Motion plays an essential role in visual processing and is widely utilised in unsupervised models like RAFT [19] to effectively capture movement in videos. Sestini et. al. [16] introduced FUN-SIS, which integrates realistic instrument segmentation masks (referred to as *shape-priors*) from various datasets with optical flow information. Despite its novelty, FUN-SIS's dependency on the *shape-priors* and its non-end-to-end nature limited its applicability. Recently, Lian et. al. [7] brought forward RCF, a fully unsupervised model for video object segmentation (VOS), demonstrating the critical role of motion. The success of RCF highlights the importance of optical flow quality in unsupervised learning, establishing a new benchmark for instrument segmentation without manual annotations.

Unlike the high-quality optical flow found in natural videos, surgical footage often suffers from low flow quality due to the factors such as dark areas, abrupt movements, and stationary instruments, as shown in Fig. 1. Following the approach of the RCF model [7], we aim to reduce the impact of low-quality optical flow on our model. Liao et al. [8] highlighted the importance of accurately segmenting boundaries, a task complicated by the blurred boundaries in surgical optical flow. Inspired by Super-BPD [20], we employ angular measurements to detect these unclear boundaries in the flow, ensuring the model focuses on the most reliable region. Additionally, the frequent presence of low-quality optical flow in surgery videos presents a notable challenge. To address this, we drop the most difficult samples in each batch. Furthermore, the subtle movement of certain instruments, often not captured by optical flow, can reduce the model sensitivity. By varying frame rates beyond the standard, we enhance the detection of these less visible movements.

2 Method

We target the challenge of unsupervised surgical instrument segmentation within a video stream of T frames, denoted as $X_T = \{\boldsymbol{x}_i | i \in \{1, ..., T\}\}$, where $\boldsymbol{x}_i$ represents the image of the i-th frame. The objective is to create a mapping function f such that $f(x_i)$ produces a mask that extracts the area of the surgical instrument. It is important to note that instrument mask annotations are not employed during the training stage.
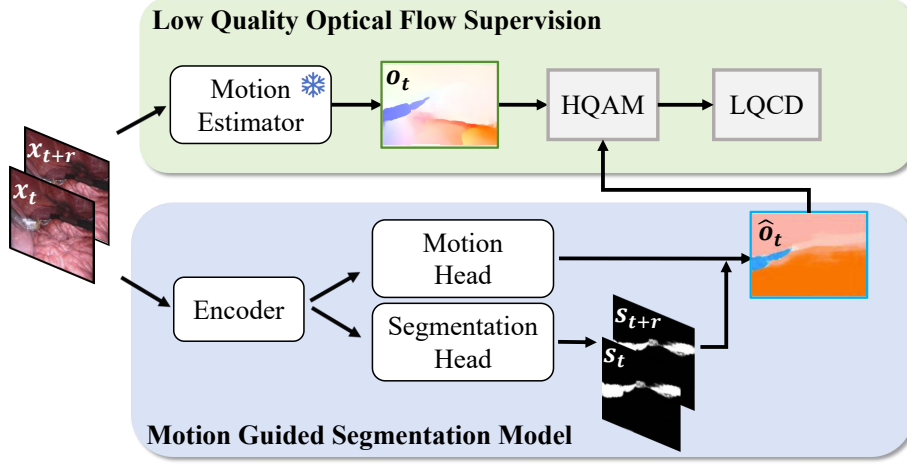
Fig. 2. Overview of our proposed unsupervised instrument segmentation framework. Two frames, separated by a random interval r, are fed into both a motion-guided segmentation model (e.g. RCF [7]) and a pre-trained Motion Estimator (e.g. RAFT [19]) that generates pseudo flow maps o_t. The proposed HQAM and LQCD modules refine these pseudo flow maps, yielding a robust supervision.

