

Learning Motion Flows for Semi-supervised Instrument Segmentation from Robotic Surgical Video

Zixu Zhao¹, Yueming Jin¹✉, Xiaojie Gao¹, Qi Dou^{1,2}, and Pheng-Ann Heng^{1,3}

¹ Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Hong Kong, China
{zxxhao, ymjjin}@cse.cuhk.edu.hk

² Shun Hing Institute of Advanced Engineering, CUHK, Hong Kong, China

³ T Stone Robotics Institute, CUHK, Hong Kong, China

Abstract. Performing low hertz labeling for surgical videos at intervals can greatly releases the burden of surgeons. In this paper, we study the semi-supervised instrument segmentation from robotic surgical videos with sparse annotations. Unlike most previous methods using unlabeled frames individually, we propose a dual motion based method to wisely learn motion flows for segmentation enhancement by leveraging temporal dynamics. We firstly design a flow predictor to derive the motion for jointly propagating the frame-label pairs given the current labeled frame. Considering the fast instrument motion, we further introduce a flow compensator to estimate intermediate motion within continuous frames, with a novel cycle learning strategy. By exploiting generated data pairs, our framework can recover and even enhance temporal consistency of training sequences to benefit segmentation. We validate our framework with binary, part, and type tasks on 2017 MICCAI EndoVis Robotic Instrument Segmentation Challenge dataset. Results show that our method outperforms the state-of-the-art semi-supervised methods by a large margin, and even exceeds fully supervised training on two tasks⁴.

Keywords: Semi-supervised Segmentation · Motion Flow · Surgical Video

1 Introduction

By providing the context perceptive assistance, semantic segmentation of surgical instrument can greatly benefit robot-assisted minimally invasive surgery towards superior surgeon performance. Automatic instrument segmentation also serves as a crucial cornerstone for more downstream capabilities such as tool pose estimation [12], tracking and control [4]. Recently, convolutional neural network (CNN) has demonstrated new state-of-the-arts on surgical instrument segmentation thanks to the effective data-driven learning [7,13,21,10]. However, these

⁴ Our code is available at <https://github.com/zxxhaoeric/Semi-InstruSeg/>

methods highly rely on abundant labeled frames to achieve the full potential. It is expensive and laborious especially for high frequency robotic surgical videos, entailing the frame-by-frame pixel-wise annotation by experienced experts.

Some studies tend to utilize extra signals to generate parsing masks, such as robot kinematic model [3,16], weak annotations of object stripe and skeleton [6], and simulated surgical scene [15]. However, additional efforts are still required for other signal access or creation. Considerable effort has been devoted to utilizing the large-scale unlabeled data to improve segmentation performance for medical image analysis [24,2,23]. For example, Bai et al. [2] propose a self-training strategy for cardiac segmentation, where the supervised loss and semi-supervised loss are alternatively updated. Yu et al. [23] raise an uncertainty-aware mean-teacher framework for 3D left atrium segmentation by learning the reliable knowledge from unlabeled data. In contrast, works focusing on the effective usage of unlabeled surgical video frames are limited. The standard mean teacher framework has recently been applied to the semi-supervised liver segmentation by computing the consistency loss of laparoscopic images [5]. Ross et al. [20] exploit a self-supervised strategy by using GAN-based re-colorization on individual unlabeled endoscopic frames for model pretraining.

Unfortunately, these semi-supervised methods propose to capture the information based on separate unlabeled video frames, failing to leverage the inherent temporal property of surgical sequences. Given 50% labeled frames with a labeling interval of 2, a recent approach [10] indicates that utilizing temporal consistency of surgical videos benefit semi-supervised segmentation. Optical flows are used to transfer predictions of unlabeled frames to adjacent position whose labels are borrowed to calculate semi-supervised loss. Yet this method heavily depends on accurate optical flow estimation and fails to provide trustworthy semi-supervision in model with some erroneous transformations.

