2024 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)
July 15-19, 2024. Boston, MA, USA

Iterative Morphological Training Set Decomposition for Endoscopic Tool Segmentation

Yicheng Zhu, Xiaoyi Wu, Sylvia Tan, Cuiling Sun, Sulagna Saha, Yun-Hsuan Su, and Kevin Huang

***Abstract*— This paper proposes a modified method for training tool segmentation networks for endoscopic images by parsing training images into two disjoint sets: one for rectangular representations of endoscopic images and one for polar. Previous work [1], [2] demonstrated that certain endoscopic images may be better segmented by a U-Net network trained on the original rectangular representation of images alone, and others performed better with polar representations. This work extends that observation to the training images and seeks to intelligently decompose the aggregate training data into disjoint image sets — one ideal for training a network to segment original, rectangular endoscopic images and the other for training a polar segmentation network. The training set decomposition consists of three stages: (1) initial data split and models, (2) image reallocation and transition mechanisms with retraining, and (3) evaluation. In (2), two separate frameworks for parsing polar vs. rectangular training images were investigated, with three switching metrics utilized in both. Experiments comparatively evaluated the segmentation performance (via Sørenson Dice coefficient) of the in-group and out-of-group images between the set-decomposed models. Results are encouraging, showing improved aggregate in-group Dice scores as well as image sets trending towards convergence.**

***Index Terms*— tool segmentation; endoscopy; image processing; robot-assisted minimally invasive surgery; U-Net**

I. Introduction

Laparoscopic keyhole surgery presents many important benefits over open surgery, including but not limited to decreased patient pain and healing time. With the incorporation of a teleoperated, highly-articulated surigcal robot, additional benefits can be realized. These include potential for remote operations, improved surgeon control (e.g. scaling of surgeon motions), and intelligent augmentations such as jitter reduction to name a few. However, perception and situational awareness can be compromised due to constrained fields of view, the dynamic surgical scene, combined with lack of realistic force feedback. It is envisioned that computer vision may be able to help remedy several of these drawbacks, including providing haptic feedback. The first step towards that end is separation of background tissue pixels from tool pixels, i.e. semantic image/tool segmentation [3].

A. Related Work

1) Tool Segmentation: Tool segmentation serves a fundamental role in RMIS [4]. It provides basic information to support robot automation and scene recognition in surgery [5]. Surgical instrument tracking and segmentation, which could inform precise navigation, are essential steps in computer-assisted surgical systems [6], and potentially inform haptic feedback, which may optimally guide surgeons in minimally invasive surgery [7], [8]. Previous studies on surgical telemanipulators conclude that the lack of haptic (especially tactile) feedback is one of the major limitations of computer-assisted surgical systems [9]. Accurate isolation of tools and correct positioning stands to greatly improve performance in robot-assisted operating rooms. Traditional approaches first transform input images into a more complex feature space that considers both color and texture, then apply feature extraction and selection to segment surgical instruments from surrounding tissue [10]. Robot kinematics can also provide a useful prior for segmentation approaches [11], [12]. A widely adopted segmentation network used in the biomedical imaging field is the U-Net [13]–[15].

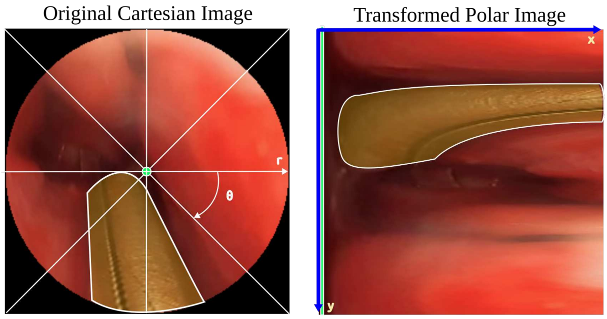

Fig. 1. A side-by-side comparison of a sample endoscopic image (left) and its polar approximation (right) about the image center (green dot). The white outline shows the ground truth border of the surgical tool. Note that the target shape post transformation more closely resembles a rectangle.

2) Morphological Transform for Tool Segmentation: The authors' previous work investigated spatially rearranged endoscopic image data via a polar transform approximation [1]. This approach is based on the observation that rigid and straight laparoscopic tools, which appear as wedges under perspective projection, typically end near the image center where tool-tissue interaction is most likely to occur. (see Fig. 1), and that image segmentation kernels are rectangular in shape. Other biomedical segmentation methods have also utilized a polar transformation, for example, for isolating the optic disk from retinal imaging [16], [17].

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2101107. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSF.

