Towards Precise Pose Estimation in Robotic Surgery: Introducing Occlusion-Aware Loss

Jihun Park[†1], Jiuk Hong[1], Jihun Yoon[2], Bokyung Park[2], Min-Kook Choi[2], and Heechul Jung[‡1⋆]

[1] Department of Artificial Intelligence, Kyungpook National University, Daegu, Republic of Korea
[2] hutom, Seoul, Republic of Korea
{jihun1142[†], heechul[‡]}@knu.ac.kr

Abstract. Accurate pose estimation of surgical instruments is crucial for analyzing robotic surgery videos using computer vision techniques. However, the scarcity of suitable public datasets poses a challenge in this regard. To address this issue, we have developed a new private dataset extracted from real gastric cancer surgery videos. The primary objective of our research is to develop a more sophisticated pose estimation algorithm for surgical instruments using this private dataset. Additionally, we introduce a novel loss function aimed at enhancing the accuracy of pose estimation, with a specific emphasis on minimizing root mean squared error. Leveraging the YOLOv8 model, our approach significantly outperforms existing methods and state-of-the-art techniques, thanks to the enhanced occlusion-aware loss function. These findings hold promise for improving the precision and safety of robotic-assisted surgeries.

Keywords: Surgical Instrument · Robotic Surgery · Pose Estimation · Occlusion-aware Loss

1 Introduction

The advent of robotic surgery has revolutionized the field of medicine, offering unprecedented precision, flexibility, and control, thereby enhancing the efficacy of surgical procedures and improving patient outcomes [1]. A cornerstone of robotic surgery is the accurate pose estimation of surgical instruments, a task that relies heavily on advancements in deep learning technologies.

Traditional approaches for pose estimation rely on either handcrafted features [8,9,10,11] or the utilization of three-dimensional coordinates [3,4,5] to ensure robustness against adverse visible conditions in surgical videos. Additionally, methods utilizing two-dimensional coordinates [6,7] have been introduced. These approaches are typically developed based on public datasets such as the EndoVis challenge dataset [12] or the RMIT dataset [11]. However, these datasets often fail to accurately reflect real-world surgical environments, as illustrated in Fig. 1.

⋆ Corresponding author

Fig. 1. Example of the existing datasets and our private dataset for pose estimation. (a) and (b) represent EndoVis challenge dataset [12] and RMIT dataset [11], respectively. The existing datasets are artificial and not obtained from actual surgical videos; rather, they are simulated. (c) represents our dataset to facilitate our research effectively.

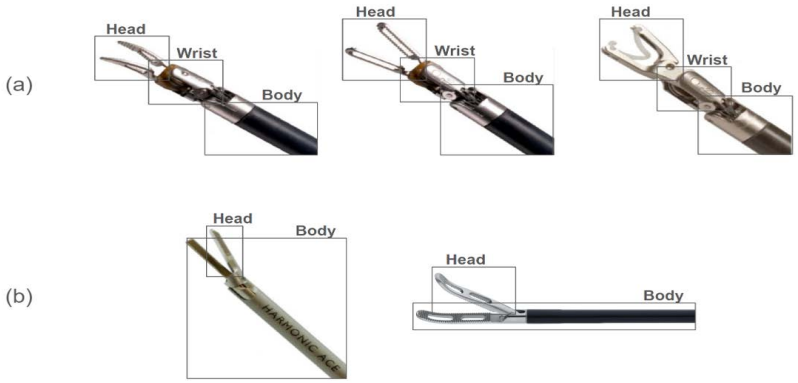

Fig. 2. Illustration of surgical instruments in our dataset. (a) represents three-part instruments such as head, wrist, and body. (b) shows two-part instruments such as head and body.

Recognizing the critical need for such datasets, our research endeavors to bridge this gap by introducing a new private dataset meticulously curated from video footage of actual gastric cancer surgeries. This dataset is specifically designed to capture the dynamic and unpredictable nature of in-the-wild scenarios, offering a rich resource for the development, training, and evaluation of advanced pose estimation models.