In this paper, we propose a dual motion based semi-supervised framework for instrument segmentation by leveraging the self-contained sequential cues in surgical videos. Given sparsely annotated sequences, our core idea is to derive the motion flows for annotation and frame transformation that recover the temporal structure of raw videos to boost semi-supervised segmentation. Specifically, we firstly design a flow predictor to learn the motion between two frames with a video reconstruction task. We propose a joint propagation strategy to generate frame-label pairs with learned flows, alleviating the misalignment of pairing propagated labels with raw frames. Next, we design a flow compensator with a frame interpolation task to learn the intermediate motion flows. A novel unsupervised cycle learning strategy is proposed to optimize models by minimizing the discrepancy between forward predicted frames and backward cycle reconstructions. The derived motion flows further propagate intermediate frame-label pairs as the augmented data to enhance the sequential consistency. Rearranging the training sequence by replacing unlabeled raw frames with generated data pairs, our framework can greatly benefit segmentation performance. We extensively evaluate the method on surgical instrument binary, part, and type segmentation tasks on 2017 MICCAI EndoVis Challenge dataset. Our method consistently

outperforms state-of-the-art semi-supervised segmentation methods by a large margin, as well as exceeding the fully supervised training on two tasks.

2 Method

Fig. 1 illustrates our dual motion-based framework. It uses raw frames to learn dual motion flows, one for recovering original annotation distribution (top branch) and the other for compensating fast instrument motions (bottom branch). We ultimately use learned motion flows to propagate aligned frame-label pairs as a substitute for unlabeled raw frames in video sequences for segmentation training.

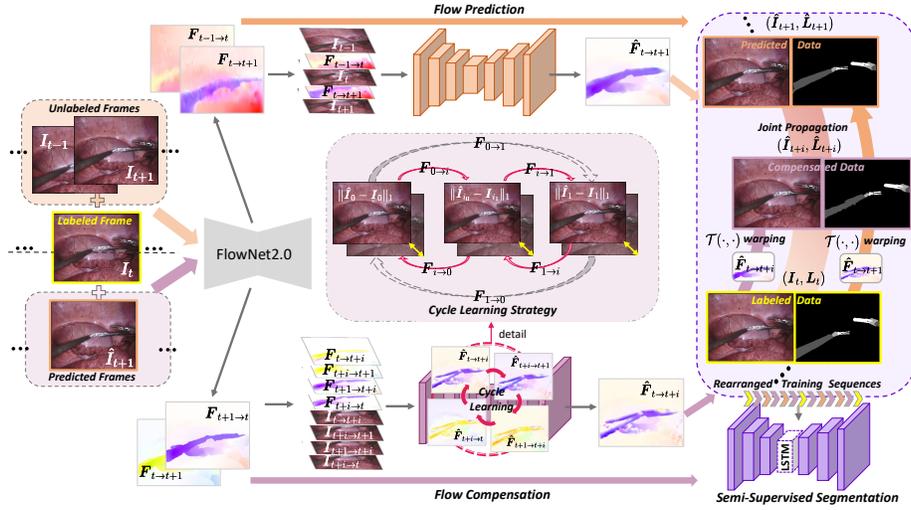


Fig. 1. The illustration of the proposed framework. We learn motion flows along *flow prediction* branch and *flow compensation* branch successively, which are used to joint propagate the aligned data pairs for semi-supervised segmentation.

2.1 Flow Prediction for Joint Propagation

With a video having T frames as $\mathbf{I} = \{I_0, I_1, \dots, I_{T-1}\}$, we assume that \mathbf{I} is labeled with intervals. For instance, only $\{I_0, I_5, I_{10}, \dots\}$ are labeled with interval 4, accounting for 20% labeled data. This setting is reasonable to clinical practice, as it is easier for surgeons to perform low hertz labeling. Sharing the spirit with [10], we argue that the motion hidden within the continuous raw frames can be applied to corresponding instrument masks. Therefore, we first derive the motion flow from raw frames with a video reconstruction task, as shown in *Flow*

