# Learnable Query Initialization for Surgical Instrument Instance Segmentation

Rohan Raju Dhanakshirur[1(✉)], K. N. Ajay Shastry[1], Kaustubh Borgavi[1], Ashish Suri[2], Prem Kumar Kalra[1], and Chetan Arora[1]

[1] Indian Institute of Technology Delhi, New Delhi, India
rohanrd@sit.iitd.ac.in, rohanrd28296@gmail.com
[2] AIIMS, New-Delhi, India

**Abstract.** Surgical tool classification and instance segmentation are crucial for minimally invasive surgeries and related applications. Though most of the state-of-the-art for instance segmentation in natural images use transformer-based architectures, they have not been successful for medical instruments. In this paper, we investigate the reasons for the failure. Our analysis reveals that this is due to incorrect query initialization, which is unsuitable for fine-grained classification of highly occluded objects in a low data setting, typical for medical instruments. We propose a class-agnostic Query Proposal Network (QPN) to improve query initialization inputted to the decoder layers. Towards this, we propose a deformable-cross-attention-based learnable Query Proposal Decoder (QPD). The proposed QPN improves the recall rate of the query initialization by 44.89% at 0.9 IOU. This leads to an improvement in segmentation performance by 1.84% on Endovis17 and 2.09% on Endovis18 datasets, as measured by ISI-IOU. The source code can be accessed at https://aineurosurgery.github.io/learnableQPD.

## 1 Introduction

**Background:** Minimally invasive surgeries (MIS) such as laparoscopic and endoscopic surgeries have gained widespread popularity due to the significant reduction in the time of surgery and post-op recovery [16,21]. Surgical instrument instance segmentation (SIIS) in these surgeries opens up the doors to increased precision and automation [18]. However, the problem is challenging due to the lack of large-scale well-annotated datasets, occlusions of tool-tip (the distinguishing part of the surgical instrument), rapid changes in the appearance, reflections due to the light source of the endoscope, smoke, blood spatter etc. [5].

**The Challenge:** Most modern techniques for SIIS [11,18,31] are multi-stage architectures, with the first stage generating region proposals (rectilinear boxes)

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_70.


H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 728–738, 2023.
https://doi.org/10.1007/978-3-031-43996-4_70

and the second stage classifying each proposal independently. Unlike natural images, rectilinear bounding boxes are not an ideal choice for medical instruments, which are long, thin, and often visible diagonally in a bounding box. Thus, the ratio of the visible tool area to the area of the bounding box is highly skewed in medical scenarios. E.g., the ratio is 0.45 for the Endovis17 dataset and 0.47 for Endovis18, the two popular MIS datasets. In contrast, it is 0.60 for the MS-COCO dataset of natural images. The ratio is important because lower numbers imply more noise due to background and a more difficult classification problem.

**Current Solution Strategy:** Recently, S3Net [4] adapted the MaskRCNN [14] backbone to propose a 3-stage architecture. Their third stage implements hard attention based on the predicted masks from the second stage and re-classifies the proposals. The hard attention avoids distraction due to the presence of large background regions in the proposal boxes, allowing them to outperform all the previous state of the art for medical instruments or natural images.

**Our Observation:** In the last few years, attention-based transformer architectures have outperformed CNN architectures for many computer-vision-based tasks. Recent transformer-based object detection models implement deformable attention [6,19,33] which predicts sampling points to focus attention on the fine-grained features in an image. One expects that this would allow transformer architectures to concentrate only on the tool instead of the background, leading to high accuracy for medical instrument instance segmentation. However, in our experiments, as well as the ones reported by [4], this is not observed. We investigate the reasons and report our findings on the probable causes. We also propose a solution strategy to ameliorate the problem. Our implementation of the strategy sets up a new state of the art for the SIIS problem.

