From Forks to Forceps: A New Framework for Instance Segmentation of Surgical Instruments

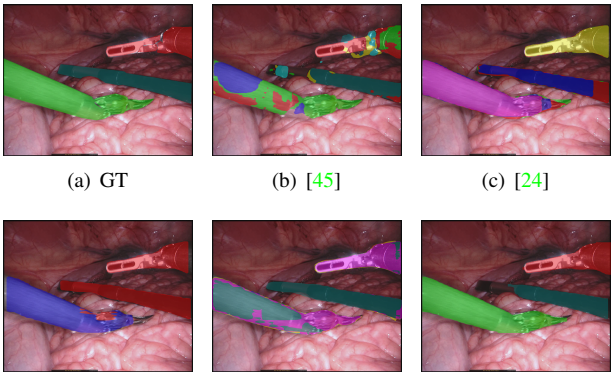
Britty Baby[1,2], Daksh Thapar[3], Mustafa Chasmai[1], Tamajit Banerjee[1], Kunal Dargan[1], Ashish Suri[2,1], Subhashis Banerjee[4,1], Chetan Arora[1]
[1]IIT Delhi, India, [2]AIIMS, New Delhi, India, [3]IIT Mandi, India, [4]Ashoka University, India

Abstract

*Minimally invasive surgeries and related applications demand surgical tool classification and segmentation at the instance level. Surgical tools are similar in appearance and are long, thin, and handled at an angle. The fine-tuning of state-of-the-art (*SOTA*) instance segmentation models trained on natural images for instrument segmentation has difficulty discriminating instrument classes. Our research demonstrates that while the bounding box and segmentation mask are often accurate, the classification head misclassifies the class label of the surgical instrument. We present a new neural network framework that adds a classification module as a new stage to existing instance segmentation models. This module specializes in improving the classification of instrument masks generated by the existing model. The module comprises multi-scale mask attention, which attends to the instrument region and masks the distracting background features. We propose training our classifier module using metric learning with* arc loss *to handle low inter-class variance of surgical instruments. We conduct exhaustive experiments on the benchmark datasets EndoVis2017 and EndoVis2018. We demonstrate that our method outperforms all (more than 18)* SOTA *methods compared with, and improves the* SOTA *performance by at least 12 points (20%) on the EndoVis2017 benchmark challenge and generalizes effectively across the datasets. Project page with source code is available at* nets-iitd.github.io/s3net.

1. Introduction

The computer vision community has significantly progressed in designing semantic and instance segmentation algorithms in recent years. One of the reasons for the success is the availability of large datasets [43, 1, 5, 14]. On the other hand, due to the advantages of small incisions and rapid recovery, minimally invasive surgeries (MIS) are increasingly accepted in various surgical specialties [53, 26].



(a) GT (b) [45] (c) [24]
(d) [21] (e) [39] (f) Ours

Figure 1. Instrument segmentation produced by various competitive methods on a sample from the EV17 dataset [3]. Each instrument class is shown in a different color. Note that ISINet [21] gets the segmentation right but classifies incorrectly. We identify instrument misclassification as the primary reason for low performance of the SOTA techniques, and propose various architecture modification for accurate classification. To illustrate the severity of the problem, substituting the predicted class label of a MaskR-CNN object with the ground truth label improved the model's AP50 score from 0.65 to 0.90.

Automated segmentation of a surgical instrument in MIS is an area of active research with high utility. The surgical instrument segmentation poses various challenges depending on the dataset acquisition source, type of surgery and instruments/ tools involved, image resolution, dataset size, tool statistics, challenging conditions (occlusions, rapid appearance changes, specular reflections, smoke, blur, blood spatter) [9].

Most surgical datasets and algorithms structure instrument segmentation as semantic segmentation, which classifies each pixel as one of the instrument class [45, 23, 28, 31, 37, 49]. Due to disconnected regions and occluded/overlapping instruments, the task of assigning an instance label to the semantic segmentation output is non-trivial. However, obtaining an instance-level mask of the manipulating instruments is essential for most surgical in-

strument segmentation applications that depend on instrument tracking [18, 34, 46, 30, 7]. Hence, we argue for formulating the task as multi-class instance segmentation.

The research in surgical instrument segmentation is largely driven by the EndoVis2017 dataset [3], which is a robotic instrument dataset containing annotations for different instrument types. The dataset contains seven instruments, all of which have a thin, long, tube-like structure. ISINet [18] fine-tunes a MaskRCNN [24] backbone for instance segmentation with a reported Challenge `IoU` score of 0.55. TraSeTR [58] uses a transformer architecture that exploits tracking cues to assist surgical instrument segmentation with a Challenge `IoU` score of 0.6.

Contribution 1: We investigated the reasons for low `IoU` scores of `SOTA` algorithms on medical instrument segmentation. We found that these methods give a reasonable output for the bounding box and segmentation mask but often misclassify the output box/mask (Fig. 1). We believe that our observation is analogous to the one reported by [50] for natural images. The authors have reported that in the dataset with a long tail, the `SOTA` techniques for object detection in natural images give correct region proposals for less frequent classes but often misclassify them. We posit that due to significant visual differences between natural objects and medical instruments, a deep neural network model that does cross-domain fine-tuning is unable to develop robust features for classification. However, since the bounding box and mask predictions are based on more robust features such as edges, these predictions generalize more easily. Therefore, there is a need for a specialized module in these techniques that focuses on acquiring the attributes necessary for efficiently classifying surgical instruments. Hence, we propose adding a dedicated classification module as a new stage in the existing techniques, which decouples the classification from the bounding box and mask prediction and specializes in classifying classes from the tail of a distribution.

Contribution 2: A deeper investigation found a variation in aspect ratio and orientation between natural images and MIS. While in natural images, the width-to-height ratio is usually around 0.5, surgical instruments are mostly two or greater. Further, natural objects appear mostly vertical in an image and fit well in rectilinear bounding boxes. On the other hand, surgical instruments are used obliquely and appear across a bounding box's diagonal. The instrument's aspect ratio and oblique appearance reduce its proportion in the bounding box area and brings in a distracting background. To make matters worse, given the small operating regions in an MIS, a proposed bounding box may contain multiple tools, further complicating the classification task. The finding motivates the need for classification based on mask-based attention rather than the existing bounding box-based one. Hence, we propose to include mask-based attention in the proposed specialized classification module.