In addition to addressing the dataset scarcity, our research introduces an innovative loss function tailored to enhance the accuracy of pose estimation for robotic surgical instruments. By focusing on minimizing the root mean squared error, this new loss function aims to refine the precision of instrument localization. The adoption of the YOLOv8 [2] model, known for its efficiency and effectiveness in object detection, further complements our approach. The integration of an occlusion-aware loss mechanism within this model is pivotal in

overcoming the inherent challenges of instrument occlusion and the dynamic nature of surgical settings.

Through comprehensive experiments and evaluations, our approach demonstrates significant improvements over existing baseline models and current state-of-the-art methods. By leveraging our custom-built dataset and the proposed loss function in conjunction with the YOLOv8 model, we achieve outstanding accuracy in the pose estimation of robotic surgical instruments. This breakthrough not only addresses the immediate challenges in the field but also lays the groundwork for further advancements in the integration of artificial intelligence with robotic-assisted surgeries. The implications of our research extend beyond the immediate improvements in surgical precision and safety, promising to catalyze innovation and enhance the overall efficacy of robotic surgery practices.

2 Our Approach

2.1 Dataset Construction

In this study, we developed our dataset[1] utilizing robotic surgery videos based on Da Vinci robotic surgical systems. The resolution of the video is 1280×1024, which represents our dataset is high-quality. These videos encompass various surgical procedures where six common types of instruments were employed: cadiere forceps, maryland bipolar forceps, medium-large clip applier, small clip applier, curved grasper, and harmonic ace. These instruments were categorized into two groups: three-part instruments and two-part instruments (See Fig. 2)[2] Among the three-part instruments, the cadiere forceps were annotated with seven keypoint coordinates, while the maryland bipolar forceps, medium-large clip applier, and curved grasper each had six key points. The two-part instruments were annotated with five keypoint coordinates.

Each keypoint's location was annotated using two-dimensional coordinates of (x, y), where x and y represent position in the image. Additionally, to enhance the precision of the pose estimation, visibility information was also assigned to each keypoint. Using this information, we build a new loss function to improve the performance of pose estimation in terms of root mean squared error (RMSE). The visibility of keypoints was categorized as follows: a keypoint that is not visible in the image was labeled as 0; a keypoint that is present in the image but occluded was labeled as 1; and a keypoint that is fully visible and not occluded within the image was labeled as 2. See Fig. 3 for details of the annotation rules.

The dataset comprises surgical videos of ten patients, which have been segmented into individual frames, resulting in a total of $125,467$ images. For the

[1] This dataset was created with the approval from the Institutional Review Board of Kyungpook National University (IRB No.KNU2023-0483).

[2] https://pdf.medicalexpo.com/pdf/intuitive-surgical/instrument-accessory-catalog/75060-153581.html,
https://www.karlstorz.com/dz/en/product-detail-page.htm?productID=1000120469&cat=1000119605

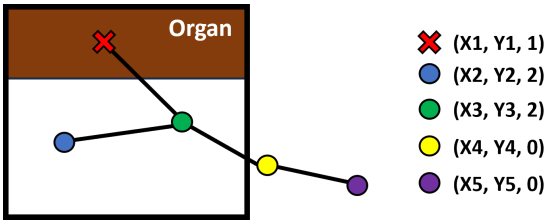

Fig. 3. Illustration of annotation details. The red cross denotes the area occluded by organs. The blue and green points represent the presence of keypoints in the image. The yellow and purple points are outside of the image. From (X1, Y1) to (X5, Y5), each value denotes a two-dimensional coordinate of each keypoint. The third part of each coordinate represents the visibility of the keypoint.

division of the dataset into train and test sets, we selected three patients as a test set to ensure that the number of keypoints maintained approximately a $2:1$ ratio between the train and test sets, while also ensuring a roughly $5:1$ ratio for the number of class-wise instruments. This was done to guarantee that there were no keypoints or instrument classes unique to either the train set or the test set. The training set consisted of $83,252$ images, while the test set comprised $42,215$ images.

2.2 YOLOv8

YOLOv8, developed by Ultralytics in 2023, is widely used for object detection, classification, segmentation, and human pose estimation. A notable enhancement is the C2F module with gradient shortcut connections, which improves the feature extractor's information flow and boosts the accuracy of detection tasks by effectively merging features with contextual information. The backbone also features the spatial pyramid pooling fast (SPPF) module, using three max-pooling layers for efficient feature map pooling with reduced computational effort and latency.