**Contributions:** **(1)** We investigate the reason for the failure of transformer-based object detectors for the medical instrument instance segmentation tasks. Our analysis reveals that incorrect query initialization is to blame. We observe that recall of an instrument based on the initialized queries is a lowly 7.48% at 0.9 IOU, indicating that many of the relevant regions of interest do not even appear in the initialized queries, thus leading to lower accuracy at the last stage. **(2)** We observe that CNN-based object detectors employ a non-maximal suppression (NMS) at the proposal stage, which helps spread the proposal over the whole image. In contrast in transformer-based detection models, this has been replaced by taking the highest confidence boxes. In this paper, we propose to switch back to NMS-based proposal selection in transformers. **(3)** The NMS uses only bounding boxes and does not allow content interaction for proposal selection. We propose a Query Proposal Decoder block containing multiple layers of self-attention and deformable cross-attention to perform region-aware refinement of the proposals. The refined proposals are used by a transformer-based decoder backbone for the prediction of the class label, bounding box, and segmentation mask. **(4)** We show an improvement of 1.84% over the best-performing SOTA

technique on the Endovis17 and 2.09% on the Endovis18 dataset as measured by ISI-IOU.

## 2 Related Work

**Datasets:** Surgical tool recognition and segmentation has been a well-explored research topic [11]. Given the data-driven nature of recent computer vision techniques, researchers have also proposed multiple datasets for the problem. Twinanda et al. [28] have proposed an 80-video dataset of cholecystectomy surgeries, with semantic segmentation annotation for 7 tools. Al Hajj et al. [1] proposed a 50-video dataset of phacoemulsification cataract surgeries with 21 tools annotated for bounding boxes. Cadis [12] complements this dataset with segmentation masks. Ross et al. [25] in 2019 proposed a 30-video dataset corresponding to multiple surgeries with one instrument, annotated at image level for its presence. Endovis datasets, Endovis 2017 (EV17) [3] and Endovis 2018 (EV18) [2] have gained popularity in the recent past. Both of them have instance-level annotations for 7 tools. EV17 is a dataset of 10 videos of the Da Vinci robot, and EV18 is a dataset of 15 videos of abdominal porcine procedures.

**Instance Segmentation Techniques for Medical Instruments:** Multiple attempts have been made to perform instance segmentation using these datasets. [11] and [18] use a MaskRCNN-based [14] backbone pre-trained on natural images and perform cross-domain fine-tuning. Wang et al. [31] assign categories to each pixel within an instance and convert the problem into a pixel classification. They then use the ResNet backbone to solve the problem. [29] modified the MaskRCNN architecture and proposed a Sample Consistency Network to bring closer the distribution of the samples at training and test time. Ganea et al. [10] use the concept of few-shot learning on top of MaskRCNN to improve the performance. Wentao et al. [8] add a mask prediction head to YoLo V3 [23] All these algorithms use CNN-based architectures, with ROI-Align, to crop the region of interest. Since the bounding boxes are not very tight in surgical cases due to the orientation of the tools, a lot of background information is passed along with the tool, and thereby the classification performance is compromised. Baby et al. [4] use a third-stage classifier on top of MaskRCNN to correct the misclassified masks and improve the performance.

**Transformer-based Instance Segmentation for Natural Images:** On the other hand, transformer-based instance segmentation architectures [6,13,17,19] generate sampling points to extract the features and thereby learn more localised information. This gives extra leverage to transformer architectures to perform better classification. [17] propose the first transformer-based end-to-end instance segmentation architecture. They predict low-level mask embeddings and combine them to generate the actual masks. [13] learn the location-specific features by providing the information on position embeddings. [6] uses a deformable-multi-head-attention-based mechanism to enrich the segmentation task. Mask DINO [19] utilizes better positional priors as originally proposed in [33]. They also

perform box refinement at multiple levels to obtain the tight instance mask. In these architectures, the query initialization is done using the *top-k* region proposals based on their corresponding classification score. Thus ambiguity in the classification results in poor query initialization, and thereby the entire mask corresponding to that instance is missed. This leads to a significant reduction in the recall rate of these models.

## 3 Proposed Methodology

**Backbone Architecture:** We utilize Mask DINO [19] as the backbone architecture for our model. As illustrated in Fig. 1, it uses ResNet [15] as the feature extractor to generate a multi-scale feature map at varying resolutions. These feature maps are then run through an encoder to generate an enhanced feature map with the same resolution as the original feature map. The enhanced feature maps are used to generate a set of region proposals. Then, the region proposals are sorted based on their classification logit values. Mask DINO uses a $d$ dimensional query vector to represent an object's information. The top $n_q$ generated region proposals are used to initialize a set of $n_q$ queries. These queries are then passed through a series of decoder layers to obtain a set of refined queries. These refined queries are used for the purpose of detection, classification, and segmentation.