Prediction branch in Fig. 1. Given the sequence $I_{t':t+1}$, we aim to estimate the motion flow $\hat{F}_{t \rightarrow t+1}$ that can translate the current frame I_t to future frame I_{t+1} :

$$\hat{F}_{t \rightarrow t+1} = \mathcal{G}(I_{t':t+1}, F_{t'+1:t+1}), \quad \hat{I}_{t+1} = \mathcal{T}(I_t, \hat{F}_{t \rightarrow t+1}), \quad (1)$$

where \mathcal{G} is a 2D CNN based flow predictor with the input $I_{t':t}$. Optical flows F_i between successive frames I_i and I_{i-1} are calculated by FlowNet2.0 [8]. \mathcal{T} is a forward warping function which is differentiable and implemented with bilinear interpolation [25]. Instead of straightforwardly relying on the optical flow [10], which suffers from the undefined problem for the dis-occluded pixels, we aim to learn the motion vector (u, v) as a precise indicator for annotation propagation which can effectively account for the gap. Intuitively, the instrument mask follows the same location shift as its frame. For an unlabeled frame I_{t+1} , we can borrow the adjacent annotation L_t and use the derived flow for its label propagation:

$$\hat{L}_{t+1} = \mathcal{T}(L_t, \hat{F}_{t \rightarrow t+1}). \quad (2)$$

Directly pairing the propagated label with the original future frame (I_{t+1}, \hat{L}_{t+1}) for our semi-supervised segmentation may encounter the mis-alignment issue in the region whose estimated motion flows are inaccurate. Motivated by [26], we introduce the concept of joint propagation into our semi-supervised setting. We pair the propagated label with predicted future frame $(\hat{I}_{t+1}, \hat{L}_{t+1})$, while leaving the original data I_{t+1} merely for motion flow generation. Such joint propagation avoids introducing the erroneous regularization towards network training. Furthermore, we can bi-directionally apply the derived motion flow with multiple steps, obtaining $(\hat{I}_{t-k:t+k}, \hat{L}_{t-k:t+k})$ with k steps ($k = 1, 2, 4$ in our experiments). The superior advantage of joint propagation can be better demonstrated when performing such multi-step propagation in a video with severely sparse annotations, as it alleviates accumulating errors within derived motion flows.

Supervised Loss Functions. The overall loss function of flow predictor is:

$$\mathcal{L}_{Pred} = \lambda_1 \mathcal{L}_1 + \lambda_p \mathcal{L}_p + \lambda_s \mathcal{L}_s, \quad (3)$$

consisting of the primary loss, i.e., L1 loss $\mathcal{L}_1 = \|\hat{I}_{t+1} - I_{t+1}\|_1$, which can capture subtle modification rather than L2 loss [17]; perceptual loss \mathcal{L}_p to retain structural details in predictions, detailed definition in [11]; smooth loss $\mathcal{L}_s = \|\nabla F_{t \rightarrow t+1}\|_1$ to encourage neighboring pixels to have similar flow values [9]. We empirically set the weights as $\lambda_1 = 0.7$, $\lambda_p = 0.2$, and $\lambda_s = 0.1$ for a more robust combination in eliminating the artifacts and occlusions than [17].

2.2 Flow Compensation via Unsupervised Cycle Learning

The fast instrument motions between successive frames always occur even in a high frequency surgical video. Smoothing large motions thus improves the sequential consistency, as well as adds data variety for semi-supervised segmentation. In this scenario, we try to compensate motion flows with a frame interpolation task. However, existing interpolation approaches are not suitable for

our case. Either the optical flow based methods [18,9] rely on consistent motion frequency, or the kernel based method [14] contradicts the alignment prerequisite. Hence, we propose an unsupervised flow compensator with a novel cycle learning strategy, which forces the model to learn intermediate flows by minimizing the discrepancy between forward predicted frames and their backward cycle reconstructions.

Given two continuous frames I_0 and I_1 , our ultimate goal is to learn a motion flow $\hat{F}_{0 \rightarrow i}$ with a time $i \in (0, 1)$ to jointly propagate the intermediate frame-label pair (\hat{I}_i, \hat{L}_i) . In the *Flow Compensation* branch, we first use the pretrained FlowNet2.0 to compute bi-directional optical flows $(F_{0 \rightarrow 1}, F_{1 \rightarrow 0})$ between two frames. We then use them to approximate the intermediate optical flows \mathbf{F} :

$$\begin{aligned} F_{i \rightarrow 0} &= -(1-i)F_{0 \rightarrow 1} + i^2 F_{1 \rightarrow 0}, & F_{1 \rightarrow i} &= F_{1 \rightarrow 0} - F_{i \rightarrow 0}, \\ F_{i \rightarrow 1} &= (1-i)^2 F_{0 \rightarrow 1} - i(1-i)F_{1 \rightarrow 0}, & F_{0 \rightarrow i} &= F_{0 \rightarrow 1} - F_{i \rightarrow 1}. \end{aligned} \quad (4)$$