The YOLOv8 head, comprising convolutional and linear layers, processes these feature maps to generate the final output. Its anchor-free approach for object detection, which directly regresses bounding box coordinates, streamlines the architecture, and potentially increases localization accuracy by eliminating reliance on pre-defined anchor boxes.

Additionally, YOLOv8 introduces a pose estimation variant, which adds a pose head to the standard architecture. This model predicts keypoints for each object, calculating the position and confidence value for each keypoint. To train such model, the keypoint loss function is used. The loss function is based on the Euclidean distance between predicted and actual keypoint positions. This function assigns weights to each keypoint and normalizes them based on the object's scale, enhancing the model's focus on discernible features. Finally, the loss function $\mathcal{L}_{\text{pose}}$ is defined as follows:

$$\mathcal{L}_{\text{pose}}(s,i,j,k) = 1 - \frac{\sum_{n=1}^{N} \exp\left(-\frac{d_n^2}{2s^2k_n^2}\right)\delta(v_n > 0)}{\sum_{n=1}^{N}\delta(v_n > 0)}, \tag{1}$$

where N represents the number of keypoints, d_n denotes the Euclidean distance between predicted and ground truth location for n-th keypoint, k_n refers to the specific weights assigned to each keypoint, and s indicates the scale of an object. v_n is the binary value of $\{0,1\}$ to represent visibility of each point. If the v_n is 0, the point is not utilized in computing the loss value. i and j represent the indices of keypoints and related data points. The δ is a indicator used to include values in the calculation only when a specific condition (e.g., visibility) is true.

The keypoint's confidence score is also trained based on the visibility of keypoints: keypoints that are visible or occluded are assigned a confidence score of 1, while keypoints outside the image or not present are assigned a score of 0. The keypoint confidence loss $\mathcal{L}_{\text{conf}}$ is as follows:

$$\mathcal{L}_{\text{conf}} = -\frac{1}{N}\sum_{n=1}^{N}[v_n \cdot \log(\sigma(c_n)) + (1 - v_n) \cdot \log(1 - \sigma(c_n))], \tag{2}$$

where c_n represents the model output for each keypoint's confidence value. σ represents a sigmoid function.

2.3 Occlusion-Aware Loss

In this study, we propose a new loss function to enhance the performance of keypoint prediction performance in YOLOv8. The loss function is an occlusion-aware loss function $\mathcal{L}_{\text{occ}}$ that utilizes a cross-entropy function, defined as follows:

$$\mathcal{L}_{\text{occ}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{3} t_n^c \log(p_n^c), \tag{3}$$

where t_n^c is c-th element of one-hot encoded vector of n-th keypoint. The value represents our annotated visibility value of $\{0,1,2\}$ already mentioned in Fig. 3. This means that our occlusion-aware loss classifies into three categories. Thanks to $\mathcal{L}_{\text{occ}}$, the network can learn whether the point is visible, occluded, or outside the image. p_n^c denotes the output value produced by the YOLOv8 model. Finally, we utilize total loss function $\mathcal{L}_{\text{total}}$ as follows:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{pose}} + \beta\mathcal{L}_{\text{conf}} + \gamma\mathcal{L}_{\text{occ}}, \tag{4}$$

where α, β and γ are parameters to control the strength of each loss function.

3 Experiments

3.1 Implementation Details

All experiments in this study were implemented using the PyTorch framework and conducted using four NVIDIA RTX A6000 GPUs. The total number of

Table 1. Quantitative comparison between baseline models and our models.