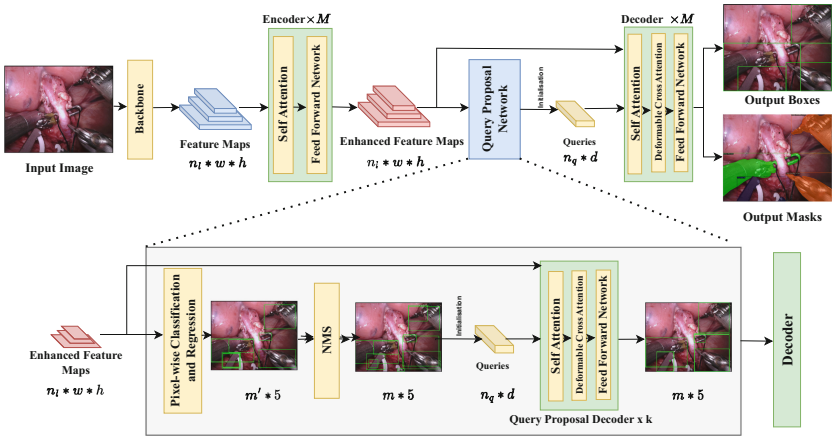


**Fig. 1.** Proposed Query Proposal Network

**Problems with Mask DINO:** The model predicts various outputs using the queries initialized with the top $n_q$ region proposals. Our analysis reveals that most false negative outputs are due to the queries initialized with significantly fewer to no region proposals corresponding to the missed objects. During the initialization procedure, Mask DINO sorts region proposals based on their classification logit values. The surgical instruments resemble one another and have

few distinguishing features. In case of label confusion, which happens often in medical instruments, the encoder outputs a proposal with low confidence. When sorted, these low-confidence proposals get deleted. Hence, in the proposed architecture we argue for class-independent proposal selection, which does not rely on the classification label or its confidence at this stage.

**Query Proposal Network: Spatial Diversification of the Proposals:** The proposed Query Proposal Network (QPN) is shown in Fig. 1. QPN takes the enhanced feature maps as input and performs a pixel-wise classification and regression to obtain the initial region proposals. These initial region proposals undergo Non-Maximal Suppression (NMS) in order to remove the duplicates, but more importantly, output the proposal boxes which are spread all through the image. This is important because, given the complexity of the classification in medical instruments, we do not wish to overly rely on the classification label and would rather explore more regions of interest in the decoder. Hence, we choose top $k$ proposals based on the NMS instead of the label confidence.

**Table 1.** The comparison of the proposed methodology against the other SOTA architectures for the Endovis 17 [3] dataset. Here rows 4–17 are obtained from [4]. Note that [34] is a video instance segmentation paper and uses video information to perform segmentation.