Such approximation suits well in smooth regions but poorly around boundaries, however, it can still serve as an essential initialization for subsequent cycle learning. The approximated flows \mathbf{F} are used to generate warped frames $\hat{\mathbf{I}}$, including forward predicted frames \hat{I}_{i_0}, \hat{I}_1 and backward reconstructed frames \hat{I}_{i_1}, \hat{I}_0 :

$$\hat{I}_{i_0} = \mathcal{T}(I_0, F_{0 \rightarrow i}), \hat{I}_1 = \mathcal{T}(\hat{I}_{i_0}, F_{i \rightarrow 1}), \hat{I}_{i_1} = \mathcal{T}(\hat{I}_1, F_{1 \rightarrow i}), \hat{I}_0 = \mathcal{T}(\hat{I}_{i_1}, F_{i \rightarrow 0}). \quad (5)$$

We then establish a flow compensator that based on a 5-stage U-Net to refine motion flows with cycle consistency. It takes the two frames (I_0, I_1) , four initial approximations \mathbf{F} , and four warped frames $\hat{\mathbf{I}}$ as input, and outputs four refined flows $\hat{\mathbf{F}}$, where $\hat{F}_{0 \rightarrow i}$ is applied on I_0 for joint frame-label pair generation.

Unsupervised Cycle Loss. The key idea is to learn the motion flow that can encourage models to satisfy cycle consistency in time domain. Intuitively, we expect that the predicted \hat{I}_1 and reconstructed \hat{I}_0 are well overlapped with the original raw data I_1 and I_0 . Meanwhile, two intermediate frames warped along a cycle, i.e., \hat{I}_{i_1} and \hat{I}_{i_0} , should show the similar representations. Keeping this in mind, we use L1 loss to primarily constrain the inconsistency of each pair:

$$\mathcal{L}_1^c = \lambda_0 \|\hat{I}_0 - I_0\|_1 + \lambda_i \|\hat{I}_{i_0} - \hat{I}_{i_1}\|_1 + \lambda_1 \|\hat{I}_1 - I_1\|_1. \quad (6)$$

To generate sharper predictions, we add the perceptual loss \mathcal{L}_p^c on the three pairs (perceptual loss definition in [11]). Our overall unsupervised cycle loss is defined as $\mathcal{L}_{cycle} = \mathcal{L}_1^c + \lambda_p \mathcal{L}_p^c$, where we empirically set $\lambda_0 = 1.0$, $\lambda_i = 0.8$, $\lambda_1 = 2.0$, and $\lambda_p = 0.01$. Our cycle regularization can avoid relying on the immediate frames and learn the motion flow in a completely unsupervised way.

2.3 Semi-supervised Segmentation

For semi-supervised segmentation, we study the sparsely annotated video sequences $\mathbf{I} = \{I_0, I_1, \dots, I_{T-1}\}$ with a label interval h . The whole dataset consists of labeled subset $\mathcal{D}_L = \{(I_t, L_t)\}_{t=hn}$ with N frames and unlabeled subset $\mathcal{D}_U = \{I_t\}_{t \neq hn}$ with $M = hN$ frames. Using consecutive raw frames, our flow

predictor learns motion flows with a video reconstruction task, which are used to transfer the adjacent annotations for the unlabeled data. With the merit of joint propagation, we pair the generated labels and frames, obtaining the re-labeled set $\mathcal{D}_R = \{\hat{I}_t, \hat{L}_t\}_{t \neq hn}$ with M frames. Subsequently, our flow compensator learns the intermediate motion flow with an unsupervised video interpolation task. We can then extend the dataset by adding $\mathcal{D}_C = \{\tilde{I}_{t_0}, \tilde{L}_{t_0}\}_{t=1}^{T-1}$ with $N+M-1$ compensated frames with interpolation rate as 1. Our flow predictor and compensator are designed based on U-Net, with network details in supplementary. We finally consider $\mathcal{D}_L \cup \mathcal{D}_R \cup \mathcal{D}_C$ as the training set for semi-supervised segmentation. For the network architecture, we basically adopt the same backbone as [10], i.e., U-Net11 [19] with pretrained encoders from VGG11 [22]. Excitingly, different from other semi-supervised methods, our motion flow based strategy retains and even enhances the inherent sequential consistency. Therefore, we can still exploit temporal units, such as adding convolutional long short term memory layer (ConvLSTM) at the bottleneck, to increase segmentation performance.

