Phase-Informed Tool Segmentation for Manual Small-Incision Cataract Surgery

Bhuvan Sachdeva[1,2]*†, Naren Akash[1]*, Tajamul Ashraf[1], Simon Müller[3], Thomas Schultz[3,4], Maximilian W. M. Wintergerst[3], Niharika Singri[2], Kaushik Murali[2]†, and Mohit Jain[1]

[1] Microsoft Research, Bengaluru, India
[2] Sankara Eye Hospital, Bengaluru, India
[3] University of Bonn, Bonn, Germany
[4] Lamarr Institute for Machine Learning and Artificial Intelligence, Bonn, Germany

Abstract. Cataract surgery is the most common surgical procedure globally, with a disproportionately higher burden in developing countries. While automated surgical video analysis has been explored in general surgery, its application to ophthalmic procedures remains limited. Existing research primarily focuses on Phaco cataract surgery, an expensive technique not accessible in regions where cataract treatment is most needed. In contrast, Manual Small-Incision Cataract Surgery (MSICS) is the preferred low-cost alternative in high-volume settings and for complex cases. However, no dataset exists for MSICS. To address this gap, we introduce Sankara-MSICS, the first comprehensive dataset containing 53 surgical videos annotated for 18 surgical phases and 3,527 frames with 13 surgical tools at the pixel level. We also present ToolSeg, a novel framework that enhances tool segmentation with a phase-conditional decoder and a semi-supervised setup leveraging pseudo-labels from foundation models. Our approach significantly improves segmentation performance, achieving a 38.1% increase in mean Dice scores, with notable gains for smaller and less prevalent tools. The code is available at `https://github.com/Sri-Kanchi-Kamakoti-Medical-Trust/ToolSeg`.

Keywords: Segmentation · Cataract Surgery · Pseudo labeling

1 Introduction

Cataract is the leading cause of preventable blindness worldwide, with surgery being the standard treatment. Over 26 million individuals undergo cataract surgery annually [4], making it one of the most common surgeries. Established techniques include Phacoemulsification (Phaco) and Manual Small Incision Cataract Surgery (MSICS). Unlike laparoscopic surgery, where most computer vision methods have been developed [20], cataract surgery pose unique challenges. It involves delicate micro-instruments in a highly reflective ocular environment [14], resulting in specular distortions [3,5]. Additionally, the small

*Equal contribution.

†Corresponding author: b-bsachdeva@microsoft.com and kaushik@sankaraeye.com

instruments cause significant foreground imbalance (e.g., Figure 1 (left)), and the transparent ocular tissues, combined with microscope use, create complex optical conditions, thus complicating image analysis.

With increasing demand, expanding surgical capacity, safety, and efficiency is crucial [6,25]. Automatic (real-time) surgical video analysis can advance surgeon skill development, improve training with targeted feedback, ensure quality control, and detect anomalies [24]. This requires robust temporal and spatial understanding, which rely on accurate phase detection and tool segmentation.

Prior work in computer vision-based surgical analysis has largely focused on Phaco procedures [16]. Despite Phaco being the preferred technique, its reliance on expensive technology and infrastructure limits accessibility in low-and-middle-income countries. Hence, in resource-limited and high-volume settings, MSICS is favored. MSICS is also the preferred method for challenging cases, such as brunescent hard, hypermature, and intumescent cataracts [1]. Despite its prevalence and advantages, MSICS has been largely neglected in the development of datasets for automated surgical video analysis. To date, no public dataset exists specifically for MSICS instrument segmentation [15,16].

In this paper, we introduce *Sankara-MSICS*, the first large-scale dataset on MSICS. It includes 3,527 frames from 53 *in vivo* human cataract surgery videos, annotated with pixel-level labels for 13 surgical tools and corresponding phases across 18 surgical stages. Analysis of Sankara-MSICS reveals a strong correlation between surgical phases and tool presence. Building on this insight, we propose *ToolSeg*, a novel framework that leverages surgical phase information as a prior for tool segmentation. Additionally, we leverage Meta's SAM 2 model [19] to generate pseudo-labels for unlabeled frames, expanding our dataset size from 3,527 to 24,405 frames without additional training. *ToolSeg* outperforms existing methods by 38.1% DSC, with notable improvements in classifying and segmenting less prevalent tools. To validate its generalizability, we applied it to a Phaco dataset, CaDIS [10], and observed significant improvements.