| Method | Conference | Instrument Classes IOU | | | | | | | Ch. IOU | ISI. IOU | MC. IOU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BF | PF | LND | VS | GR | MCS | UP | | | |
| Dataset EV17 | | | | | | | | | | | |
| SimCaL [30] | ECCV20 | 39.44 | 38.01 | 46.74 | 16.52 | 1.90 | 1.98 | 13.11 | 49.56 | 45.71 | 23.78 |
| CondInst [27] | ECCV20 | 44.29 | 38.03 | 47.38 | 24.77 | 4.51 | 15.21 | 15.67 | 59.02 | 52.12 | 27.12 |
| BMaskRCNN [7] | ECCV20 | 32.89 | 32.82 | 41.93 | 12.66 | 2.07 | 1.37 | 14.43 | 49.81 | 38.81 | 19.74 |
| SOLO [31] | NeurIPS20 | 22.05 | 23.17 | 41.07 | 7.68 | 0.00 | 11.29 | 4.60 | 35.41 | 33.72 | 15.79 |
| ISINET [11] | MICCAI20 | 38.70 | 38.50 | 50.09 | 27.43 | 2.01 | 28.72 | 12.56 | 55.62 | 52.20 | 28.96 |
| SCNet [29] | AAAI21 | 43.96 | 29.54 | 48.75 | 22.89 | 1.19 | 4.90 | 14.47 | 48.17 | 46.92 | 25.98 |
| MFTA [10] | CVPR21 | 31.16 | 35.07 | 39.9 | 12.05 | 2.28 | 6.08 | 11.61 | 46.16 | 41.77 | 20.27 |
| Detectors [22] | CVPR21 | 48.54 | 34.36 | 49.72 | 20.33 | 2.04 | 8.92 | 10.58 | 50.93 | 47.38 | 24.93 |
| Orienmask [8] | ICCV21 | 40.42 | 28.78 | 44.48 | 12.11 | 3.91 | 15.18 | 12.32 | 42.09 | 39.27 | 23.22 |
| Querylnst [9] | ICCV21 | 20.87 | 12.37 | 46.75 | 10.48 | 0.52 | 0.39 | 4.58 | 33.59 | 33.06 | 15.32 |
| FASA [32] | ICCV21 | 20.13 | 18.81 | 39.12 | 8.34 | 0.68 | 2.17 | 3.46 | 34.38 | 29.67 | 13.24 |
| Mask2Former [6] | CVPR22 | 19.60 | 20.22 | 45.44 | 11.95 | 0.00 | 1.48 | 22.10 | 40.39 | 39.84 | 17.78 |
| TraSeTR* [34] | ICRA22 | 45.20 | **56.70** | 55.80 | 38.90 | 11.40 | 31.30 | 18.20 | 60.40 | 65.20 | 36.79 |
| S3Net [4] | WACV23 | **75.08** | 54.32 | 61.84 | 35.50 | 27.47 | 43.23 | 28.38 | 72.54 | 71.99 | 46.55 |
| Mask DINO [19] | Archive 22 | 64.14 | 45.29 | 78.96 | 59.35 | 21.90 | 36.70 | 30.72 | 75.96 | 77.63 | 43.86 |
| Proposed architecture | | 70.61 | 45.84 | **80.01** | **63.41** | **33.64** | **66.57** | **35.28** | **77.8** | **79.58** | **49.92** |

**Query Proposal Network: Content-based Inter-proposal Interaction:** The top $k$ region proposals from NMS are used to initialize the queries for the proposed Query Proposal Decoder (QPD). Note that the NMS works only on the basis of box coordinates and does not take care of content embeddings into account. We try to make up for the gap through the QPD module. The QPD module consists of self-attention, cross-attention and feed-forward layers. The

self-attention layer of QPD allows the queries to interact with each other and the duplicates are avoided. We use the standard deformable cross-attention module as proposed in [35]. Unlike traditional cross-attention-based mechanisms, this method attends only to a fixed number of learnable key points around the middlemost pixel of every region proposal (query) irrespective of the size of the feature maps. This allows us to achieve better convergence in larger feature maps. Thus, the cross-attention layer allows the interaction of queries with the enhanced feature maps from the encoder and feature representation for each query is obtained. The feed-forward layer refines the queries based on the feature representation obtained in the previous layer. We train the QPD layers using the mask and bounding box loss as is common in transformer architectures [19]. Note that no classification loss is back-propagated. This allows the network to perform query refinement irrespective of the classification label and retains queries corresponding to the object instances, which were omitted due to ambiguity in classification. The queries outputted by the QPD module are used to initialize the standard decoder network of Mask DINO.

**Table 2.** The comparison of the proposed methodology against the other SOTA architectures for the Endovis 18 [2] dataset. Here rows 4–17 are obtained from [4] Note that [34] is a video instance segmentation paper and uses video information to perform segmentation.