Contribution 3: Surgical instruments show inter-class appearance similarity and contain long shafts; the only distinguishing characteristic may be the instruments' tips. Therefore, generic cross entropy-based training of classifiers in contemporary architecture is unsuitable for fine-grained classification of surgical tools. Recent literature suggests that for small datasets, it is beneficial to separate representation learning and classification stages [54]. The first can be achieved using a contrastive loss, followed by fine-tuning for the classification. We follow a similar approach and train our proposed classification module using arc loss [16], followed by fine-tuning with cross-entropy loss.

Results: We conduct exhaustive experiments on the benchmark Robot-assisted surgery datasets EndoVis2017 (`EV17`), and EndoVis2018 (`EV18`). The proposed method generalizes well on all these datasets, outperforms the instance segmentation methods with varying backbones, and achieves at least 12 points (20%) improvement over the `SOTA` on the benchmark EndoVis2017 challenge.

2. Related Work

The application of object detection, segmentation, and tracking in the field of MIS extends to various surgical branches like gynecology, ophthalmology, and neurosurgery [8, 55, 38, 40]. Researchers have contributed different datasets in this regard as well [2, 42, 22]. Many techniques have been developed using the Endovis challenge datasets. The instrument segmentation problem has been formulated using both semantic [45, 23, 28, 31, 37, 49], as well as instance segmentation [32, 21, 29, 33]. The approach for this segmentation can be supervised, semi-supervised or unsupervised methods. The semi-supervised/unsupervised methods handle the data annotation scarcity in the medical domain and explore domain adaptation of the model to surgical scenario [44, 35, 57, 36]. In this work, we are focusing on the supervised instance segmentation problem.

Semantic Segmentation of Medical Instruments: TernausNet uses U-Net architecture [41], on a pre-trained VGG11 or VGG16 backbone [45]. It shows the best performance on binary segmentation but performs poorly on the classification of the instrument type. U-NetPlus uses a modified Encoder-Decoder-based U-Net architecture and data augmentation techniques to improve performance [23]. Some methods explore real-time instrument semantic segmentation [28, 49]. PAANet aggregate multi-scale attentive features [37] and MF-TAPNet integrates flow-based temporal priors to an attention pyramid network [31]. All the methods discussed above use a single-stage approach for

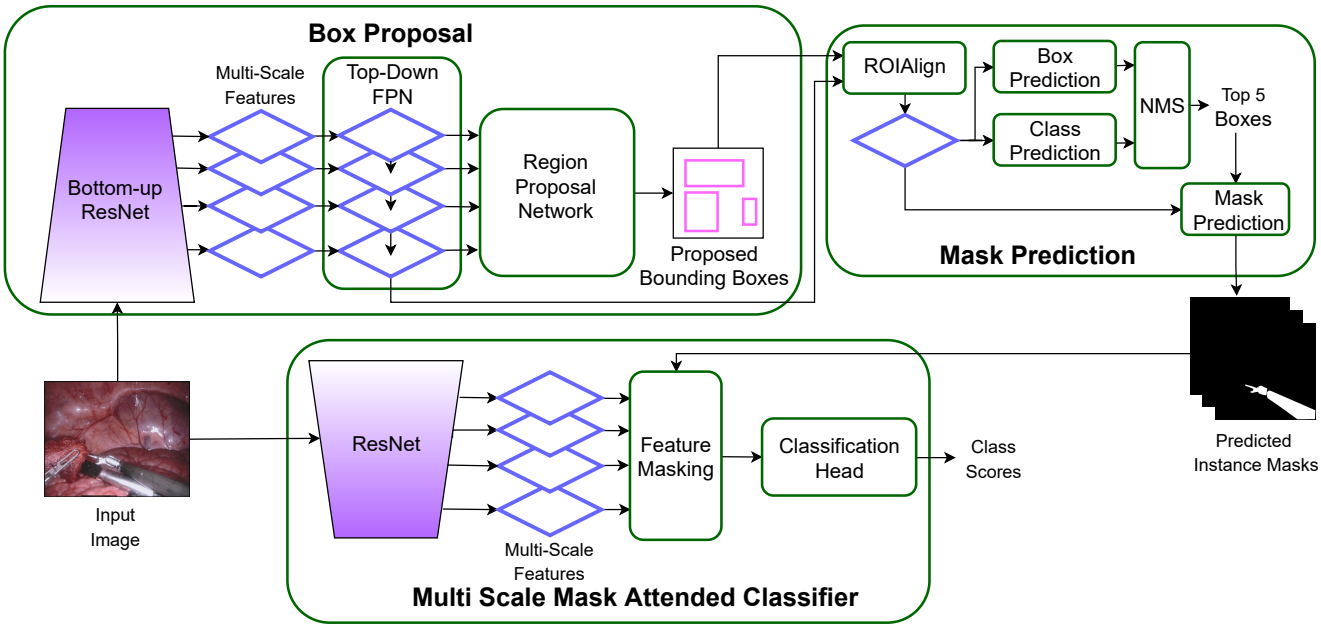


Figure 2. Architecture of the proposed 3-stage neural network model, named `S3Net`, for the instrument segmentation. Whereas the first two stages are similar to the state of the art, we introduce a third stage, named `MSMA`, specializing in classification. We make several innovations in the design of `MSMA` as described in the main text.

semantic segmentation, which often over segments an instrument to multiple classes (see [45] in Fig. 1).

Instance Segmentation of Medical Instruments: The formulation has been explored in two ways: cross-domain fine-tuning of models pre-trained on natural images [24, 32, 21, 58], and custom-designed models for the task [29, 33]. For the first category, researchers have primarily used MaskRCNN [24]. Kong et al. [32] adapted MaskRCNN by optimizing the anchor scales for instrument types. ISINet [21] has used fine-tuned MaskRCNN along with a temporal consistency module to exploit the sequential nature of the data. The improved performance in ISINet is due to the non-maximum suppression of regions across various classes and retaining only the highest predicted class for any instrument instance. Their temporal consistency module improves marginally over their instance selection heuristic. TraSeTR [58] is a transformer-based track-to-segment method that incorporates tracking cues for instance segmentation of instruments. This technique relies on the second stage's classification predictions and adds identity matching and contrastive query learning to address surgical instruments with huge temporal variations. AP-MTL [29] has proposed an Encoder-Decoder architecture for real-time instance segmentation. They have shown improvement over a domain-adapted MaskRCNN. Mask-then-classify [33] also used an encoder-decoder network, along with a classifier that uses features from the segmentation stage to classify the pixel-wise instances. The approach uses a single-stage network and is prone to classification errors if there are errors in the masks and vice-versa.