2 The Sankara-MSICS Dataset

The Sankara-MSICS dataset consists of 53 cataract surgery videos recorded at *Sankara Eye Hospital*, *Bangalore, India*, from October 2023 to October 2024. Each video, averaging 15 min 39 s ± 7 min 38 s, was captured at 30 fps with a 1920 x 1080 resolution using a microscope-mounted video camera.

Two resident ophthalmologists at *Sankara Eye Hospital* defined 18 MSICS surgical phases, and annotated all videos with start and stop timestamps for each phase. Frames were uniformly extracted across phases, with additional frames for underrepresented tools to address class imbalance. Resident ophthalmologists then segmented and labeled the frames using a SAM [12]-based annotation tool to minimize manual effort. The final dataset consists of 3,527 frames, annotated for 13 surgical tools (at the pixel-level) and 18 surgical phases.

Table 1 compares Sankara-MSICS with existing cataract surgery datasets. While Sankara-MSICS has fewer manually annotated frames than CaDIS [10],

Table 1: Comparison of Sankara-MSICS with other cataract surgery datasets. Note: CaDIS [10] statistics are as per instrument classes based on Task 2 and 3.

Dataset	**Surgery Type**	**Size**		**Annotations**		**Test Videos**
		Videos	Frames	Tools	Phases	
Sankara-MSICS	MSICS	53	3527	13	18	(7-9/split)
CaDIS [10]	Phaco	25	4670	13	14	3
Cata7 [17]	Phaco	7	2500	10	-	2
InSegCat [7]	Phaco	-	843	11	10	-
Cataract-1K [8]	Phaco	30	2256	9	13	(6/split)

it contains the *highest number of surgical videos* (53), more than twice that of CaDIS (25). The larger number of surgeries increases variability and diversity, which is important for robust model development. The number of test videos in Sankara-MSICS is notably higher than the rest. We also release phase annotations and unannotated frames to support multi-task learning.

3 ToolSeg: Our Proposed Method

3.1 Phase Tool Correlation

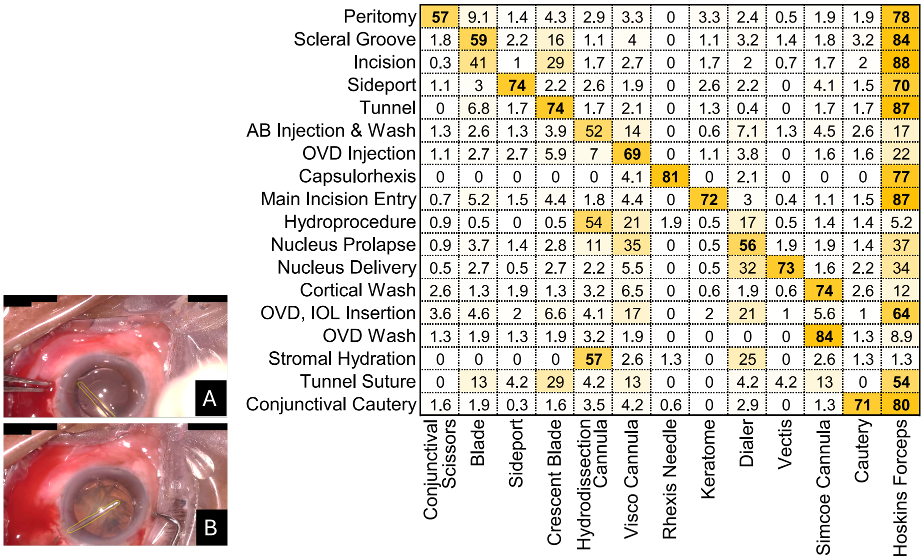
Analysis of Sankara-MSICS reveals a strong correlation between surgical phases and tool presence (Figure 1). For instance, tool Vectis appears exclusively in phase Nucleus Delivery, and Cautery in Conjunctival Cautery. In contrast, tools such as Hoskins Forceps and Crescent Blade are used in multiple phases. Based on these insights, we propose an approach that leverages surgical phase information as a prior for tool segmentation.