| Method | Conference | Instrument Classes IOU | | | | | | | Ch. IOU | ISI. IOU | MC. IOU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BF | PF | LND | SI | CA | MCS | UP | | | |
| Dataset EV18 | | | | | | | | | | | |
| SimCaL [30] | ECCV20 | 73.67 | 40.35 | 5.57 | 0.00 | 0.00 | 89.84 | 0.00 | 68.56 | 67.58 | 29.92 |
| CondInst [27] | ECCV20 | 77.42 | 37.43 | 7.77 | 43.62 | 0.00 | 87.8 | 0.00 | 72.27 | 71.55 | 36.29 |
| BMaskRCNN [7] | ECCV20 | 70.04 | 28.91 | 9.97 | 45.01 | 4.28 | 86.73 | 3.31 | 68.94 | 67.23 | 35.46 |
| SOLO [31] | NeurIPS20 | 69.46 | 23.92 | 2.61 | 36.19 | 0.00 | 87.97 | 0.00 | 65.59 | 64.88 | 31.45 |
| ISINET [31] | MICCAI20 | 73.83 | 48.61 | 30.98 | 37.68 | 0.00 | 88.16 | 2.16 | 73.03 | 70.97 | 40.21 |
| SCNet [29] | AAAI21 | 78.40 | 47.97 | 5.22 | 29.52 | 0.00 | 86.69 | 0.00 | 71.74 | 70.99 | 35.40 |
| MFTA [10] | CVPR21 | 71.00 | 31.62 | 3.93 | 43.48 | 9.90 | 87.77 | 3.86 | 69.20 | 67.97 | 35.94 |
| Detectors [22] | CVPR21 | 73.94 | 46.85 | 0.00 | 0.00 | 0.00 | 79.92 | 0.00 | 66.69 | 65.06 | 28.67 |
| Orienmask [8] | ICCV21 | 68.95 | 38.66 | 0.00 | 31.25 | 0.00 | 91.21 | 0.00 | 67.69 | 66.77 | 32.87 |
| QueryInst [9] | ICCV21 | 74.13 | 31.68 | 2.30 | 0.00 | 0.00 | 87.28 | 0.00 | 66.44 | 65.82 | 27.91 |
| FASA [32] | ICCV21 | 72.82 | 37.64 | 5.62 | 0.00 | 0.00 | 89.02 | 1.03 | 68.31 | 66.84 | 29.45 |
| Mask2Former [6] | CVPR22 | 69.35 | 24.13 | 0.00 | 0.00 | 0.00 | 89.96 | 10.29 | 65.47 | 64.69 | 27.67 |
| TraSeTR* [34] | ICRA22 | 76.30 | 53.30 | **46.5** | 40.6 | **13.90** | 86.3 | **17.5** | 76.20 | | **47.77** |
| S3Net [4] | WACV23 | 77.22 | 50.87 | 19.83 | 50.59 | 0.00 | 92.12 | 7.44 | 75.81 | 74.02 | 42.58 |
| Mask DINO [19] | Archive 22 | 82.35 | 57.67 | 0.83 | **60.46** | 0.00 | 90.73 | 0.00 | 75.63 | 76.39 | 41.73 |
| Proposed architecture | | **82.8** | **60.94** | 19.96 | 49.70 | 0.00 | **93.93** | 0.00 | **77.77** | **78.43** | 43.84 |

**Implementation Details:** We train the proposed architecture using three kinds of losses, classification loss $\mathcal{L}_{\text{cls}}$, box regression loss $\mathcal{L}_{\text{box}}$, and the mask prediction loss $\mathcal{L}_{\text{mask}}$. We use focal loss [20] as $\mathcal{L}_{\text{cls}}$. We use $\ell_1$ and GIOU [24] loss for $\mathcal{L}_{\text{box}}$. For $\mathcal{L}_{\text{mask}}$, we use cross entropy and IOU (or dice) loss. The total loss is:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{cls}} + \lambda_2 \ell_1 + \lambda_3 \mathcal{L}_{\text{giou}} + \lambda_4 \mathcal{L}_{\text{ce}} + \lambda_5 \mathcal{L}_{\text{dice}} \tag{1}$$

Through hyper-parameter tuning, we set $\lambda = [0.19, 0.24, 0.1, 0.24, 0.24]$. We use a batch size of 8. The initial learning rate is set to 0.0001, which drops by 0.1 after every 20 epochs. We set 0.9 as the Nesterov momentum coefficient. We train the network for 50 epochs on a server with 8 NVidia A100, 40 GB GPUs. Besides QPN we use the exact same architecture as proposed in MaskDINO [19]. However, we perform transfer learning using the MaskDINO pre-trained weights, and therefore, we do not use "GT+noise" as an input to the decoder.