In this work, we focus on the misclassification challenges of the bounding-box-based instance segmentation methods when fine-tuned for instrument segmentation. We propose adding a novel specialized classification module to mitigate the challenges.

3. Proposed Architecture

As such, our main contribution, the specialized classification module as a new stage, can be inserted into any existing instance segmentation model. However, we have based our model on the MaskRCNN [24] backbone for validation. The MaskRCNN backbone contains two stages corresponding to the region proposal network (RPN) and a classification head generating masks and labels for each proposal. We insert our classification module as the third stage in the MaskRCNN and replace the labels generated by the second stage with the ones generated by our module. The architecture of our proposed Three Stage Deep Neural Network (`S3Net`) is shown in Fig. 2.

Notation: For a given input frame I_i, the first stage, (Box Proposal) extracts l bounding box proposals, where $B_{i,j}$ is the j^{th} proposal in the i^{th} frame. The second stage (Mask Prediction), predicts the mask $P_{i,j,\hat{c}}$ for each instrument using the bounding box proposals. Here, $\hat{c}$ refers to the class predicted at this stage.

Processing of second stage output: The first stage of MaskRCNN and other similar models is called Region Pro-

posal Network (RPN) and typically outputs many overlapping regions corresponding to a single object instance in the image. The classification head of the MaskRCNN remains weak even after cross-domain fine-tuning. Hence, many of these overlapping boxes get classified as different classes. A typical non-maximal suppression (NMS) step does not reject overlapping boxes corresponding to different classes. If not addressed, this leads to many false-positive predictions by MaskRCNN and other SOTA techniques that we have compared within our experiments. Hence, we modify our implementation's standard NMS step to reject overlapping segments in an image across the classes.

Handling misclassification: In a two-stage network, we observe that the proposals generated by the first stage RPN are inaccurate. However, these proposals get refined by the bounding box regression head in the second stage, leading to higher bounding box and mask accuracy after the second stage. However, the classification is performed on the regions cropped out of inaccurate region proposals from the first stage and remains fragile, which makes classification the bottleneck in the instrument segmentation accuracy. Based upon the insight gained from our analysis, we propose a new deep neural network paradigm that uses the first two stages from a standard instance segmentation method but contains an additional third stage specializing in classification based on the masks. We call the proposed classifier as *Multi-Scale Mask Attended Classifier (MSMA)*, which updates/corrects the class predictions from the first two stages. Let $\hat{c}$ denote the class label, and $P_{i,j,\hat{c}}$ denote the mask output corresponding to a region proposal. Then the objective of MSMA is to take the original image and mask $P_{i,j,\hat{c}}$ as the input and refine the class label from $\hat{c}$ to a more accurate label c. The final mask (with updated class label) is denoted as $P_{i,j,c}$. As described earlier, the bounding box rectangular regions for a medical instrument contain lots of background and the pixels of other instruments. This is due to the instrument shape and how instruments are typically used in a surgery. This distracts the classification head and leads to errors. Hence, instead of using rectangular region proposals, we introduce spatial mask attention in MSMA to emphasize the region belonging to the instrument only.

During training, we use ground truth masks corresponding to the instance, and while testing, we use the mask predicted by the second stage. This hard-mask attention is performed on multi-scale features of the image. This helps our model focus on the correct instrument spatial region in the image, leading to more accurate classification of the mask generated in the second stage of MaskRCNN. Further, to effectively train with a small dataset, we separate learning feature representation and classification in the proposed third stage. We first perform metric learning with arc loss function, followed by a learning classifier with categorical cross-entropy. Below we describe the proposed MSMA module.

Multi Scale Mask Attended (MSMA) Classifier: Convolutional Feature masking [15] was proposed by Dai et al. to exploit shape information to separate objects from the stuff. We adopt this in an instance segmentation framework to separate instruments from the background/ overlapping instruments and improve the classifier. We explore the decoupling property of the mask and the classification head and use a dedicated neural network with multi-scale mask attention for classification. Our proposed paradigm is shown in Fig. 2. It takes as input the original RGB image I_i and the predicted masks $P_{i,j,\hat{c}}$ of each instrument instance. A ResNet [25] backbone is used to extract multi-scale features from I_i. Then, the mask $P_{i,j,\hat{c}}$ is multiplied by each feature to create multi-scale mask-attended features. The masked features are then merged using another 1×1 convolution, creating a single feature map for each instance. Note that if multiple instances of a class are predicted in a frame, then the MSMA classifier is run separately for each instance.

We learn an embedding layer over the masked feature map, which outputs an embedding, $E_{i,j}$, for each instrument instance. Each $E_{i,j}$ is then used to classify the instrument present in the mask, giving us a new class label c for the mask. For training MSMA classifier, we utilize arc loss [16], as defined below:

$$\mathcal{L} = -\frac{1}{C}\sum_{c=1}^{c=C} \log \frac{e^{\cos(\theta_c + m)}}{e^{\cos(\theta_c + m)} + \sum_{j=1, j\neq c}^{C} e^{\cos\theta_j}}.$$

Here C is the number of classes, and m is the angular margin enforced between features of different classes. Further, θ_j is the angle formed between the Embedding feature E_j and the weight vector of the j^{th} neuron in the final fully connected layer. The arc loss is adapted from the face recognition domain to the surgical domain where the inter-class variance is low; the arc loss tries to maximize the distance between the features of the classes, thereby increasing the classification accuracy. Unlike categorical cross-entropy loss, which computes the dot product between $E_{i,j}$ and each weight vector, the arc loss only depends on the angle between them. Using arc loss removes the effect of the magnitude of the weight vector for the final decision. Since the magnitude of weight vectors is unbounded, they can easily become biased for a class with more samples. Hence, the arc loss handles class imbalance in the data by removing the dependency over the magnitude of weight vectors. Moreover, the arc loss forms a metric-based angular cluster for each class rather than learning a decision boundary between various classes. This is ensured by the angular margin m, resulting in better intra-class compactness and inter-class separability despite data scarcity.