3 Experiments

Dataset and Evaluation Metrics. We validate our method on the public dataset of Robotic Instrument Segmentation from 2017 MICCAI EndoVis Challenge [1]. The video sequences with a high resolution of 1280×1024 are acquired from *da Vinci Xi* surgical system during different porcine procedures. We conduct all three sub-tasks of this dataset, i.e., binary (2 classes), part (4 classes) and type (8 classes), with gradually fine-grained segmentation for an instrument. For direct and fair comparison, we follow the same evaluation manner in [10], by using the released 8×225 -frame videos for 4-fold cross-validation, also with the same fold splits. Two metrics are adopted to quantitatively evaluate our method, including mean intersection-over-union (IoU) and Dice coefficient (Dice).

Implementation Details. The framework is implemented in Pytorch with NVIDIA Titan Xp GPUs. The parameters of pretrained FlowNet2.0 are frozen while training the overall framework with Adam optimizer. The learning rate is set as $1e-3$ and divided by 10 every 150 epochs. We randomly crop 448×448 sub-images as the framework input. For training segmentation models, we follow the rules in [10]. As for the ConvLSTM based variant, the length of input sequence is 5. The initial learning rate is set as $1e-4$ for ConvLSTM layer while $1e-5$ for other network components. All the experiments are repeated 5 times to account for the stochastic nature of DNN training.

Comparison with Other Semi-supervised Methods. We implement several state-of-the-art semi-supervised segmentation methods for comparison, including ASC [14] (interpolating labels with adaptive separable convolution), MF-TAPNet [10] (propagating labels with optical flows), self-training method [2], Recolor [20] (GAN-based re-colorization for model initialization), and UA-MT [23] (uncertainty-aware mean teacher). We conduct experiments under the setting of 20% frames being labeled with annotation interval as 4. Most above methods are difficult to gain profit from temporal units except ASC, due to the uncontinuous

Table 1. Comparison of instrument segmentation results on three tasks (mean±std).