## 4 Results and Discussion

**Evaluation Methodology:** We demonstrate the performance of the proposed methodology on two benchmark Robot-assisted endoscopic surgery datasets Endovis 2017 [3] (denoted as EV17), and Endovis 2018 [2] (denoted as EV18) as performed by [4]. EV17 is a dataset of 10 videos (1800 images) obtained from the Da Vinci robotic system with 7 instruments. We adopt the same four-fold cross-validation strategy as shown by [26] for the EV17 dataset. EV18 is a real

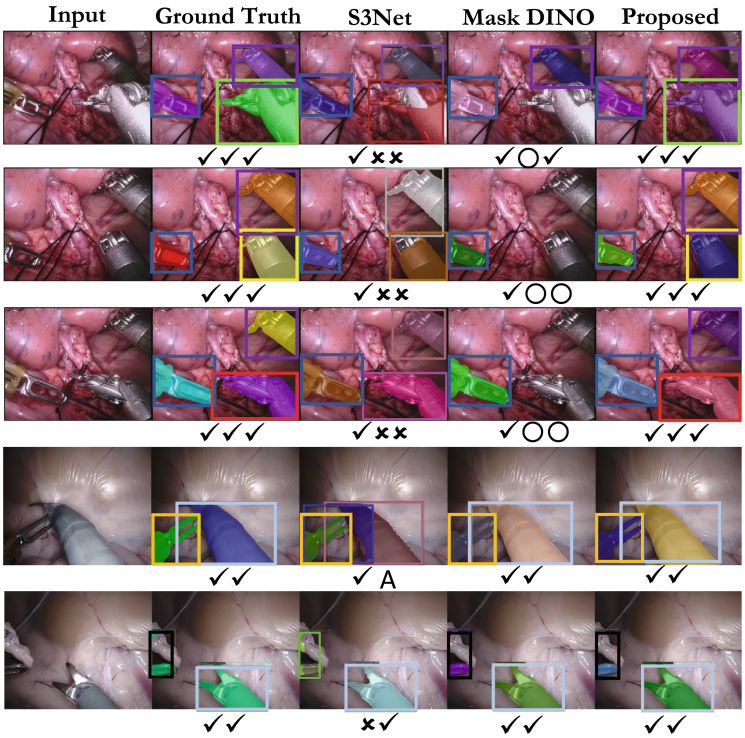


**Fig. 2.** The qualitative analysis of instance segmentation by the proposed methodology against the other SOTA algorithms: Here, ✓ indicates that the instrument is classified and segmented correctly. *O* indicates missed instance, × indicates incorrect classification and *A* indicates ambiguous instance.

surgery dataset of 15 videos containing 1639 training and 596 validation images, with 7 instruments. [4] corrected the misclassified ground truth labels and added instance-level annotations to this dataset. We evaluate the performance of the proposed algorithm using challenge IOU as proposed by [3], as well as ISI IOU and Mean class IOU as reported in [11]. We also report per instrument class IOU as suggested by [4].

**Quantitative Results on EV17, EV18, and Cadis:** The results of our technique for the EV17 dataset are shown in Table 1 and for the EV18 dataset are shown in Table 2. It can be observed that the proposed technique outperforms the best-performing SOTA methods by 1.84% in terms of challenge IOU for EV17 and 1.96% for EV18 datasets. Due to the paucity of space, we show the performance of the proposed methodology on the Cadis [12] dataset in the supplementary material. The qualitative analysis of instance segmentation by the proposed methodology against the other SOTA algorithms is shown in Fig. 2. We demonstrate improved performance in the occluded and overlapping cases. We observe a testing speed of 40 FPS on a standard 40GB Nvidia A100 GPU cluster.

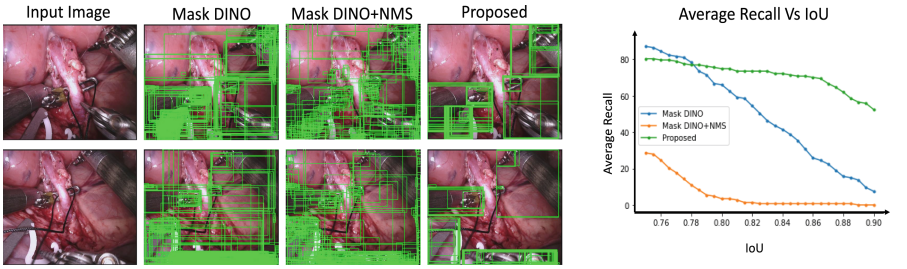


**Fig. 3.** Proposed improvement in the Query Initialization. Left: two sample images and their corresponding query initialization by different architectures. Right: shows the recall rate at various IOU thresholds

**Table 3.** Ablation analysis on EV18 dataset to demonstrate the importance of each block in the proposed architecture and to check for the most optimal number of Query Proposal Decoder (QPD) layers.