We adopt the first stage of the state-of-the-art (SOTA) RCF model [7] as our backbone, thanks to its fully end-to-end design and independence from *shape-priors*. However, the optical flow in surgical videos often suffers from dim lighting, abrupt instrument or camera movements, and partially stationary objects (Fig. 1), creating significant challenges for unsupervised segmentation. To address these issues, we introduce a **High-Quality Area Matching (HQAM)** block that emphasizes reliable flow regions, a **Low-Quality Case Dropping (LQCD)** mechanism to discard frames with severely compromised flow, and a variable frame-rate scheme to better capture subtle instrument motions. The overall architecture of our method is illustrated in Fig. 2.

2.1 High-Quality Area Matching

Accurate motion cues derived from optical flow are essential for unsupervised instrument segmentation. However, in surgical videos, these cues often fail to capture the interior of instruments, particularly in low-light or partially stationary scenarios (Fig. 1), resulting in missed motion signals. If such incomplete cues are used for training, they risk propagating errors. In contrast, regions closed to the flow changes abruptly (*motion boundaries*) typically provide more reliable signals. Building on this insight, we propose a novel High-Quality Area Matching (HQAM) module, as shown in Fig. 3, that selectively emphasizes these high-contrast motion areas. By directing supervision toward the most trustworthy regions, HQAM mitigates the impact of missed interior motion and significantly improves segmentation performance.
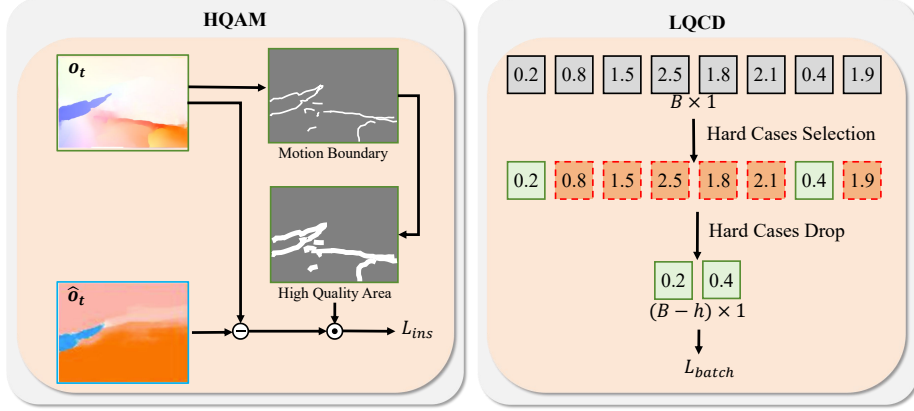
Fig. 3. Illustration of our HQAM and LQCD modules. HQAM derives a boundary-based mask from pseudo optical flows o_t, isolating reliable *high-quality* regions to guide segmentation. Meanwhile, LQCD ranks each frame in a batch by its per-frame loss and discards the top h "hard cases", removing globally low-quality motion signals.

Initially, we transform the optical flow representation $o_i \in \mathbb{R}^{H \times W \times 2}$, encoding motion vectors in two dimensions (horizontal and vertical movements), into a directional format $\theta_i \in \mathbb{R}^{H \times W}$, where H and W represent the height and width of the image frame, respectively. This conversion calculates the angle θ_i^p of the optical flow at each pixel position $p = (j, k)$. Specifically, the direction angle θ_i^p at position p is obtained through the formula:

$$\theta_i^p = \arctan 2(o_i^{j,k,x}, o_i^{j,k,y}),$$

where $o_i^{j,k,x}$ and $o_i^{j,k,y}$ are the horizontal and vertical components of the optical flow vector at the pixel p, respectively. Next, we define the directional difference at each pixel p as follows:

$$\delta_i^p = \max_{n^p \in \mathcal{N}_1^p} |\theta_i^p - \theta_i^{n^p}|,$$

where $\mathcal{N}_1^p$ signifies the 4-neighbourhood of pixel p, including the adjacent pixels in both the vertical and horizontal orientations. To demarcate boundary regions, we introduce a difference threshold $\alpha = \frac{\pi}{12}$. This leads to the creation of the boundary mask M_i, where a value of 1 is assigned to pixels that meet the condition $\delta_i^p > \alpha$, indicating the presence of a boundary, while a value of 0 denotes areas without boundaries. To enhance the supervision diversity, a simple dilation technique is employed to broaden the areas under supervision. Specifically, for every boundary pixel satisfying $\delta_i^p > \alpha$, we set $M_i(\mathcal{N}_d^p)$ to 1, where d represents the dilation distance. Consequently, we define the instance-level optical flow loss by averaging the per-pixel loss over the entire image domain Ω:

$$\mathcal{L}_{\text{ins}} = \frac{\sum_{p \in \Omega} M_i(p) \left\| o_i(p) - \hat{o}_i(p) \right\|^2}{\sum_{p \in \Omega} M_i(p)}.$$

2.2   Low-Quality Cases Drop

While HQAM addresses local flow inaccuracies, we observe that low-quality optical flow often spans the entire frame rather than being confined to specific regions. In other words, frames with degraded local flow typically exhibit inferior quality overall. To tackle this issue, we propose a frame-level drop mechanism Low-Quality Cases Drop (LQCD), as shown in the right of Fig. 3, that discards globally problematic frames. Specifically, for a batch of B training images, we identify and remove the top h frames with the highest losses (the "hard cases"), leaving us with a subset $\mathcal{S}_{\mathrm{remain}}$ of $B - h$ frames. Our final batch-level loss is then computed as:

$$\mathcal{L}_{\mathrm{batch}} = \frac{1}{|\mathcal{S}_{\mathrm{remain}}|} \sum_{x_i \in \mathcal{S}_{\mathrm{remain}}} \mathcal{L}_{\mathrm{ins}}(x_i).$$

By removing globally low-quality flow cases, we strengthen the reliability of frame-level supervision and reduce error propagation.

2.3   Variable Frame Rates Training Input

Despite establishing reliable supervision mechanisms, instruments that frequently remain stationary pose a challenge to effective training due to the scarcity of meaningful optical flow supervision samples in these regions. To overcome this limitation, we introduce a strategy of feeding training images with variable frame rates, (x_i, x_{i+r}), where the interval between adjacent frames, r, is a random number within the range of 1 to 3, instead of being fixed at $r = 1$. This variability ensures that the optical flow can capture the motion of instruments that typically exhibit little or no movement, thereby facilitating more effective training.

3   Experiments

3.1   Datasets and Evaluation Metrics

We implemented our experiments on datasets from the MICCAI EndoVis 2017 Robotic Instrument Segmentation Challenge. This dataset features 8 sets, each with 225 frames of stereo camera footage recorded by the da Vinci Xi surgical robot during various pig surgery procedures. Each frame has been meticulously annotated by experts, with identifed distinct parts of robotic surgical instruments. These instruments have been further categorised into rigid axes, articulated wrists, grommets, and a miscellaneous category for additional instruments like laparoscopic tools or drop-in ultrasound probes. We explored two variations of the EndoVis 2017 dataset for our study: EndoVis 2017 VOS and EndoVis 2017 Challenge. The VOS version includes laparoscopic tools and ultrasound probes as per Sestini et al. [16], while the Challenge version adheres to the original competition guidelines by excluding the 'other label' category. In the evaluation, we followed the same conventions and the same data partitioning and performed a 4-fold cross-validation on both the VOS and the challenge

Table 1. Comparison with SOTA supervised methods and unsupervised methods. Mean IoU (%) is reported. And the results of prior works are quoted.

Annotation	*Shape-priors*	Stage	Method	EndoVis 2017VOS	EndoVis 2017Challenge
100%	No	1	TernausNet [18]	89.06	83.60
100%	No	1	MF-TAPNet [6]	89.61	87.56
0%	Yes	1	FUN-SIS(Stage1)	40.08	37.03
0%	Yes	2	FUN-SIS(Stage2)	74.78	68.31
0%	Yes	3	FUN-SIS(Stage3)[16]	83.77	76.25
0%	No	1	AGSD [9]	71.47	67.85
0%	No	1	RCF(Stage1)	46.09	46.17
0%	No	2	RCF(Stage2)[7]	49.18	49.19
0%	No	1	RCF(Stage1) + **Ours**	**75.07**	**72.07**

versions of the dataset [6]. Our main evaluation metric is the mean Intersection-over-Union (mIoU), consistent with the benchmarks set in the MICCAI EndoVis 2017 Challenge and subsequent research in the field.