Methods	Frames used		Binary segmentation		Part segmentation		Type segmentation	
	Label	Unlabel	IoU (%)	Dice (%)	IoU (%)	Dice (%)	IoU (%)	Dice (%)
U-Net11	100%	0	82.55±12.51	89.76±9.10	64.87±14.46	76.08±13.05	36.83±26.36	45.48±28.16
U-Net11*	100%	0	83.17±12.01	90.22±8.83	64.96±14.12	76.57±12.44	40.31±24.38	49.57±25.39
TernausNet [21]	100%	0	83.60±15.83	90.01±12.50	65.50±17.22	75.97±16.21	33.78±19.16	44.95±22.89
MF-TAPNet [10]	100%	0	87.56±16.24	93.37±12.93	67.92±16.50	77.05±16.17	36.62±22.78	48.01±25.64
ASC [14]	20%	80%	78.51±13.40	87.17±9.88	59.07±14.76	70.92±13.97	30.19±17.65	41.70±20.62
ASC*	20%	80%	78.33±12.67	87.04±12.85	58.93±14.61	70.76±13.40	30.60±16.55	41.88±22.24
Self-training [2]	20%	80%	79.32±12.11	87.62±9.46	59.30±15.70	71.04±14.04	31.00±25.12	42.11±24.52
Re-color [20]	20%	80%	79.85±13.55	87.78±10.10	59.67±15.14	71.51±15.13	30.72±25.66	41.47±25.30
MF-TAPNet [10]	20%	80%	80.06±13.26	87.96±9.57	59.62±16.01	71.57±15.90	31.55±18.72	42.35±22.41
UA-MT [23]	20%	80%	80.68±12.63	88.20±9.61	60.11±14.49	72.18±13.78	32.42±21.74	43.61±26.30
Our Dual MF	20%	80%	83.42±12.73	90.34±9.25	61.77±14.19	73.22±13.25	37.06±25.03	46.55±27.10
Our Dual MF*	20%	80%	84.05±13.27	91.13±9.31	62.51±13.32	74.06±13.08	43.71±25.01	52.80±26.16
U-Net11	30%	0	80.16±13.69	88.14±10.14	61.75±14.40	72.44±13.41	31.96±27.98	38.52±31.02
Our Single MF	30%	70%	83.70±12.47	90.46±8.95	63.02±14.80	74.49±13.76	39.38±25.54	48.49±26.92
Our Dual MF	30%	70%	84.12±13.18	90.77±9.45	63.82±15.63	74.74±13.84	39.61±26.45	48.80±27.67
Our Dual MF*	30%	70%	84.62±13.54	91.63±9.13	64.89±13.26	76.33±12.61	45.83±21.96	56.11±22.33
U-Net11	20%	0	76.75±14.69	85.75±11.36	58.50±14.65	70.70±13.95	23.53±24.84	26.74±27.17
Our Single MF	20%	80%	83.10±12.18	90.15±8.83	61.20±14.10	72.49±12.94	36.72±23.62	46.50±25.09
Our Dual MF	20%	80%	83.42±12.73	90.34±9.25	61.77±14.19	73.22±13.25	37.06±25.03	46.55±27.19
Our Dual MF*	20%	80%	84.05±13.27	91.13±9.31	62.51±13.32	74.06±13.08	43.71±25.01	52.80±26.16
U-Net11	10%	0	75.93±15.03	85.09±11.77	55.24±15.27	67.78±14.97	15.87±16.97	19.30±19.99
Our Single MF	10%	90%	82.05±14.35	89.23±10.65	57.91±14.51	69.28±14.79	30.24±21.33	40.12±24.21
Our Dual MF	10%	90%	82.70±13.21	89.74±9.56	58.29±15.60	69.54±15.23	31.28±19.53	41.01±21.91
Our Dual MF*	10%	90%	83.10±12.45	90.02±8.80	59.36±14.38	70.20±13.96	33.64±20.19	43.20±22.70

Note: * denotes that the temporal unit ConvLSTM is added at the bottleneck of the segmentation network.

labeled input or network design. We use the same network backbone (U-Net11) among these methods for fair comparison. Table 1 compares our segmentation results with other semi-supervised methods. We also report fully supervised results of U-Net11 as upper bound, as well as two benchmarks TernausNet [21], and MF-TAPNet for reference. Among the semi-supervised methods, UA-MT achieves slightly better performance as it draws out more reliable information from unlabeled data. Notably, our method consistently outperforms UA-MT across three tasks by a large margin, i.e., 2.68% in IoU and 2.24% in Dice on average. After adding the temporal unit, results of ASC degrade instead on two tasks due to the inaccurate interpolated labels. As our semi-supervised method can enhance sequential consistency by expanded frame-label pairs, our results can be further improved with ConvLSTM, even surpassing the fully supervised training (U-Net11*) by 0.91% Dice on binary task and 3.23% Dice on type task.

Analysis of Our Semi-supervised Methods. For 6×225 -frame training videos in each fold, we study the frames labeled at an interval of 2, 4, and 8, resulting in 30%, 20%, and 10% annotations. Table 1 also lists results of three ablation settings: (1) Our Single MF: U-Net11 trained by set $\{\mathcal{D}_L \cup \mathcal{D}_R\}$ with Flow Prediction; (2) Our Dual MF: U-Net11 trained by set $\{\mathcal{D}_L \cup \mathcal{D}_R \cup \mathcal{D}_C\}$ with Flow Prediction and Compensation; (3) Our Dual MF*: U-Net11 embedded with ConvLSTM and trained by set $\{\mathcal{D}_L \cup \mathcal{D}_R \cup \mathcal{D}_C\}$. It is observed that under all annotation ratios, compared with U-Net11 trained by labeled set \mathcal{D}_L alone, our flow based framework can progressively boost the semi-supervised performance with generated annotations. We gain the maximum benefits in the severest condition (10% labeling), where our Single MF has already largely improved the segmentation by