3.2 Preliminary: Surgery Phase Recognition

To leverage surgical phase information for tool segmentation, we employ the Multi-Stage Temporal Convolutional Network (MS-TCN++) [13] to predict the phase of each frame. It processes the video at full temporal resolution, ensuring smooth and consistent predictions through a multi-stage design. The initial stage with dual dilated layers generates a preliminary phase prediction, which is iteratively refined by subsequent stages with dilated residual layers. We train the MS-TCN++ model on I3D features from our dataset, achieving an accuracy of 61.5% in phase prediction.

3.3 Phase-Conditioned Segmentation Network

We propose *ToolSeg*, an encoder-decoder architecture for surgical tool segmentation, in which the decoder is conditioned on the surgical phase to improve segmentation accuracy (Figure 2). To achieve this, we introduce the Phase-informed Conditional Decoder (PCD) layer at each decoder level. It consists



	Conjunctival Scissors	Blade	Sideport	Crescent Blade	Hydrodissection Cannula	Visco Cannula	Rhexis Needle	Keratome	Dialer	Vectis	Simcoe Cannula	Cautery	Hoskins Forceps
Peritomy	**57**	9.1	1.4	4.3	2.9	3.3	0	3.3	2.4	0.5	1.9	1.9	**78**
Scleral Groove	1.8	**59**	2.2	16	1.1	4	0	1.1	3.2	1.4	1.8	3.2	**84**
Incision	0.3	41	1	29	1.7	2.7	0	1.7	2	0.7	1.7	2	**88**
Sideport	1.1	3	**74**	2.2	2.6	1.9	0	2.6	2.2	0	4.1	1.5	**70**
Tunnel	0	6.8	1.7	**74**	1.7	2.1	0	1.3	0.4	0	1.7	1.7	**87**
AB Injection & Wash	1.3	2.6	1.3	3.9	52	14	0	0.6	7.1	1.3	4.5	2.6	17
OVD Injection	1.1	2.7	2.7	5.9	7	**69**	0	1.1	3.8	0	1.6	1.6	22
Capsulorhexis	0	0	0	0	0	4.1	**81**	0	2.1	0	0	0	**77**
Main Incision Entry	0.7	5.2	1.5	4.4	1.8	4.4	0	**72**	3	0.4	1.1	1.5	**87**
Hydroprocedure	0.9	0.5	0	0.5	54	21	1.9	0.5	17	0.5	1.4	1.4	5.2
Nucleus Prolapse	0.9	3.7	1.4	2.8	11	35	0	0.5	**56**	1.9	1.9	1.4	37
Nucleus Delivery	0.5	2.7	0.5	2.7	2.2	5.5	0	0.5	32	**73**	1.6	2.2	34
Cortical Wash	2.6	1.3	1.9	1.3	3.2	6.5	0	0.6	1.9	0.6	**74**	2.6	12
OVD, IOL Insertion	3.6	4.6	2	6.6	4.1	17	0	2	21	1	5.6	1	**64**
OVD Wash	1.3	1.9	1.3	1.9	3.2	1.9	0	0	0	0	**84**	1.3	8.9
Stromal Hydration	0	0	0	0	**57**	2.6	1.3	0	25	0	2.6	1.3	1.3
Tunnel Suture	0	13	4.2	29	4.2	13	0	0	4.2	4.2	13	0	**54**
Conjunctival Cautery	1.6	1.9	0.3	1.6	3.5	4.2	0.6	0	2.9	0	1.3	**71**	**80**