3.2   Implementation Details

We chose the RCF framework [7] as the backbone for our model, leveraging a ResNet50 [5] for feature extraction, which feeded into both a segmentation and a residual prediction head. The optical flow was derived using the RAFT model, pretrained on synthetic datasets - FlyingChairs [4] and FlyingThings [12], without any human-annotated data. Our implementation was based on PyTorch and runs on a single NVIDIA A100-PCIE-40GB GPU. The batch size was set to 8, with the $h = 6$ hard cases to ensure frame-level supervision quality. For boundary dilation, a kernel size of $d = 7$ was used.

3.3   Comparison with State-of-the-art Methods

Our findings on the EndoVis 2017 VOS and EndoVis 2017 Challenge datasets are presented in Table 1. We compared our method with the AGSD [9] and FUN-SIS [16] methods, which were both tailored to surgical video segmentation, and RCF [7], that was originally designed to segment natural images. Our method advanced the RCF model, delivering an end-to-end solution without relying on *shape-priors* for annotations. The enhancements we had implemented lead to substantial performance enhancements, showing a 28.98 percentage points (pp) and 25.93 pp increase over the original RCF model on the EndoVis 2017 VOS and Challenge datasets, respectively. This demonstrated our approach's capability to refine the use of low-quality optical flow. Moreover, when measured against AGSD, our fully unsupervised method presented gains of 3.6 pp and 4.22 pp for the respective datasets, marking the first instance of a purely unsupervised model using optical flow outperforming those dependent on crafted pseudo-labels. Our
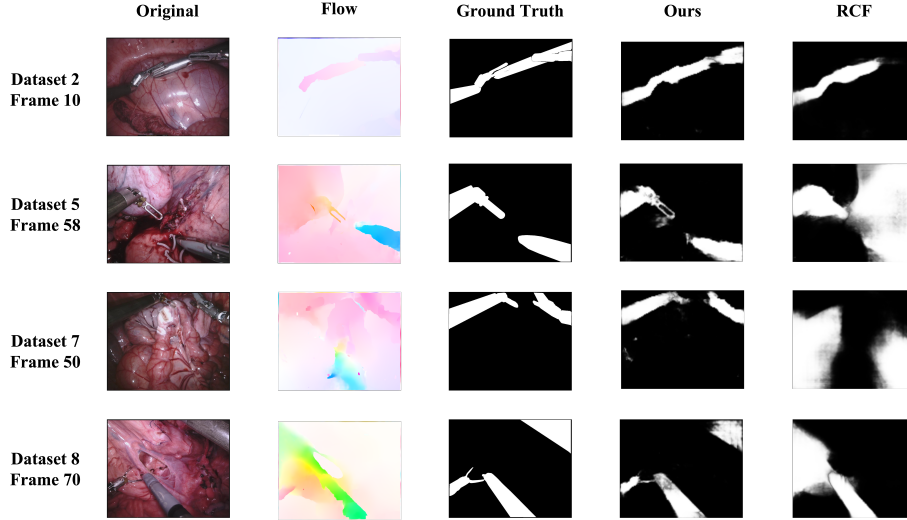
Fig. 4. Qualitative comparisons with the baseline model RCF, showing (a) optical flow pseudo-labels obtained by RAFT prediction (b) Ground Truth from EndoVis 2017, offering (c) Prediction masks of our method (d) Prediction masks of RCF.

model also showed an edge over the FUN-SIS's second stage, surpassing it by 0.29 pp and 3.76 pp on the VOS and Challenge versions, respectively.

Fig. 4 showcases comparative visual examples between our method and the baseline RCF model. Although RCF struggles with the typically low optical flow in surgical videos, our model excels, showing impressive results even in areas with challenging flow conditions. Particularly in detailed areas, such as the instrument head at case frame 58, our approach not only matches but also surpasses the ground truth.