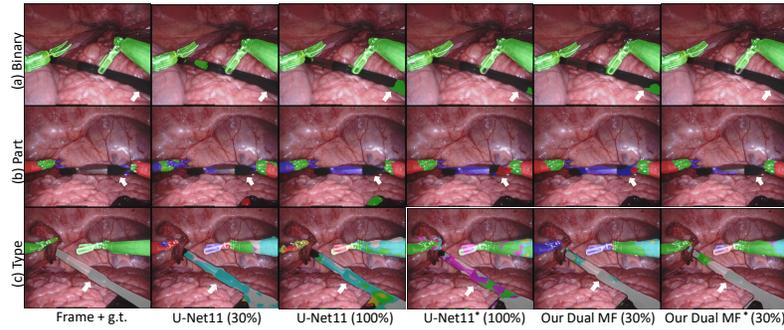


Fig. 2. Visualization of instrument (a) binary, (b) part, and (c) type segmentation. From left to right, we present frame with ground-truth, results of fully supervised training and our semi-supervised methods. * denotes incorporating ConvLSTM units.

6.12% IoU and 4.14% Dice (binary), 2.67% IoU and 1.50% Dice (part), 14.37% IoU and 20.82% Dice (type). Leveraging compensated pairs, our Dual MF with 30% and 20% labels is even able to exceed the full annotation training by 1-3% IoU or Dice, corroborating that our method can recover and adjust the motion distribution for better network training. It can be further verified using temporal units. We only see slight improvements in fully supervised setting (the first two rows) because some motion inconsistency existed in original videos decreases the model learning capability of temporal cues. Excitingly, the increment is obvious between our Dual MF and Dual MF*, especially for the toughest type segmentation. For instance, IoU and Dice can be boosted by 6.65% and 6.25% in 20% labeling case. Fig. 2 shows some visual results. Our Dual MF* can largely suppress misclassified regions in Ultrasound probes for binary and part tasks, and

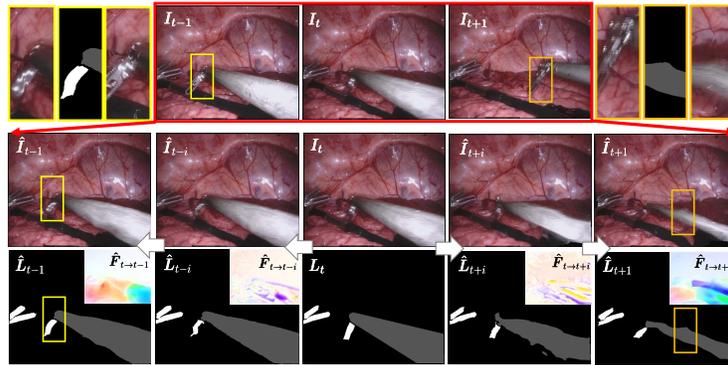


Fig. 3. Example of rearranged training sequence with propagation step $k = 1$.

achieve more complete and consistent type segmentation. It is even better at distinguishing hard mimics between instruments than fully supervised U-Net11*.

Analysis of Frame-Label Pairs. Our joint propagation can alleviate the misalignment issue from label propagation. In Fig. 3, labels in certain regions, like jaw (yellow) and shaft (orange) of instruments, fail to align with the original frames (first row) due to imprecision in learned flows, but correspond well with propagated frames (second row) as they experience the same transformation. The good alignment is crucial for segmentation. Besides, our learned flows can propagate instruments to a more reasonable position with smooth motion shift. The fast instrument motion is slowed down from I_t to \hat{I}_{t+1} with smoother movement of Prograsp Forceps (orange), greatly benefiting ConvLSTM training.

4 Conclusions

We propose a flow prediction and compensation framework for semi-supervised instrument segmentation. Interestingly, we study the sparsely annotated surgical videos from the fresh perspective of learning the motion flow. Large performance gain over state-of-the-art semi-supervised methods demonstrates the effectiveness of our framework. Inherently our method can recover the temporal structure of raw videos and be applied to surgical videos with high motion inconsistency.

Acknowledgments. This work was supported by Key-Area Research and Development Program of Guangdong Province, China (2020B010165004), Hong Kong RGC TRS Project No.T42-409/18-R, National Natural Science Foundation of China with Project No. U1813204, and CUHK Shun Hing Institute of Advanced Engineering (project MMT-p5-20).