Fig. 1: (Left) Despite visual similarity, Image A shows a Hydrodissection Cannula tool (Hydroprocedure phase), while B shows a Dialer (OVD, IOL Insertion phase). *ToolSeg* leverages surgical phase information to accurately classify and segment tools. (Right) Phase-tool co-occurrence matrix.

of three key components: Phase-aware Affine Feature Transform (PAFT), Dynamic Feature Blending Factor (DFBF), and Context-Aware Adaptive Gating (CGate). PAFT modulates feature maps channel-wise, DFBF applies spatial modulation, and CGate combines these modulations. We propose two variants of phase-specific conditioning: **PCD-Basic**, where only PAFT is applied, and **PCD-Gated**, where PAFT is combined with DFBF and CGate to further enhance feature modulation and segmentation accuracy.

To leverage phase-specific information for tool localization, the PAFT module conditions the segmentation network on the predicted surgical phase by learning phase-specific, channel-wise shift and scale embeddings. For each phase p, a pair of learnable embeddings—γ_p (shift) and β_p (scale)—captures phase-related priors, such as phase-tool correlation and tool co-occurrence. These embeddings are applied to each input feature map f as: $f' = \gamma_p \odot f + \beta_p$. This enables the network to adaptively adjust its feature maps *channel-wise* based on the current phase, improving tool segmentation by emphasizing phase-relevant features.

To effectively integrate PAFT-modulated features with spatial features, we introduce a blending factor α, computed based on input feature map f and phase p. It is formulated as: $\alpha = \frac{H(f)+\eta_p}{2}$, where H is a convolutional operation applied to f and η_p is a learnable phase embedding.

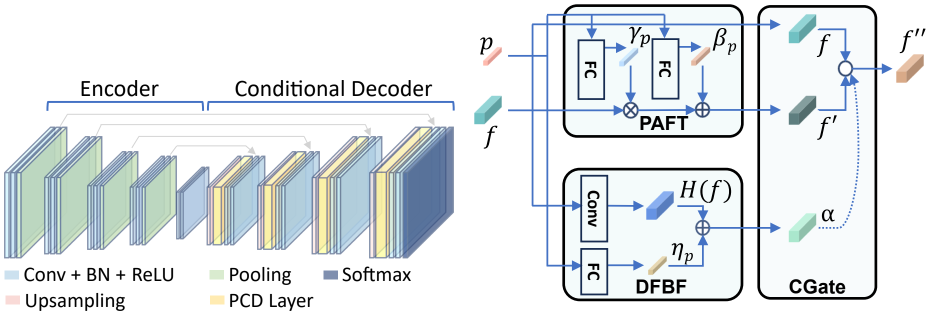


Fig. 2: Overview of ToolSeg architecture (left) and key components of Phase-informed Conditional Decoder layer: PAFT, DFBF, and CGate (right).

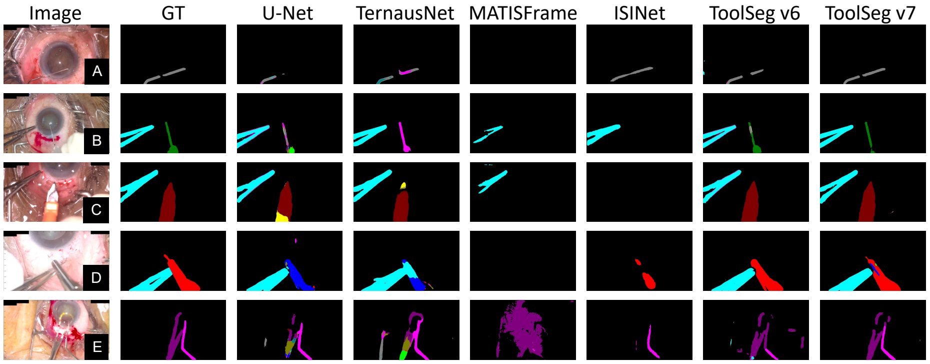


Fig. 3: Qualitative results of *ToolSeg* compared with SOTA methods. (A) Hydrodissection Cannula, (B) Hoskins Forceps, Rhexis Needle, (C) Hoskins Forceps, Keratome, (D) Hoskins Forceps, Blade, and (E) Vectis, Dialer.