3.4   Ablation Study

To ensure fairness and generalisability, all our experiments employ a 4-fold cross-validation on the EndoVis 2017 VOS dataset. We conducted extensive ablation studies to evaluate the impact of our proposed three key components: HQAM, LQCD and variable frame rates training input (Variable).

Different Components of the Method. Table 2 presents the ablation study results of our proposed approach. Starting with the baseline model (first row), we observe a modest performance of 46.09% mIoU. Introducing LQCD alone raises the mIoU to 47.15%, confirming that discarding frames with severely degraded optical flow helps reduce noisy supervision. When we further integrate HQAM, the performance jumps significantly to 74.47% mIoU. This demonstrates the importance of leveraging more reliable boundary cues for instrument segmentation in unsupervised settings. Finally, adding the variable frame-rate strategy produces a slight yet consistent improvement, achieving our best result of 75.07%

Table 2. The effects of different components in method.

Table 3. Different parameters of HQAM.

Function			Metric
LQCD	HQAM	Variable	mIoU(%)
			46.09
✓			47.15
	✓		70.66
✓	✓		74.47
✓	✓	✓	75.07

Parameters		Metric
α	d	mIoU(%)
	7	68.34
$\frac{\pi}{12}$	3	66.04
	1	67.81
	7	68.74
$\frac{\pi}{6}$	3	66.03
	1	66.21
	7	**70.66**
$\frac{\pi}{3}$	3	67.03
	1	64.74

mIoU, which indicates that adapting to different temporal intervals provides complementary gains. Overall, these findings highlight the effectiveness of each component and underscore their synergistic contributions to improving unsupervised instrument segmentation.

Parameter Analysis of HQAM. We investigate the impact of two key HQAM hyperparameters: the angle threshold $\alpha \in \left\{ \frac{\pi}{12}, \frac{\pi}{6}, \frac{\pi}{3} \right\}$ and the dilation kernel size $d \in \{1, 3, 7\}$. In these experiments, HQAM is the only enhancement used (i.e., both LQCD and variable frame rates are disabled). As shown in Table 3, the best performance (70.66%) is achieved at $\alpha = \frac{\pi}{3}$ and $d = 7$, a substantial improvement over the 46.09% baseline without HQAM. Although performance does vary across hyperparameters, an expected outcome in an *unsupervised* setting, every tested configuration with HQAM significantly surpasses the baseline, highlighting the method's overall robustness and effectiveness. In our default setting, we adopt more conservative parameters ($\alpha = \frac{\pi}{12}$, $d = 7$) for broader generalization in diverse surgical scenarios. Nevertheless, the consistent gains under different hyperparameter choices underscore HQAM's adaptability and confirm its value in unsupervised instrument segmentation.

4   Conclusion

In this study, we presented an unsupervised surgical instrument segmentation method that integrates High-Quality Area Matching (HQAM), Low-Quality Cases Dropping (LQCD), and variable frame rates to address the challenges posed by low-quality optical flow. Evaluated on the MICCAI EndoVis 2017 VOS and Challenge datasets, our approach achieves mIoU scores of 0.75 and 0.72, respectively, effectively reducing the reliance on manual annotations in clinical environments. Notably, our method features a plug-and-play design, allowing easy extension to other motion-driven tasks such as unsupervised depth estimation. While our experiments reveal certain hyperparameter sensitivities, common