Model	RMSE		
	Visible	Occluded	Average
HRNet [15]	175.634	179.236	176.386
OTPose [16]	136.753	206.270	154.965
YOLOv8-n [2]	18.048	40.746	25.109
YOLOv8-s [2]	17.982	34.113	22.693
YOLOv8-m [2]	17.119	32.774	21.713
YOLOv8-l [2]	16.142	32.319	20.963
YOLOv8-x [2]	16.282	29.843	20.196
YOLOv8-n with $\mathcal{L}_{\text{occ}}$ (ours)	18.031	33.821	22.63
YOLOv8-s with $\mathcal{L}_{\text{occ}}$ (ours)	15.972	32.062	20.761
YOLOv8-m with $\mathcal{L}_{\text{occ}}$ (ours)	16.172	32.606	21.092
YOLOv8-l with $\mathcal{L}_{\text{occ}}$ (ours)	**15.472**	**27.850**	**19.028**
YOLOv8-x with $\mathcal{L}_{\text{occ}}$ (ours)	16.533	31.878	21.045

epochs is 30, and the batch size is 128. For optimization, we used two different optimizers: stochastic gradient descent (SGD) and AdamW. AdamW was utilized for the first $10,000$ iterations, after which the optimizer was switched to SGD for the remaining iterations. We set the parameters α, β, and γ as follows: $\alpha = 12$, $\beta = 0$, and $\gamma = 1$ (See Table 2 why we set β as 0.). We use LambdaLR as a learning rate scheduler. The function is as follows:

$$\lambda(i) = (1 - \frac{i}{e}) \times 0.99 + 0.01, \tag{5}$$

where e and i are the total number of epochs and the current epoch number, respectively.

In this study, two types of data augmentation techniques such as mosaic [13] and mixup [14], were applied to effectively improve the generalization performance of the model. Mosaic augmentation is a process that combines multiple (e.g., four or nine) images into a single composite mosaic image. Mixup creates a new image by overlaying two images while also blending their respective labels.

3.2 Experimental Results

Table 1 shows the evaluation results, in terms of RMSE. To observe the performance under conditions where keypoints are clearly visible and when they are occluded, we conducted experiments in two scenarios: Visible and Occluded.

Based on our experiments, the original YOLOv8 models demonstrated superior performance compared to other models, including HRNet [15] and OTPose [16], which are representative models for human pose estimation tasks. We experimented with various sizes of YOLOv8 models, such as nano (n), small (s), medium (m), large (l), and xlarge (x) models. The YOLOv8-x model achieved a commendable performance of 20.196. In the case of our YOLOv8, we augmented the original YOLOv8 with $\mathcal{L}_{\text{occ}}$, and observed an improvement in performance.

Table 2. Performance comparison according to β and γ.

Model	# of params	β, γ	RMSE		
			Visible	Occluded	Average
YOLOv8-n	3.3M	0, 0	18.048	40.746	25.109
		1, 1	18.136 ∇0.49%	39.759 Δ2.42%	24.80 Δ1.23%
		0, 1	**18.031** Δ0.09%	**33.821** Δ17.00%	**22.63** Δ9.87%
YOLOv8-s	11.6M	0, 0	17.982	34.113	22.693
		1, 1	16.430 Δ8.63%	**31.091** Δ8.86%	21.427 Δ5.58%
		0, 1	**15.972** Δ11.18%	32.062 Δ6.01%	**20.761** Δ8.51%
YOLOv8-m	26.4M	0, 0	17.119	32.774	21.713
		1, 1	16.773 Δ2.02%	**30.773** Δ6.11%	**20.788** Δ4.26%
		0, 1	**16.172** Δ5.53%	32.606 Δ0.51%	21.092 Δ2.86%
YOLOv8-l	44.4M	0, 0	16.142	32.319	20.963
		1, 1	16.177 ∇0.22%	28.722 Δ11.13%	20.323 Δ3.05%
		0, 1	**15.472** Δ4.15%	**27.850** Δ13.83%	**19.028** Δ9.23%
YOLOv8-x	69.4M	0, 0	**16.282**	29.843	**20.196**
		1, 1	16.558 ∇1.70%	**29.463** Δ1.27%	20.241 ∇0.22%
		0, 1	16.533 ∇1.54%	31.878 ∇6.82%	21.045 ∇4.20%

Particularly, there was a significant enhancement in the occluded scenario. On average, our YOLOv8-m model achieved the best performance.

Fig. 4 presents qualitative results. Our method successfully identifies keypoints even in the presence of occlusion. This indicates that our occlusion-aware loss function can aid in detecting keypoints effectively.