Table 2: Impact of *ToolSeg* components on the Sankara-MSICS dataset.

Variant	**Phase Conditioning**	**Phase Source**	**Pseudo Data**	**IoU (m±std)**	**DSC (m±std)**
v0	-	-	✗	40.90 ± 3.4	50.66 ± 3.9
v1	-	-	✓	48.68 ± 4.9	58.29 ± 5.0
v2	PCD-Basic	Predicted MSTCN++	✗	46.58 ± 5.5	55.40 ± 5.8
v3	PCD-Gated	Predicted MSTCN++	✗	48.77 ± 5.5	57.52 ± 5.4
v4	PCD-Gated	Predicted MSTCN++	✓	54.32 ± 4.4	62.70 ± 4.3
v5	PCD-Basic	Ground Truth	✗	54.26 ± 4.6	62.98 ± 4.7
v6	PCD-Gated	Ground Truth	✗	56.13 ± 4.5	64.76 ± 4.5
v7	PCD-Gated	Ground Truth	✓	**61.62 ± 3.8**	**69.96 ± 3.8**

Building on α, we design CGate to fuse the phase-modulated features f' with the original feature map f. The final output feature map is calculated as

$f'' = f' \cdot \alpha + f \cdot (1 - \alpha)$. A higher α emphasizes phase-specific information, whereas a lower α preserves the original spatial information.

3.4 Semi-Supervised Learning with SAM 2

Surgical tool segmentation models typically rely on *sparsely* annotated video datasets, where only a small fraction of frames have ground truth masks. Since precise, clinically relevant annotations require medical-trained professionals, the process is labor-intensive, time-consuming, and costly, leaving most frames unutilized. To address this, we propose a simple yet effective pseudo-label generation method within a semi-supervised learning framework, leveraging SAM 2 foundation model and targeted selection of unlabeled data.

To generate high-quality pseudo-labels, we use Meta's SAM 2 [19], a state-of-the-art interactive foundation model. SAM 2 supports prompt-based segmentation using point inputs and can propagate masks from a seed frame across an entire video, similar to object tracking. Leveraging these capabilities, we apply an iterative prompting strategy followed by mask propagation to generate six additional labelled frames for each existing frame. Thus, we generated 20,878 pseudo-labeled frames from our base dataset of 3,527 annotated frames.

We use a semi-supervised training strategy that combines annotated and pseudo-labeled data. Pseudo labels provide weak supervision, allowing the model to learn from a larger but low-quality dataset. This is followed by fine-tuning on high-quality annotated data for strong supervision, ensuring the model retains alignment with the expert-labeled data.

3.5 Experimental Setup

We use five-fold cross-validation, ensuring each fold has a distinct test set at the video level, while the remaining data is split into training (80%) and validation (20%) sets, with all frames downscaled to 480x270 resolution to optimize computation. Segmentation performance is measured using Intersection over Union (IoU) and Dice Similarity Coefficient (DSC).

Our model is based on a U-Net encoder-decoder architecture with four stages and a bottleneck layer. We train it using the AdamW optimizer with an initial learning rate of $1e - 4$. The experiments run on an NVIDIA A100 GPU with a batch size of 16 for up to 100 epochs, applying early stopping (patience: 10).

4 Results and Analysis

We assess the contributions of each component in our proposed solution by constructing multiple model variants, with results summarized in Table 2. Our baseline U-Net model without phase conditioning or pseudo data (v0) achieves a mean IoU of 40.90 and DSC of 50.66.

(i) Impact of Phase Conditioning. We train the ToolSeg model (v2, v3) with predicted phases from the MS-TCN++ model which yields improvements

Table 3: DSC-based comparison of *ToolSeg* variants for tool segmentation in Sankara-MSICS. Note: Background averages 94.51% pixel occupancy per frame.