References

1. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
2. Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D.: Semi-supervised learning for network-based cardiac mr image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 253–260. Springer (2017)
3. da Costa Rocha, C., Padoy, N., Rosa, B.: Self-supervised surgical tool segmentation using kinematic information. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 8720–8726. IEEE (2019)
4. Du, X., Allan, M., Bodenstedt, S., Maier-Hein, L., Speidel, S., Dore, A., Stoyanov, D.: Patch-based adaptive weighting with segmentation and scale (pawss) for visual tracking in surgical video. *Medical image analysis* **57**, 120–135 (2019)
5. Fu, Y., Robu, M.R., Koo, B., Schneider, C., van Laarhoven, S., Stoyanov, D., Davidson, B., Clarkson, M.J., Hu, Y.: More unlabelled data or label more data? a

- study on semi-supervised laparoscopic image segmentation. In: *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*, pp. 173–180. Springer (2019)
6. Fuentes-Hurtado, F., Kadkhodamohammadi, A., Flouty, E., Barbarisi, S., Luengo, I., Stoyanov, D.: Easylabels: weak labels for scene segmentation in laparoscopic videos. *International journal of computer assisted radiology and surgery* **14**(7), 1247–1257 (2019)
 7. García-Peraza-Herrera, L.C., Li, W., Fidon, L., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Vander Poorten, E., Stoyanov, D., Vercauteren, T., et al.: Toolnet: holistically-nested real-time segmentation of robotic surgical tools. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 5717–5722. IEEE (2017)
 8. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: FlowNet 2.0: Evolution of optical flow estimation with deep networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2462–2470 (2017)
 9. Jiang, H., Sun, D., Jampani, V., Yang, M.H., Learned-Miller, E., Kautz, J.: Super sloMo: High quality estimation of multiple intermediate frames for video interpolation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 9000–9008 (2018)
 10. Jin, Y., Cheng, K., Dou, Q., Heng, P.A.: Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 440–448. Springer (2019)
 11. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: *European conference on computer vision*. pp. 694–711. Springer (2016)
 12. Kurmann, T., Neila, P.M., Du, X., Fua, P., Stoyanov, D., Wolf, S., Sznitman, R.: Simultaneous recognition and pose estimation of instruments in minimally invasive surgery. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 505–513. Springer (2017)
 13. Milletari, F., Rieke, N., Baust, M., Esposito, M., Navab, N.: CfcM: Segmentation via coarse to fine context memory. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 667–674. Springer (2018)
 14. Niklaus, S., Mai, L., Liu, F.: Video frame interpolation via adaptive separable convolution. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 261–270 (2017)
 15. Pfeiffer, M., Funke, I., Robu, M.R., Bodenstedt, S., Strenger, L., Engelhardt, S., Roß, T., Clarkson, M.J., Gurusamy, K., Davidson, B.R., et al.: Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 119–127. Springer (2019)
 16. Qin, F., Li, Y., Su, Y.H., Xu, D., Hannaford, B.: Surgical instrument segmentation for endoscopic vision with data fusion of rediction and kinematic pose. In: *2019 International Conference on Robotics and Automation (ICRA)*. pp. 9821–9827. IEEE (2019)
 17. Reda, F.A., Liu, G., Shih, K.J., Kirby, R., Barker, J., Tarjan, D., Tao, A., Catanzaro, B.: Sdc-net: Video prediction using spatially-displaced convolution. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 718–733 (2018)

18. Reda, F.A., Sun, D., Dundar, A., Shoeybi, M., Liu, G., Shih, K.J., Tao, A., Kautz, J., Catanzaro, B.: Unsupervised video interpolation using cycle consistency. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 892–900 (2019)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
20. Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., et al.: Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *International journal of computer assisted radiology and surgery* **13**(6), 925–933 (2018)
21. 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)
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
23. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 605–613. Springer (2019)
24. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 408–416. Springer (2017)
25. Zhou, T., Tulsiani, S., Sun, W., Malik, J., Efros, A.A.: View synthesis by appearance flow. In: European conference on computer vision. pp. 286–301. Springer (2016)
26. Zhu, Y., Sapra, K., Reda, F.A., Shih, K.J., Newsam, S., Tao, A., Catanzaro, B.: Improving semantic segmentation via video propagation and label relaxation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8856–8865 (2019)