Table 3. Performance comparison according to γ.

γ	RMSE		
	Visible	Occluded	Average
0.25	16.452	33.238	21.463
0.5	16.248	30.131	20.292
1	**15.472**	**27.850**	**19.028**
2	16.273	32.082	20.960
3	15.750	31.189	20.334
4	16.702	33.266	21.625
5	17.547	38.509	24.008

Table 4. Occlusion classification test of our YOLOv8 models.

Model	Accuarcy
YOLOv8-n with $\mathcal{L}_{\text{occ}}$	93.269
YOLOv8-s with $\mathcal{L}_{\text{occ}}$	93.612
YOLOv8-m with $\mathcal{L}_{\text{occ}}$	93.763
YOLOv8-l with $\mathcal{L}_{\text{occ}}$	94.109
YOLOv8-x with $\mathcal{L}_{\text{occ}}$	94.340

3.3 Ablation Study

In this section, we conducted experiments to adjust the coefficients controlling the strength of the loss and added tests to classify visibility, allowing us to verify whether $\mathcal{L}_{\text{occ}}$ was effectively learned. Table 2 compares the performance with and without L_{conf} and L_{occ}, corresponding to experiments conducted with different combinations of β and γ being set to either 0 or 1. As indicated in the table, it

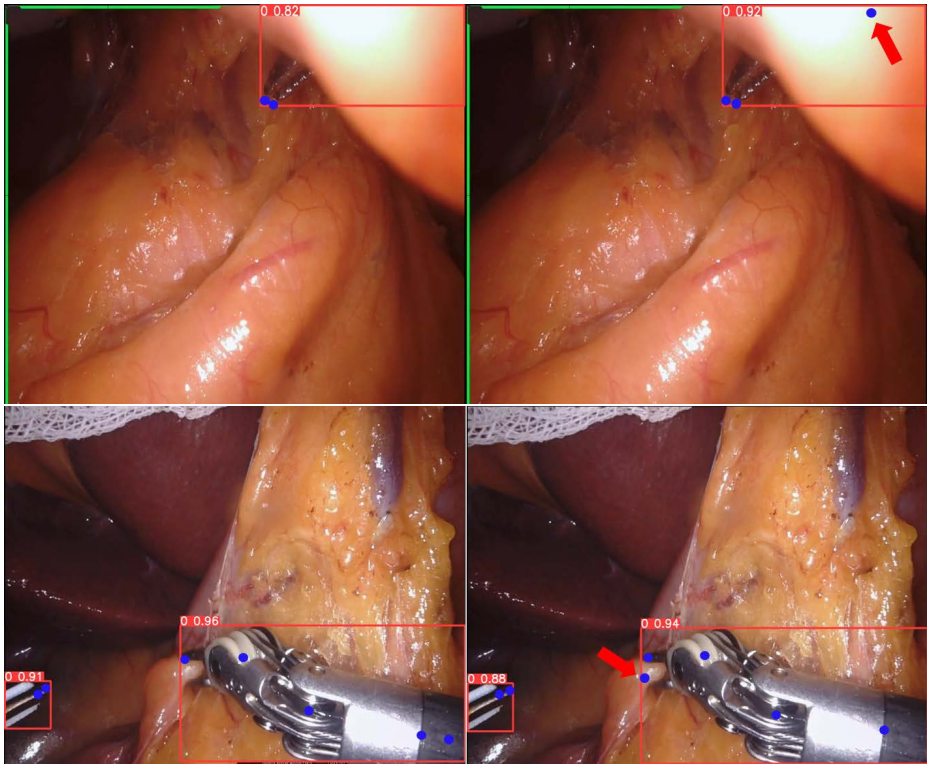

Fig. 4. Qualitative results comparing the original and our YOLOv8 models. Two images in the first column represent the results extracted from the original YOLOv8 model. The second column denotes the results based on our proposed occlusion-aware loss. The red arrow represents enhancement compared to the original YOLOv8 models.

was observed that overall performance mostly improved when both β and γ were set to 1 compared to when they were both set to 0. However, the best results were achieved when β was set to 0 and γ was set to 1 using the YOLOv8-l model. This suggests that our proposed occlusion-aware loss function can indeed replace the traditional confidence loss effectively.