4. Dataset and Evaluation

Benchmark Datasets: We have used Robot-assisted endoscopic surgery datasets EndoVis 2017 [3] (denoted as EV17), and EndoVis 2018 [4] (denoted as EV18) datasets for our experiments. **(1)** EV17 dataset contains ten videos from the da-Vinci robotic system and provides annotations of 6 robotic instruments and an ultrasound probe. We adopted the 4-fold cross-validation from [45] for fair comparison with 1800 frames (8×225). The fold-wise split makes it 1350 and 450 frames for training and validation, respectively. **(2)** EV18 is a robotic instrument clinical dataset that includes organs and surgical items like gauze and suturing thread and contains the instrument super category but not the instrument type. This dataset is additionally annotated for instrument types by [21] with seven robotic instrument types and 11 training videos, and four testing videos with 149 images each. They provide the annotations as image pixels but do not provide instance labels. We annotated the instances ourselves for our experiments.

Evaluation: We categorize the compared methods into two categories. EVS methods are the ones that have reported their accuracy on EV17 or EV18 datasets. NLI models are the instance segmentation methods proposed for natural images. For EV17 and EV18 datasets, we evaluate the performance on the Challenge IoU (Ch_IoU) metric as proposed in the EV17 challenge [6] and ISINet IoU (ISI_IoU) and mean class IoU (mcIoU) metrics proposed in [21].

5. Experiments and Results

5.1. Implementation Details

Backbones: The proposed MSMA module can be added to any existing instance segmentation method as an additional stage. To validate this, we added MSMA on two methods with very different architectures: a CNN-based MaskRCNN [24] and a newer transformer-based Mask2former [12]. The latter is used for validation of third stage and performs poorly compared to the former. We report most of the results using the MaskRCNN as the initial stage. We refer to both the models (based on MaskRCNN or Mask2former) as S3Net and explicitly specify the architecture type when using the transformer architecture.

Training: We first train the first two stages using a regression loss, cross-entropy classification loss, and per-pixel segmentation loss. We have used an ImageNet pre-trained ResNet-50-FPN model to match the SOTA [21] architecture and finetuned it for the instrument dataset. We resize each image to a size of $(1333, 800)$. Stochastic Gradient Descent with a learning rate of 20^{-2} is used to train the two stages simultaneously for 12 epochs.

For stage 3 (MSMA), we first use the pre-trained weights of ResNet-50-FPN from the box proposal module and freeze them initially to avoid over-fitting. The classification head of the MSMA classifier is first trained using ground truth instance masks for ten epochs using cross-entropy-loss. Then it is trained for 15 epochs using arc loss. After 25 epochs, we unfreeze the weights of ResNet also and train the complete MSMA classifier end-to-end for five epochs using arc loss. Finally, only the classification layer is trained using cross-entropy loss. For training the MSMA classifier, we have resized each image to $(224, 224)$. The ground truth masks were resized to $(56, 56)$ while masking the features to match the feature resolution of the last layer of block 3 in ResNet. The mask attended classification head of MSMA is trained using Adam optimizer with a learning rate of 10^{-5}, whereas the end-to-end training of MSMA is done using a learning rate of 10^{-7}.

Inference: During inference, we set the score threshold of 0.0 to accommodate all the classes and select only the top 5 instances because a typical frame contains approximately 3 to 4 instruments in the ground truth.

5.2. Analysis

Comparison with SOTA: We compare S3Net for instrument type segmentation with EVS methods that include semantic segmentation [45, 31] and instance segmentation approaches [21, 58]. For the NLI techniques, we use the source code provided by the authors to train the models for the mentioned datasets and use our inference parameters. We add region-based NMS as mentioned in Sec. 3 as a post-processing step on the predicted masks of an image and report the IoU scores after post-processing for all the models (see Tab. 1).

For EV17, S3Net outperforms all the NLI methods and the other evs-based instance segmentation methods. It improves over ISINet [21] by 30% Ch_IoU and 60 % mcIoU, showing that the mask-based classification using stage 3 improves the results by a considerable margin with only using the spatial information. Even though TraSeTR [58] explores transformer-based architecture with mask classification paradigm and also uses the temporal information, S3Net outperforms it by a margin of 20% on Ch_IoU and 26% on mcIoU achieving the SOTA results. For EV18 dataset, S3Net outperforms ISINet [21] by a slight margin of 3.8% on ChIoU and 5.8% on mcIoU. In comparison to TraSeTR, the results are slightly lesser, which shows that for this dataset, apart from classification, other determining factors like temporal information of the instruments favored a tracking-based method.

Validating Reasons for low accuracy of SOTA: In Tab. 1, we compare SOTA instance segmentation on EVS datasets. We have analyzed why the NLI models are less accurate. As noted before, we make three claims about SOTA's low