Tools	**#Instances**	**v0**	**v3**	**v4**	**v6**	**v7**
Blade	387	46.4	63.4	63.4	69.0	**74.8**
Cautery	278	63.8	60.3	70.9	73.4	**75.5**
Conjunctival Scissors	156	61.9	68.1	75.6	76.0	**82.3**
Crescent Blade	390	67.4	67.4	73.7	76.1	**80.0**
Dialer	343	40.3	41.5	49.7	41.2	**59.0**
Hoskins Forceps	2005	78.3	78.7	82.2	80.6	**84.3**
Hydrodissection Cannula	339	36.5	36.7	49.6	**57.6**	56.18
Keratome	231	46.2	67.9	70.2	75.7	**77.8**
Rhexis Needle	86	14.0	30.3	27.6	38.7	**47.2**
Sideport	240	68.7	63.7	74.9	76.8	**77.6**
Simcoe Cannula	319	46.9	57.3	66.1	65.1	**72.0**
Vectis	152	41.7	44.7	56.2	55.4	**61.8**
Visco Cannula	396	46.4	40.1	55.2	56.3	**61.0**

Table 4: Comparison of *ToolSeg* with SOTA methods on our Sankara-MSICS dataset.

Method	**IoU** (m ± std)	**DSC** (m ± std)
U-Net [21]	40.90 ± 3.4	50.66 ± 3.9
TernausNet [11]	42.76 ± 5.8	52.03 ± 6.0
ISINet [9]	27.55 ± 3.1	37.60 ± 3.7
MATIS Frame [2]	11.41 ± 6.8	18.24 ± 10.1
PAANet [22]	37.67 ± 3.2	46.93 ± 3.7
RAUNet [17]	40.71 ± 4.4	49.62 ± 5.2
HRNetV2 [23]	37.01 ± 6.5	45.33 ± 7.3
ToolSeg v6	**56.13 ± 4.5**	**64.76 ± 4.5**

Table 5: Comparison of *ToolSeg* with SOTA methods on the CaDIS dataset.

Model	**IoU** (m)	**DSC** (m)
U-Net [21]	52.69	62.84
TernausNet [11]	46.47	55.22
ISINet [9]	11.51	15.41
MATISFrame [2]	25.59	34.43
ToolSeg v1	54.65	64.36
ToolSeg v6	**60.73**	**68.63**
ToolSeg v7	59.05	67.72

of 13.9–19.2% in IoU and 9.4–13.6% in DSC. To further establish an upper bound performance we utilize ground truth phase labels, ToolSeg (v5, v6), which yields IoU gains of 32.7–37.2% and DSC gains of 24.3–27.8% over the baseline (v0). The results demonstrate that utilizing phase labels (whether predicted or ground truth) can significantly boost segmentation performance, which is further refined by the proposed gating mechanism. The performance gap between using ground truth and predicted phase labels can be bridged by improving phase prediction models, which is an open research direction and orthogonal to our contribution.

(ii) Effect of Pseudo-labeled Data. Adding pseudo-labeled data alone (v1) improves IoU by 19.0% and DSC by 15.1%, showing the benefit of utilizing otherwise unlabeled video frames. Combining the PCD-Gated model with pseudo-labeled data (v4 and v7) yields substantial gains, increasing IoU by 32.8%-50.7% and DSC by 23.8%-38.1%. This highlights the complementary benefits of phase information and semi-supervised learning.

We evaluate the impact of our semi-supervised setup by training the model with varying proportions of manually annotated data: 25%, 50%, and 100%. Using only 50% of the labeled data along with pseudo-data, our model achieves better performance (IoU: 57.1, DSC: 65.9) than a model trained with 100% labeled data alone (IoU: 56.1, DSC: 64.8). These results demonstrate the effectiveness of our semi-supervised approach in reducing annotation requirements while improving performance.