| Configuration | Instrument Classes IOU | | | | | | | Ch. IOU | ISI. IOU | MC. IOU |
|---|---|---|---|---|---|---|---|---|---|---|
| | BF | PF | LND | SI | CA | MCS | UP | | | |
| Vanilla MaskDINO | 82.35 | 57.67 | 0.83 | 60.46 | 0.00 | 90.73 | 0.00 | 75.63 | 76.39 | 41.73 |
| Mask DINO + 1 QPD | 82.41 | 66.61 | 0.00 | 40.57 | 0.00 | 89.65 | 0.00 | 75.58 | 76.28 | 39.89 |
| Mask DINO + 2 QPD | 83.71 | 60.92 | 10.55 | 50.4 | 0.00 | 92.22 | 4.84 | 76.67 | 77.51 | 43.23 |
| Mask DINO + 3 QPD | 76.24 | 61.26 | 7.06 | 48.91 | 0.00 | 92.30 | 0.00 | 74.11 | 75.00 | 40.83 |
| Mask DINO + 4 QPD | 75.43 | 45.67 | 2.67 | 48.31 | 0.00 | 94.23 | 0.00 | 71.97 | 73.46 | 38.04 |
| Mask DINO + 5 QPD | 78.25 | 27.39 | 0.92 | 47.96 | 0.00 | 93.89 | 0.00 | 71.38 | 71.79 | 35.49 |
| Mask DINO + Random Input to NMS | 81.46 | 51.73 | 25.36 | 40.73 | 0.00 | 92.73 | 0.00 | 76.19 | 76.60 | 41.70 |
| Mask DINO + Encoder Input to NMS | 82.05 | 63.02 | 15.33 | 50.61 | 0.00 | 88.71 | 0.00 | 76.21 | 76.64 | 42.82 |
| Mask DINO + NMS + 2 QPD (Proposed) | 82.80 | 60.94 | 19.96 | 49.70 | 0.00 | 93.93 | 0.00 | 77.77 | 78.43 | 43.84 |

**Evidence of Query Improvement:** The improvement in the query initialization due to the proposed architecture is demonstrated in Fig. 3. Here, we mark the centre of initialized query boxes and generate the scatter plot. The average recall rate for the EV18 dataset at 0.9 IOU for the top 300 queries in vanilla Mask DINO is 7.48%. After performing NMS, the recall rate decreases to 0.00%, but the queries get diversified. After passing the same through the proposed Query Proposal Decoder (QPD), the recall rate is observed to be 52.38%. The increase in recall rate to 52.38% indicates successful learning of QPD. It indicates that the proposed architecture is able to cover more objects in the initialized queries thereby improving the performance at the end. While the diversity of the queries is important, it is also important to ensure that the queries with higher confidence are near the ground truth to ensure better learning. Random initialization for NMS is observed with sub-optimal performance and is shown in Table 3.

**Ablation Study:** We perform an ablation on the EV18 dataset to show the importance of each block in the proposed methodology. The results of the same are summarised in Table 3. We also experiment with the number of layers in QPD. We have used the best-performing 2-layer QPD architecture for all other experiments. The reduction in performance with more QPD layers can be attributed to the under-learning of negative samples due to stricter query proposals

## 5 Conclusion

In this paper, we proposed a novel class-agnostic Query Proposal Network (QPN) to better initialize the queries for a transformer-based surgical instrument instance segmentation model. Towards this, we first diversified the queries using the non-maximal suppression and proposed a deformable-cross-attention-based learnable Query Proposal Decoder (QPD). On average, the proposed QPN improved the recall rate of the query initialization by 52.38% at 0.9 IOU. The improvement translates to an improved ISI-IOU of 1.84% and 2.09% in the publicly available Endovis 2017 and Endovis 2018 datasets, respectively.

**Acknowledgements.** We thank Dr Britty Baby for her assistance. We also thank DBT, Govt of India, and ICMR, Govt of India, for the funding under the projects BT/PR13455/CoE/34/24/2015 and 5/3/8/1/2022/MDMS(CARE) respectively.