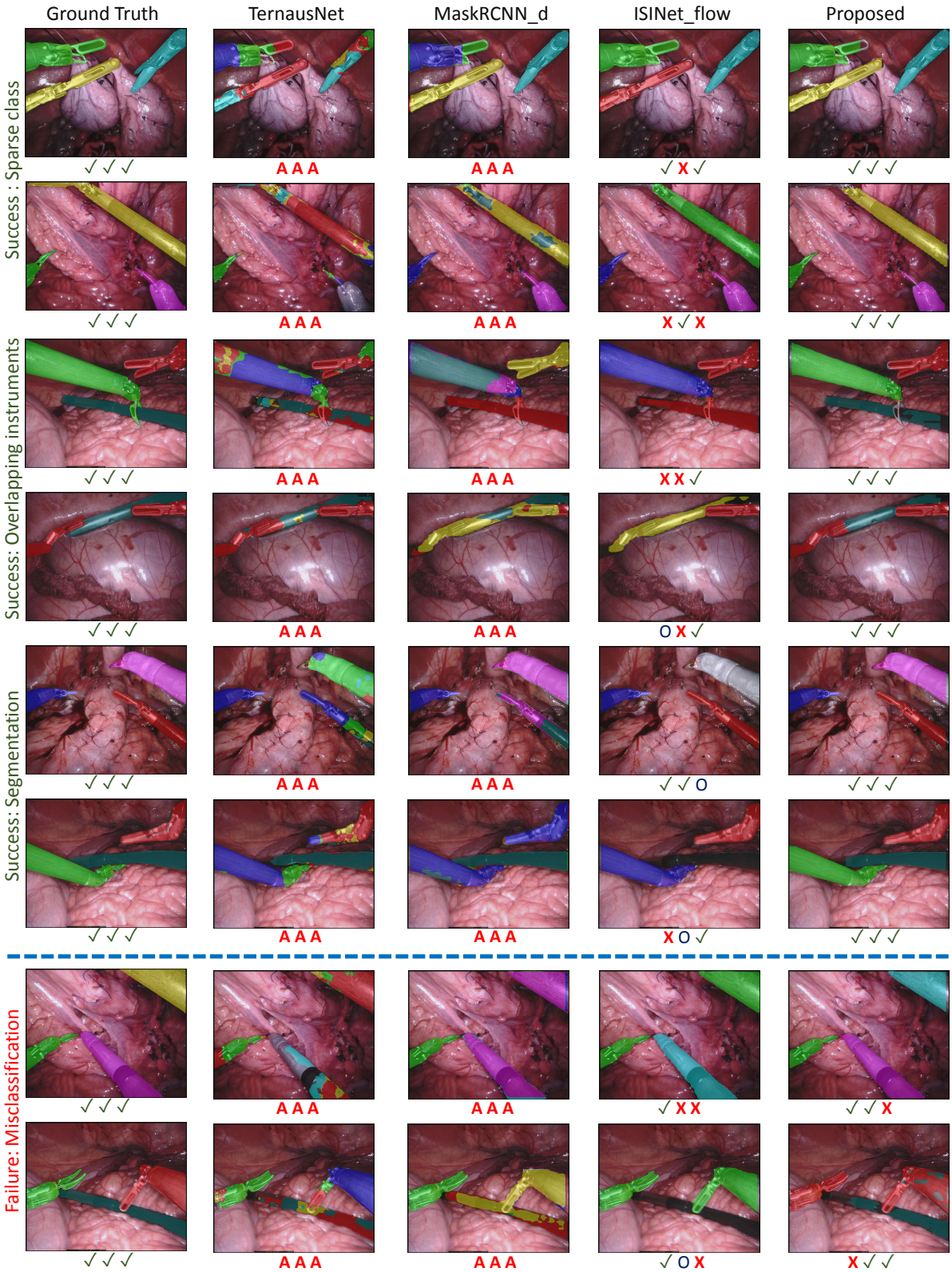


Figure 3. Qualitative Analysis for the comparison of instance segmentation: 4 symbols are used to show the results; ✓ represents the instance labeled correctly, ✗shows the misclassified instance, and 'O' represents missed instance, A letter 'A' indicates an ambiguous instance, where it is ambiguous to select the class of the instrument either due to over-segmentation or due to multiple copies of instrument classes at the same region. We show better classification in the cases of sparse class, overlapping instruments, and do not miss instrument instances. Our failure cases include cases where the instance shows the only shaft of the instrument and a significant change in instrument orientation.