We benchmark *ToolSeg* against several state-of-the-art models (Table 4), including U-Net [21], TernausNet [11], and ISINet [9], all of which have demonstrated effectiveness in various surgical contexts. *ToolSeg*, with gated phase conditioning and semi-supervised learning, achieves significantly higher performance than all benchmarked models. Among the baselines, TernausNet achieves the highest performance (IoU: 42.8, DSC: 52.0), followed by U-Net (IoU: 40.9, DSC: 52.0). ISINet and MATIS-Frame, which have shown strong performance in gasterointestinal surgery, achieve significantly lower IoU scores of 27.55 and 11.41, respectively. This contrast highlights the unique challenges of ocular surgery, where tools are smaller, often resemble each other, and blend into complex anatomical backgrounds, making segmentation difficult. We do not compare our method with video-based models such as [2,26] since these methods rely on a sequence of frames whereas our method operates on single frames. Additionally, [18] focuses on addressing class imbalance through loss function and sampling optimizations, which is orthogonal to our objective. Therefore, we exclude it from our comparisons.

Tools like the Blade, Keratome, and Rhexis Needle benefit significantly from phase priors (ToolSeg v3/v6 vs v0), as their usage is strongly associated with specific surgical phases (Figure 3). In contrast, tools like the Hoskins Forceps and Dialer show smaller gains due to their presence across multiple phases. Semi-supervised learning further boosts performance, particularly for underrepresented tools, with the Rhexis Needle and Dialer benefiting the most (ToolSeg v6 vs v7). Meanwhile, tools with higher pixel occupancy, such as the Hoskins Forceps, show smaller relative gains. These results suggest that phase priors enhance segmentation for phase-dependent tools, while semi-supervised learning benefits tools with fewer instances or lower pixel presence.

To assess the generalizability of our method, we evaluate it on the CaDIS dataset [10], which focuses on Phaco cataract surgeries and includes 13 tools across 18 surgical phases. Using the best-performing ToolSeg variant (v6), our model achieves a mean IoU of 60.7% and a DSC of 68.6%, significantly outperforming SOTA models, like U-Net, TernausNet and MATIS-Frame by 9.2%, 24.3% and 99.3% in DSC, respectively (Table 5). The semi-supervised setup (v1) alone improves the baseline (v0) by 2.4% in DSC, and incorporating gated phase conditioning with GT phases (v6) provides a substantial 15.3% IoU gain over the baseline. These findings confirm that our approach generalizes well to other surgical datasets, delivering notable performance gains. As ToolSeg focuses solely on tool segmentation, we use only the surgical tool classes from CaDIS Task II, merging anatomy and ‘other’ categories into the background. Thus, our results are not directly comparable to prior work.

5 Conclusion

We present Sankara-MSICS, the first comprehensive dataset on Manual Small-Incision Cataract Surgery, addressing a critical gap in AI-driven surgical video analysis in a widely performed but underexplored procedure. The dataset includes 3,527 frames from 53 videos with phase and tool annotations. Existing tool segmentation models struggle with accurately classifying and segmenting MSICS tools. To address this, we introduce *ToolSeg*, a novel segmentation framework that significantly improves performance by leveraging surgical phase information as a prior. Also, we use SAM 2-based label propagation to expand the dataset to 24,405 frames, reducing manual annotation efforts. We hope this work establishes a solid foundation for future work in surgical tool segmentation, ultimately advancing automated analysis for MSICS and similar procedures.

Acknowledgments. This project is funded by the German Federal Ministry for Economic Cooperation and Development (BMZ) and the Else Kröner-Fresenius-Stiftung (EKFS) and supported by the funding program Hospital Partnerships of the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) (funding line “academic”, contract number 81294577).