Additionally, we explore the influence of γ by varying from 0.25 to 5. As shown in Table 3, the best RMSE performance is achieved when $\gamma = 1$. Furthermore, to verify whether the model was effectively trained with our occlusion-aware loss, we examined the accuracy of occlusion classification. As shown in Table 4, most models achieved high accuracies, exceeding 93%. This implies that our models are capable of effectively discerning the occlusion status of keypoints. In other words, the feature space learned by YOLOv8 can be considered to include occlusion-aware features.

4 Conclusion

In conclusion, we proposed a new occlusion-aware loss function for accurate pose estimation of robotic surgical instruments. To achieve this, we constructed a new private dataset from real gastric cancer surgery videos. This dataset was specifically designed to capture the dynamic and unpredictable nature of surgical environments. Experimental results demonstrated that our proposed loss function effectively replaces the traditional confidence loss and that our YOLOv8 model performs exceptionally well in various scenarios, particularly in the presence of occlusion. These findings suggest that our model can effectively recognize the occlusion status of keypoints, opening up new possibilities for precise pose estimation in robotic-assisted surgery.

Acknowledgments. This research was equally supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI22C1496) and the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education(No. RS-2023-00241123).

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

References

1. Reddy, Kavyanjali, et al. "Advancements in Robotic Surgery: A Comprehensive Overview of Current Utilizations and Upcoming Frontiers." Cureus 15(12) (2023)
2. Contributors, Jocher, G., Chaurasia, A., Qiu, J., YOLO by Ultralytics : https://github.com/ultralytics/ultralytics. Jan 2023
3. Allan, Max, et al. "Image based surgical instrument pose estimation with multi-class labelling and optical flow." MICCAI, pp. 203-210. (2015)
4. Du, Xiaofei, et al. "Combined 2D and 3D tracking of surgical instruments for minimally invasive and robotic-assisted surgery." International journal of computer assisted radiology and surgery **11**(6), pp. 1109-1119. (2016)
5. Hasan, Md Kamrul, et al. "Detection, segmentation, and 3D pose estimation of surgical tools using convolutional neural networks and algebraic geometry." Medical Image Analysis 70, pp. 101994 (2021)
6. Du, Xiaofei, et al. "Articulated multi-instrument 2-D pose estimation using fully convolutional networks." IEEE transactions on medical imaging 37(5), pp 1276-1287 (2018)
7. Kayhan, Mert, et al. "Deep attention based semi-supervised 2d-pose estimation for surgical instruments." ICPR International Workshops and Challenges, pp. 444-460. (2021)
8. Sznitman, Raphael, et al. "Unified detection and tracking of instruments during retinal microsurgery." IEEE transactions on pattern analysis and machine intelligence 35(5) pp. 1263-1273. (2012)
9. Ye, Menglong, et al. "Real-time 3d tracking of articulated tools for robotic surgery." Medical Image Computing and Computer-Assisted Intervention, pp. 386-394 (2016)

10. Zhou, J., Payandeh, S.: Visual tracking of laparoscopic instruments. Journal of Automation and Control Engineering 2(3) (2014)
11. Rieke, Nicola, et al.: Real-time localization of articulated surgical instruments in retinal microsurgery. Medical image analysis 34, pp. 82–100 (2016)
12. MICCAI 2015 Endoscopic Vision Challenge, https://endovissub-instrument.grand-challenge.org/.
13. Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "Yolov4: Optimal speed and accuracy of object detection." arXiv preprint arXiv:2004.10934 (2020).
14. Zhang, Hongyi, et al. "mixup: Beyond empirical risk minimization." arXiv preprint arXiv:1710.09412 (2017).
15. Sun, Ke, et al. "Deep high-resolution representation learning for human pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2019)
16. Jin, Kyung-Min, Gun-Hee Lee, and Seong-Whan Lee. "OTPose: Occlusion-Aware Transformer for Pose Estimation in Sparsely-Labeled Videos." 2022 IEEE International Conference on Systems, (2022).