Method	Conference	Arch.	Ch_ IoU	ISI_ IoU	Instrument Classes IOU							mc IoU
					BF	PF	LND	VS/ SI	GR/ CA	MCS	UP	
Dataset EV17												
NLI Methods												
MaskRCNN [24]	ICCV17	R50	45.65	41.77	27.59	33.67	43.96	17.95	0.80	4.20	8.98	19.59
CascadeRCNN [10]	CVPR18	R50	49.03	39.9	33.47	32.03	44.1	16.36	1.38	3.74	10.94	20.29
HTC [11]	CVPR19	R50	43.81	40.39	35.86	27.01	46.3	14.16	1.36	7.05	9.4	20.96
MScoring_RCNN [27]	CVPR19	R50	47.63	44.54	37.95	38.48	49.43	13.55	2.57	3.93	9.52	25.23
SimCal [51]	ECCV20	R50	49.56	45.71	39.44	38.01	46.74	16.52	1.9	1.98	13.11	23.78
CondInst [47]	ECCV20	R50	59.02	52.12	44.29	38.03	47.38	24.77	4.51	15.21	15.67	27.12
BMaskRCNN [13]	ECCV20	R50	49.81	38.81	32.89	32.82	41.93	12.66	2.07	1.37	14.43	19.74
SOLO [52]	NeurIPS20	R50	35.41	33.72	22.05	23.17	41.07	7.68	0	11.29	4.6	15.79
SCNet [48]	AAAI21	R50	48.17	46.92	43.96	29.54	48.75	22.89	1.19	4.9	14.47	25.98
MFTA [20]	CVPR21	R50	46.16	41.77	31.16	35.07	39.9	12.05	2.28	6.08	11.61	20.27
DetectoRS [39]	CVPR21	R50	50.93	47.38	48.54	34.36	49.72	20.33	2.04	8.92	10.58	24.93
Orienmask [17]	ICCV21	Dknt53	42.09	39.27	40.42	28.78	44.48	12.11	3.91	15.18	12.32	23.22
QueryInst [19]	ICCV21	R50	33.59	33.06	20.87	12.37	46.75	10.48	0.52	0.39	4.58	15.32
FASA [56]	ICCV21	R50	34.38	29.67	20.13	18.81	39.12	8.34	0.68	2.17	3.46	13.24
Mask2Former [12]	CVPR22	Trfmr	40.39	39.84	19.60	20.22	45.44	11.95	0.00	1.48	22.10	17.78
S3Net (+Mask2former)		R50	53.31	51.2	49.48	29.91	70.61	32.98	19.53	18.35	49.51	38.13
EVS Methods												
TernausNet-11 [45]	ICMLA18	UNet11	35.27	12.67	13.45	12.39	20.51	5.97	1.08	1	16.76	10.17
MF-TAPNET [31]	MICCAI19	UNet	37.35	13.49	16.39	14.11	19.01	8.11	0.31	4.09	13.4	10.77
ISINET [21]	MICCAI20	R50	55.62	52.2	38.7	38.5	50.09	27.43	2.01	28.72	12.56	28.96
TraSeTR [58]	ICRA22	Trfmr	60.4	65.2	45.2	**56.7**	55.8	**38.9**	11.4	31.3	18.2	36.79
S3Net (+MaskRCNN)		R50	**72.54**	**71.99**	**75.08**	54.32	**61.84**	35.5	**27.47**	**43.23**	**28.38**	**46.55**
Dataset EV18												
NLI Methods												
MaskRCNN [24]	ICCV17	R50	69.41	67.94	72.85	43.13	0.85	32.63	0	86.16	0	33.66
CascadeRCNN [10]	CVPR18	R50	67.11	66.29	71.22	33.6	4.94	0	0	90.61	2.62	29
HTC [11]	CVPR19	R50	69.07	68.04	72.45	36.64	1.64	37.04	0	88.27	1.95	34
MScoring_RCNN [27]	CVPR19	R50	65.19	64.04	68.69	31.23	4.81	0	0	88.23	1.75	27.82
SimCal [51]	ECCV20	R50	68.56	67.58	73.67	40.35	5.57	0	0	89.84	0	29.92
CondInst [47]	ECCV20	R50	72.27	71.55	77.42	37.43	7.77	43.62	0	87.8	0	36.29
BMaskRCNN [13]	ECCV20	R50	68.94	67.23	70.04	28.91	9.97	45.01	4.28	86.73	3.31	35.46
SOLO [52]	NeurIPS20	R50	65.59	64.88	69.46	23.92	2.61	36.19	0	87.97	0	31.45
SCNet [48]	AAAI21	R50	71.74	70.99	78.4	47.97	5.22	29.52	0	86.69	0	35.4
MFTA [20]	CVPR21	R50	69.2	67.97	71	31.62	3.93	43.48	9.9	87.77	3.86	35.94
DetectoRS [39]	CVPR21	R50	66.69	65.06	73.94	46.85	0	0	0	79.92	0	28.67
Orienmask [17]	ICCV21	Dknt53	67.69	66.77	68.95	38.66	0	31.25	0	91.21	0	32.87
QueryInst [19]	ICCV21	R50	66.44	65.82	74.13	31.68	2.3	0	0	87.28	0	27.91
FASA [56]	ICCV21	R50	68.31	66.84	72.82	37.64	5.62	0	0	89.02	1.03	29.45
Mask2Former [12]	CVPR22	Trfmr	65.47	64.69	69.35	24.13	0	0	0	89.96	10.29	27.67
S3Net (+Mask2former)		R50	67.78	67.06	71.18	29.77	1.59	0	0	90.61	10.29	29.06
EVS Methods												
TernausNet-11 [45]	ICMLA18	UNet11	46.22	39.87	44.2	4.67	0	0	0	50.44	0	14.19
MF-TAPNET [31]	MICCAI19	UNet	67.87	39.14	69.23	6.1	11.68	14	0.91	70.24	0.57	24.68
ISINET [21]	MICCAI20	R50	73.03	70.97	73.83	48.61	30.98	37.68	0	88.16	2.16	40.21
TraSeTR [58]	ICRA22	Trfmr	**76.2**	_	76.3	**53.3**	**46.5**	40.6	**13.9**	86.3	**17.5**	**47.77**
S3Net (+MaskRCNN)		R50	75.81	74.02	**77.22**	50.87	19.83	**50.59**	0	**92.12**	7.44	42.58

Table 1. Performance of SOTA instance segmentation methods on EV17 and EV18 instrument segmentation datasets. (R50 represents ResNet-50-FPN, Trfmr represents Transformer, BF-Bipolar Forceps, PF-Prograsp Forceps, LND-Large Needle Driver, VS/SI- Vessel Sealer/ Suction Instrument, GR/CA- Grasping Retractor/Clip Applier, MCS-Monopolar Curved Scissors, UP-Ultrasound Probe)

segmentation accuracy. First, the present two-stage classification heads are weak and are the bottleneck in accuracy. Second, the instrument's oblique posture and design allow the background to seep into the rectangular boxes, complicating classification. Third, cross entropy-based loss makes learning visually similar instruments harder. We investigated all three claims.

For the first claim, we replace a two-stage model's (MaskRCNN) predicted label with the ground truth label. This simple change increases the mask AP50 score from

	Model	Ch_ IoU
Stage 1 & 2	Stage 1 & 2	37.97
	Stage 1 & 2_wsr	53.30
	Stage 1 & 2_maskc	57.35
	Stage 1 & 2_wma	57.09
Stage 3	Stage 3_cel	63.63
	S3Net	72.54

Table 2. Ablation Studies of the proposed S3Net on EV17

0.65 to 0.90, showing classification inaccuracy.

We evaluate Video 1 frames of EV17 containing ultrasound probes for the second claim. MaskRCNN predicts 37 ultrasound probe bounding boxes out of 224, with 26 having an IoU of 0.75. 22 of the 26 boxes had an aspect ratio greater than 3, whereas just 4 had an aspect ratio less than 3. Elongated boxes' prediction accuracy was 84%. MaskRCNN has substantially greater accuracy when the ground truth box firmly hugs the instrument.

For the third claim, we compare the outcomes after training the third stage using cross-entropy loss and arc loss-based metric learning.

5.3. Qualitative Analysis

The qualitative classification accuracy results are shown in Fig. 3. We show the comparative results of a semantic segmentation method, a typical instance segmentation method, and an EVS method qualitatively. We show better classification in the cases of sparse class and overlapping instruments. Our failure cases include cases where the instance shows only the instrument shaft and a significant change in instrument orientation. Our better performance on the sparse class of Grasping retractor is due to the metric learning-based training loss. Because we devise a mask-attention-based classifier, the network performs well in classifying overlapping regions. We only focused on improving the classification of instance segmentation and not on the temporal context because of the high dependency of the classification accuracy for the next stage of applications. A future direction of this problem can be towards improving the masks and improving the instance labels further based on the temporal information.