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

References

1. Alió, J.L., Dick, H.B., Osher, R.H. (eds.): Cataract surgery. Essentials in Ophthalmology, Springer Nature, Cham, Switzerland, 1 edn. (Jul 2022)
2. Ayobi, N., Pérez-Rondón, A., Rodríguez, S., Arbeláez, P.: Matis: Masked-attention transformers for surgical instrument segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2023)
3. Boldrey, E.E., Ho, B.T., Griffith, R.D.: Retinal burns occurring at cataract extraction. Ophthalmology **91**(11), 1297–1302 (1984)
4. Cicinelli, M.V., Buchan, J.C., Nicholson, M., Varadaraj, V., Khanna, R.C.: Cataracts. The Lancet **401**(10374), 377–389 (2023)
5. Curlin, J., Herman, C.K.: Current state of surgical lighting. The Surgery Journal **6**(02), e87–e97 (2020)
6. Dobson, G.P.: Trauma of major surgery: a global problem that is not going away (2020)
7. Fox, M., Taschwer, M., Schoeffmann, K.: Pixel-based tool segmentation in cataract surgery videos with mask r-cnn. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). pp. 565–568. IEEE (2020)
8. Ghamsarian, N., El-Shabrawi, Y., Nasirihaghighi, S., Putzgruber-Adamitsch, D., Zinkernagel, M., Wolf, S., Schoeffmann, K., Sznitman, R.: Cataract-1k dataset for deep-learning-assisted analysis of cataract surgery videos. Scientific data **11**(1), 373 (2024)
9. González, C., Bravo-Sánchez, L., Arbelaez, P.: Isinet: an instance-based approach for surgical instrument segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 595–605. Springer (2020)

10. Grammatikopoulou, M., Flouty, E., Kadkhodamohammadi, A., Quellec, G., Chow, A., Nehme, J., Luengo, I., Stoyanov, D.: Cadis: Cataract dataset for surgical rgb-image segmentation. Medical Image Analysis **71**, 102053 (2021)
11. Iglovikov, V., Shvets, A.: Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv:1801.05746 (2018)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
13. Li, S., Farha, Y.A., Liu, Y., Cheng, M.M., Gall, J.: Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. IEEE transactions on pattern analysis and machine intelligence **45**(6), 6647–6658 (2020)
14. Meek, K.M., Knupp, C.: Corneal structure and transparency. Progress in retinal and eye research **49**, 1–16 (2015)
15. Mueller, S., Sachdeva, B., Prasad, S.N., Lechtenboehmer, R., Holz, F.G., Finger, R.P., Murali, K., Jain, M., Wintergerst, M.W., Schultz, T.: Phase recognition in manual small-incision cataract surgery with ms-tcn++ on the novel sics-105 dataset. Scientific Reports **15**(1), 1–10 (2025)
16. Müller, S., Jain, M., Sachdeva, B., Shah, P.N., Holz, F.G., Finger, R.P., Murali, K., Wintergerst, M.W., Schultz, T.: Artificial intelligence in cataract surgery: A systematic review. Translational Vision Science & Technology **13**(4), 20–20 (2024)
17. Ni, Z.L., Bian, G.B., Zhou, X.H., Hou, Z.G., Xie, X.L., Wang, C., Zhou, Y.J., Li, R.Q., Li, Z.: Raunet: Residual attention u-net for semantic segmentation of cataract surgical instruments. In: International Conference on Neural Information Processing. pp. 139–149. Springer (2019)
18. Pissas, T., Ravasio, C.S., Da Cruz, L., Bergeles, C.: Effective semantic segmentation in cataract surgery: What matters most? In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24. pp. 509–518. Springer (2021)
19. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
20. Rodrigues, M., Mayo, M., Patros, P.: Surgical tool datasets for machine learning research: a survey. International Journal of Computer Vision **130**(9), 2222–2248 (2022)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
22. Srivastava, A., Chanda, S., Jha, D., Riegler, M.A., Halvorsen, P., Johansen, D., Pal, U.: Paanet: Progressive alternating attention for automatic medical image segmentation. In: 2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART). pp. 1–4. IEEE (2021)
23. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 (2019)
24. Tollefson, M.K., Ross, C.J.: Defining the standard for surgical video deidentification. JAMA surgery **159**(1), 104–105 (2024)
25. Varghese, C., Harrison, E.M., O'Grady, G., Topol, E.J.: Artificial intelligence in surgery. Nature Medicine pp. 1–12 (2024)

26. Zhao, Z., Jin, Y., Heng, P.A.: Trasetr: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery. In: 2022 International conference on robotics and automation (ICRA). pp. 11186–11193. IEEE (2022)