in unsupervised settings, this limitation also highlights a promising direction for future research aimed at enhancing the robustness and generalizability of our framework.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bodenstedt, S., Wagner, M., Katić, D., Mietkowski, P., Mayer, B., Kenngott, H., Müller-Stich, B., Dillmann, R., Speidel, S.: Unsupervised temporal context learning using convolutional neural networks for laparoscopic workflow analysis. arXiv preprint arXiv:1702.03684 (2017)
2. Bouget, D., Allan, M., Stoyanov, D., Jannin, P.: Vision-based and marker-less surgical tool detection and tracking: a review of the literature. Medical image analysis **35**, 633–654 (2017)
3. DiPietro, R., Hager, G.D.: Unsupervised learning for surgical motion by learning to predict the future. In: International conference on medical image computing and computer-assisted intervention. pp. 281–288. Springer (2018)
4. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2758–2766 (2015)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Jin, Y., Cheng, K., Dou, Q., Heng, P.A.: Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22. pp. 440–448. Springer (2019)
7. Lian, L., Wu, Z., Yu, S.X.: Bootstrapping objectness from videos by relaxed common fate and visual grouping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14582–14591 (June 2023)
8. Liao, Y., Kang, S., Li, J., Liu, Y., Liu, Y., Dong, Z., Yang, B., Chen, X.: Mobileseed: Joint semantic segmentation and boundary detection for mobile robots. arXiv preprint arXiv:2311.12651 (2023)
9. Liu, D., Wei, Y., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Unsupervised surgical instrument segmentation via anchor generation and semantic diffusion. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 657–667. Springer (2020)
10. Liu, Y., Boels, M., Garcia-Peraza-Herrera, L.C., Vercauteren, T., Dasgupta, P., Granados, A., Ourselin, S.: Lovit: Long video transformer for surgical phase recognition. arXiv preprint arXiv:2305.08989 (2023)
11. Liu, Y., Huo, J., Peng, J., Sparks, R., Dasgupta, P., Granados, A., Ourselin, S.: Skit: a fast key information video transformer for online surgical phase recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21074–21084 (2023)

12. Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4040–4048 (2016)
13. Ni, Z.L., Bian, G.B., Xie, X.L., Hou, Z.G., Zhou, X.H., Zhou, Y.J.: 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). pp. 5735–5738. IEEE (2019)
14. Ni, Z.L., Zhou, X.H., Wang, G.A., Yue, W.Q., Li, Z., Bian, G.B., Hou, Z.G.: Surginet: Pyramid attention aggregation and class-wise self-distillation for surgical instrument segmentation. Medical Image Analysis **76**, 102310 (2022)
15. Psychogyios, D., Colleoni, E., Van Amsterdam, B., Li, C.Y., Huang, S.Y., Li, Y., Jia, F., Zou, B., Wang, G., Liu, Y., et al.: Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. arXiv preprint arXiv:2401.00496 (2023)
16. Sestini, L., Rosa, B., De Momi, E., Ferrigno, G., Padoy, N.: Fun-sis: A fully unsupervised approach for surgical instrument segmentation. Medical Image Analysis **85**, 102751 (2023)
17. Shi, X., Jin, Y., Dou, Q., Heng, P.A.: Semi-supervised learning with progressive unlabeled data excavation for label-efficient surgical workflow recognition. Medical Image Analysis **73**, 102158 (2021)
18. Shvets, A.A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.I.: Automatic instrument segmentation in robot-assisted surgery using deep learning. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA). pp. 624–628. IEEE (2018)
19. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402–419. Springer (2020)
20. Wan, J., Liu, Y., Wei, D., Bai, X., Xu, Y.: Super-bpd: Super boundary-to-pixel direction for fast image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
21. Wei, M., Budd, C., Garcia-Peraza-Herrera, L.C., Dorent, R., Shi, M., Vercauteren, T.: Segmatch: A semi-supervised learning method for surgical instrument segmentation (2023)
22. Wu, Z., Lau, C.Y., Zhou, Q., Wu, J., Wang, Y., Liu, Q., Lei, Z., Liu, H.: Surgivisor: Transformer-based semi-supervised instrument segmentation for endoscopic surgery. Biomedical Signal Processing and Control **87**, 105434 (2024)
23. Yang, L., Gu, Y., Bian, G., Liu, Y.: Tmf-net: A transformer-based multiscale fusion network for surgical instrument segmentation from endoscopic images. IEEE Transactions on Instrumentation and Measurement **72**, 1–15 (2023). `https://doi.org/10.1109/TIM.2022.3225922`