5.4. Ablation studies

Tab. 2 gives the result of various ablation studies performed to understand the importance of various modules in our system. We describe the notation below:

Stage 1 & 2: Here, we report the accuracy obtained by our model after 2nd stage without the post-processing. Since our model uses the MaskRCNN backbone for the first two stages, this is essentially the accuracy of MaskRCNN using our hyper-parameters.

Stage 1 & 2 _wsr: Result of the second stage using our post-processing of non-maximal suppression across classes.

Stage 1 & 2 _maskc: In the current MaskRCNN, classification and mask prediction are performed in parallel. As per the thesis of this paper, the classification of erroneous boxes is fragile. Hence, in this experiment, we have changed the ordering of the stage 2 predictions. Now the classification is not performed parallel to mask prediction, but after the mask prediction and is done on the mask attended features.

Stage 1 & 2 _wma: We have explored whether we can use features of stage 2 for classification instead of training a separate classifier stage. In this experiment, we keep the first two stages as is, but after stage 2, use the mask attended features from stage 2 only. The difference between this and the previous configuration is that in the previous config, the original classifier of MaskRCNN was disabled, but in this one, it remains as is.

Stage 3 _cel: Here we train S3Net third stage-trained using cross-entropy instead of arc loss. The lower accuracy of this configuration serves to validate one of this paper's key observations, that the instrument's visual similarity makes it difficult to classify, and hence learning representation and classification should be disentangled using metric learning.

6. Conclusion

In this study, we investigated the reasons for the low performance of techniques developed for natural image, on the surgical instrument segmentation tasks. We also showed how carefully designed architectural innovations can successfully mitigate the challenges. We conduct exhaustive experiments on the benchmark robot-assisted surgery datasets EndoVis2017 (EV17), and EndoVis2018 (EV18). The proposed method generalizes well on all these datasets, outperforms the instance segmentation methods with varying backbones, and achieves at least 12 points (20%) improvement over the SOTA on the benchmark EV17 challenge. We conclude that adding a third classification stage improves the results for applications involving fine-grained classification, such as surgical tool segmentation. We hope that our analysis and the innovations to mitigate the challenges specific to surgical instruments will spark similar interest among researchers for the effective application of advancements in natural imaging models to surgical imaging tasks. The proposed framework can be used for downstream applications that depend on tool identification and segmentation. We plan to extend the method to include tracking cues and further improve classification accuracy.

Acknowledgements: This work was supported by Department of Biotechnology, Ministry of Science and Technology, India (Project No. BT/PR13455/CoE/34/24/2015)

References

[1] Common objects in context (coco) dataset. https://cocodataset.org/. 1

[2] Endoscopic vision challenge. http://endovis.grand-challenge.org. 2

[3] Endoscopic vision challenge: Robotic instrument segmentation sub-challenge. https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/. 1, 2, 5

[4] Endoscopic vision challenge: Robotic instrument segmentation sub-challenge. https://endovissub2018-roboticscenesegmentation.grand-challenge.org/home/. 5

[5] Imagenet dataset. https://www.image-net.org/. 1

[6] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019. 5

[7] Britty Baby, Vinkle Kumar Srivastav, Ramandeep Singh, Ashish Suri, and Subhashis Banerjee. Neuro-endo-activity-tracker: An automatic activity detection application for neuro-endo-trainer: Neuro-endo-activity-tracker. In *ICACCI*, pages 987–993. IEEE, 2016. 2

[8] Sebastian Bodenstedt, Max Allan, Anthony Agustinos, Xiaofei Du, Luis Garcia-Peraza-Herrera, Hannes Kenngott, Thomas Kurmann, Beat Müller-Stich, Sebastien Ourselin, Daniil Pakhomov, et al. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. *arXiv preprint arXiv:1805.02475*, 2018. 2

[9] David Bouget, Max Allan, Danail Stoyanov, and Pierre Jannin. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. *Medical image analysis*, 35:633–654, 2017. 1

[10] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018. 7

[11] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, pages 4974–4983, 2019. 7

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 5, 7

[13] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving mask r-cnn. In *ECCV*, pages 660–676. Springer, 2020. 7

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset. In *CVPR Workshop on the Future of Datasets in Vision*, volume 2, 2015. 1

[15] Jifeng Dai, Kaiming He, and Jian Sun. Convolutional feature masking for joint object and stuff segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3992–4000, 2015. 4

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 2, 4

[17] Wentao Du, Zhiyu Xiang, Shuya Chen, Chengyu Qiao, Yiman Chen, and Tingming Bai. Real-time instance segmentation with discriminative orientation maps. In *CVPR*, pages 7314–7323, 2021. 7

[18] Xiaofei Du, Thomas Kurmann, Ping-Lin Chang, Maximilian Allan, Sebastien Ourselin, Raphael Sznitman, John D Kelly, and Danail Stoyanov. Articulated multi-instrument 2-d pose estimation using fully convolutional networks. *IEEE transactions on medical imaging*, 37(5):1276–1287, 2018. 2

[19] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *CVPR*, pages 6910–6919, 2021. 7

[20] Dan Andrei Ganea, Bas Boom, and Ronald Poppe. Incremental few-shot instance segmentation. In *CVPR*, pages 1185–1194, 2021. 7

[21] Cristina González, Laura Bravo-Sánchez, and Pablo Arbelaez. Isinet: an instance-based approach for surgical instrument segmentation. In *MICCAI*, pages 595–605. Springer, 2020. 1, 2, 3, 5, 7

[22] Maria Grammatikopoulou, Evangello Flouty, Abdolrahim Kadkhodamohammadi, Gwenol'e Quellec, Andre Chow, Jean Nehme, Imanol Luengo, and Danail Stoyanov. Cadis: Cataract dataset for image segmentation. *arXiv preprint arXiv:1906.11586*, 2019. 2

[23] SM Hasan and Cristian A Linte. U-netplus: a modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instrument. *arXiv preprint arXiv:1902.08994*, 2019. 1, 2

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2, 3, 5, 7

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[26] HS Himal. Minimally invasive (laparoscopic) surgery. *Surgical Endoscopy And Other Interventional Techniques*, 16(12):1647–1652, 2002. 1

[27] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *CVPR*, pages 6409–6418, 2019. 7