## References

1. Al Hajj, H., et al.: CATARACTS: challenge on automatic tool annotation for cataract surgery. MIA **52**, 24–41 (2019)
2. Allan, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
3. Allan, M., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)

4. Baby, B., et al.: From forks to forceps: a new framework for instance segmentation of surgical instruments. In: WACV, pp. 6191–6201 (2023)
5. Bouget, D., Allan, M., Stoyanov, D., Jannin, P.: Vision-based and marker-less surgical tool detection and tracking: a review of the literature. MIA **35**, 633–654 (2017)
6. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.: Mask2former for video instance segmentation. arXiv preprint arXiv:2112.10764 (2021)
7. Cheng, T., Wang, X., Huang, L., Liu, W.: Boundary-preserving mask R-CNN. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12359, pp. 660–676. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58568-6_39
8. Du, W., Xiang, Z., Chen, S., Qiao, C., Chen, Y., Bai, T.: Real-time instance segmentation with discriminative orientation maps. In: ICCV, pp. 7314–7323 (2021)
9. Fang, Y., et al.: Instances as queries. In: ICCV, pp. 6910–6919 (2021)
10. Ganea, D.A., Boom, B., Poppe, R.: Incremental few-shot instance segmentation. In: CVPR, pp. 1185–1194 (2021)
11. González, C., Bravo-Sánchez, L., Arbelaez, P.: ISINet: an instance-based approach for surgical instrument segmentation. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 595–605. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_57
12. Grammatikopoulou, M., et al.: Cadis: cataract dataset for image segmentation. arXiv preprint arXiv:1906.11586 (2019)
13. Guo, R., Niu, D., Qu, L., Li, Z.: Sotr: segmenting objects with transformers. In: ICCV, pp. 7157–7166 (2021)
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV, pp. 2961–2969 (2017)
15. 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)
16. Himal, H.: Minimally invasive (laparoscopic) surgery. Surg. Endosc. Interv. Tech. **16**, 1647–1652 (2002)
17. Hu, J., et al.: ISTR: end-to-end instance segmentation with transformers. arXiv preprint arXiv:2105.00637 (2021)
18. Kong, X., et al.: Accurate instance segmentation of surgical instruments in robotic surgery: model refinement and cross-dataset evaluation. IJCARS **16**(9), 1607–1614 (2021)
19. Li, F., Zhang, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y., et al.: Mask DINO: towards a unified transformer-based framework for object detection and segmentation. arXiv preprint arXiv:2206.02777 (2022)
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV, pp. 2980–2988 (2017)
21. Westebring-van der Putten, E.P., Goossens, R.H., Jakimowicz, J.J., Dankelman, J.: Haptics in minimally invasive surgery-a review. Minim. Invasive Therapy Allied Technol. **17**(1), 3–16 (2008)
22. Qiao, S., Chen, L.C., Yuille, A.: Detectors: detecting objects with recursive feature pyramid and switchable atrous convolution. In: CVPR, pp. 10213–10224 (2021)
23. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)

24. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: a metric and a loss for bounding box regression. In: CVPR, pp. 658–666 (2019)
25. Ross, T., et al.: Comparative validation of multi-instance instrument segmentation in endoscopy: results of the ROBUST-MIS 2019 challenge. MIA **70**, 101920 (2021)
26. Shvets, A.A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.I.: Automatic instrument segmentation in robot-assisted surgery using deep learning. In: 2018 17th ICMLA, pp. 624–628. IEEE (2018)
27. Tian, Z., Shen, C., Chen, H.: Conditional convolutions for instance segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 282–298. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_17
28. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE TMI **36**(1), 86–97 (2016)
29. Vu, T., Kang, H., Yoo, C.D.: SCNet: training inference sample consistency for instance segmentation. In: AAAI, vol. 35, pp. 2701–2709 (2021)
30. Wang, T., et al.: The devil is in classification: a simple framework for long-tail instance segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12359, pp. 728–744. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58568-6_43
31. Wang, X., Zhang, R., Kong, T., Li, L., Shen, C.: SOLOv2: dynamic and fast instance segmentation. ANIPS **33**, 17721–17732 (2020)
32. Zang, Y., Huang, C., Loy, C.C.: FASA: feature augmentation and sampling adaptation for long-tailed instance segmentation. In: ICCV, pp. 3457–3466 (2021)
33. Zhang, H., et al.: DINO: detr with improved denoising anchor boxes for end-to-end object detection. In: 11th ICLR (2022)
34. Zhao, Z., Jin, Y., Heng, P.A.: Trasetr: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery. In: ICRA, pp. 11186–11193. IEEE (2022)
35. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)