[28] Mobarakol Islam, Daniel Anojan Atputharuban, Ravikiran Ramesh, and Hongliang Ren. Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning. *IEEE Robotics and Automation Letters*, 4(2):2188–2195, 2019. 1, 2

[29] Mobarakol Islam, VS Vibashan, and Hongliang Ren. Ap-mtl: Attention pruned multi-task learning model for real-time instrument detection and segmentation in robot-assisted surgery. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8433–8439. IEEE, 2020. 2, 3

[30] Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 691–699. IEEE, 2018. 2

[31] Yueming Jin, Keyun Cheng, Qi Dou, and Pheng-Ann Heng. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In *MICCAI*, pages 440–448. Springer, 2019. 1, 2, 5, 7

[32] Xiaowen Kong, Yueming Jin, Qi Dou, Ziyi Wang, Zerui Wang, Bo Lu, Erbao Dong, Yun-Hui Liu, and Dong Sun. Accurate instance segmentation of surgical instruments in robotic surgery: model refinement and cross-dataset evaluation. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–8, 2021. 2, 3

[33] Thomas Kurmann, Pablo Márquez-Neila, Max Allan, Sebastian Wolf, and Raphael Sznitman. Mask then classify: multi-instance segmentation for surgical instruments. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–10, 2021. 2, 3

[34] Thomas Kurmann, Pablo Marquez Neila, Xiaofei Du, Pascal Fua, Danail Stoyanov, Sebastian Wolf, and Raphael Sznitman. Simultaneous recognition and pose estimation of instruments in minimally invasive surgery. In *MICCAI*, pages 505–513. Springer, 2017. 2

[35] Daochang Liu, Yuhui Wei, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Unsupervised surgical instrument segmentation via anchor generation and semantic diffusion. In *MICCAI*, pages 657–667. Springer, 2020. 2

[36] Jie Liu, Xiaoqing Guo, and Yixuan Yuan. Prototypical interaction graph for unsupervised domain adaptation in surgical instrument segmentation. In *MICCAI*, pages 272–281. Springer, 2021. 2

[37] Zhen-Liang Ni, Gui-Bin Bian, Guan-An Wang, Xiao-Hu Zhou, Zeng-Guang Hou, Hua-Bin Chen, and Xiao-Liang Xie. Pyramid attention aggregation network for semantic segmentation of surgical instruments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11782–11790, 2020. 1, 2

[38] Zhen-Liang Ni, Gui-Bin Bian, Xiao-Hu Zhou, Zeng-Guang Hou, Xiao-Liang Xie, Chen Wang, Yan-Jie Zhou, Rui-Qi Li, and Zhen Li. Raunet: Residual attention u-net for semantic segmentation of cataract surgical instruments. In *International Conference on Neural Information Processing*, pages 139–149. Springer, 2019. 2

[39] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *CVPR*, pages 10213–10224, 2021. 1, 7

[40] Fangbo Qin, Shan Lin, Yangming Li, Randall A Bly, Kris S Moe, and Blake Hannaford. Towards better surgical instrument segmentation in endoscopic vision: multi-angle feature aggregation and contour supervision. *IEEE Robotics and Automation Letters*, 5(4):6639–6646, 2020. 2

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2

[42] Tobias Ross, Annika Reinke, Peter M Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc Filimon, Patrick Scholz, Thuy Nuong Tran, et al. Robust medical instrument segmentation challenge 2019. *arXiv preprint arXiv:2003.10299*, 2020. 2

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[44] Manish Sahu, Anirban Mukhopadhyay, and Stefan Zachow. Simulation-to-real domain adaptation with teacher–student learning for endoscopic instrument segmentation. *International journal of computer assisted radiology and surgery*, 16(5):849–859, 2021. 2

[45] Alexey A Shvets, Alexander Rakhlin, Alexandr A Kalinin, and Vladimir I Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 624–628. IEEE, 2018. 1, 2, 3, 5, 7

[46] Raphael Sznitman, Carlos Becker, and Pascal Fua. Fast part-based classification for instrument detection in minimally invasive surgery. In *MICCAI*, pages 692–699. Springer, 2014. 2

[47] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *ECCV*, pages 282–298. Springer, 2020. 7

[48] Thang Vu, Haeyong Kang, and Chang D Yoo. Scnet: Training inference sample consistency for instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2701–2709, 2021. 7

[49] Jiacheng Wang, Yueming Jin, Liansheng Wang, Shuntian Cai, Pheng-Ann Heng, and Jing Qin. Efficient global-local memory for real-time instrument segmentation of robotic surgical video. In *MICCAI*, pages 341–351. Springer, 2021. 1, 2

[50] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *ECCV*, pages 728–744. Springer, 2020. 2

[51] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *ECCV*, pages 728–744. Springer, 2020. 7

[52] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *arXiv preprint arXiv:2003.10152*, 2020. 7

[53] Eleanora P Westebring-van der Putten, Richard HM Goossens, Jack J Jakimowicz, and Jenny Dankelman. Haptics in minimally invasive surgery–a review. *Minimally Invasive Therapy & Allied Technologies*, 17(1):3–16, 2008. 1

[54] Xingxu Yao, Dongyu She, Haiwei Zhang, Jufeng Yang, Ming-Ming Cheng, and Liang Wang. Adaptive deep metric learning for affective image retrieval and classification. *IEEE Transactions on Multimedia*, 2020. 2

[55] Sabrina Madad Zadeh, Tom Francois, Lilian Calvet, Pauline Chauvet, Michel Canis, Adrien Bartoli, and Nicolas Bourdel. Surgai: deep learning for computerized laparoscopic image understanding in gynaecology. *Surgical endoscopy*, 34(12):5377–5383, 2020. 2

[56] Yuhang Zang, Chen Huang, and Chen Change Loy. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. *arXiv preprint arXiv:2102.12867*, 2021. 7

[57] Zixu Zhao, Yueming Jin, Xiaojie Gao, Qi Dou, and Pheng-Ann Heng. Learning motion flows for semi-supervised instrument segmentation from robotic surgical video. In *MICCAI*, pages 679–689. Springer, 2020. 2

[58] Zixu Zhao, Yueming Jin, and Pheng-Ann Heng. Trasetr: Track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery. *arXiv preprint arXiv:2202.08453*, 2022. 2, 3, 5, 7