Yicheng Zhu, Xiaoyi Wu, Sylvia Tan and Kevin Huang are with Smith College, Picker Engineering Program, 100 Green Street Northampton, MA 01063 USA {yzhu19, cwu78, stan, kevinhuang}@smith.edu

Cuiling Sun, Sulagna Saha and Yun-Hsuan Su are with Mount Holyoke College, Department of Computer Science, 50 College St, South Hadley, MA 01075 USA {sun25c, saha23s, msu}@mtholyoke.edu

979-8-3503-9154-1/24/$31.00 ©2024 IEEE

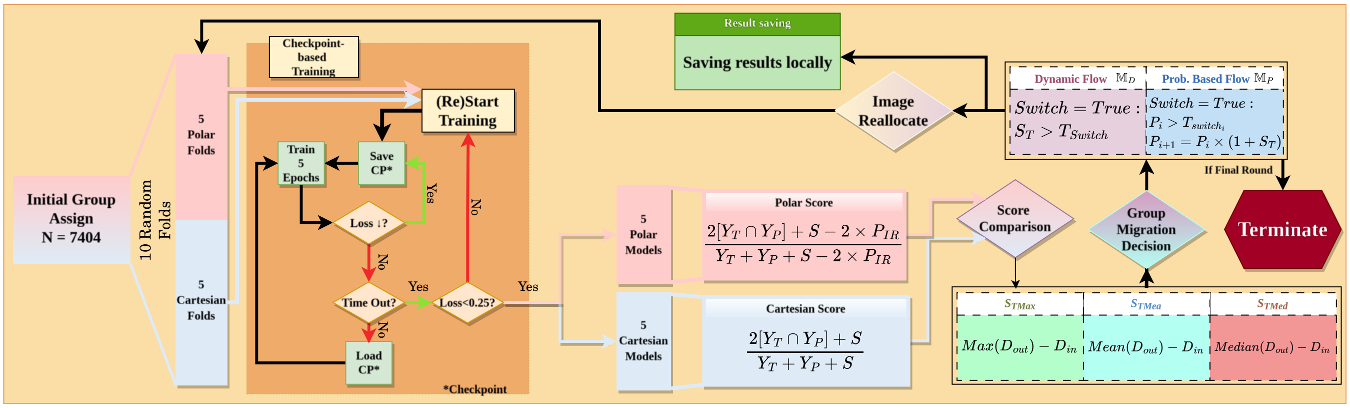

Fig. 2. An overview of the overall workflow of the experiment including the initial group allocation, checkpoint-based training, strategic score generation, switch tendency selection, and migration framework selection.

In a follow-up study by the authors, a novel transformation was proposed that allowed the polar transform center to deviate from the image center [2], and automated tool-tip and vanishing point detection methods were used to generate suitable transform centers. Previous research found that some images were better recognized by models trained on polar-arranged images, despite both polar and rectangular models being trained on the same datasets. It is hypothesized that if certain testing images are more suited to one model type, training images should also be similarly categorized. Decomposing the aggregate training dataset into two disjoint sets for training polar and rectangular models is a major modification of the authors' previous methods and is the core motivation for the research presented here.

B. Contributions

This paper presents and evaluates a novel method for decomposing endoscopic image datasets into two categories for the purposes of training tool segmentation networks. As shown in Fig. 2, the separation aims to enhance the segmentation performance by training two separate networks: (1) using polar-approximation and (2) original rectangular image spatial representations. To the best of the authors' knowledge, this work is the first to present simultaneously:

i) a method for iteratively decomposing training images into two distinct sets for polar and rectangular image segmentation networks;
ii) two distinct image migration frameworks - (1) Dynamic Flow $\mathbb{M}_D$, and (2) Probability-Based Flow $\mathbb{M}_P$ - for transferring or parsing images between iterations;
iii) a comparative evaluation between the various models trained with the aforementioned decomposed data sets.

II. Methods

A. System Hardware

All training and evaluation of segmentation networks were implemented on a machine equipped with an Intel Core i9 24-core processor with 64GB DDR4 RAM, NVIDIA GeForce RTX 4090 graphics card, and running Ubuntu 20.04.3, 64-bit operating system. Both training and testing were executed using hardware acceleration with GPU-runtime as specified by the system hardware to improve training speed.

B. Dataset

The data used in this work were obtained from the University of Washington Sinus Surgery Cadaver/Live Data set [18], [19]. Images were obtained using the Karl Storz Hopkins 4mm 0°endoscope and Stryker 1088 high definition camera at 30 frames per second. Images were captured from real surgeries (live and cadaveric) with smoke, blood, occlusions, reflections, motion blur, and other features of real endoscopic imaging. Furthermore, the dataset is labeled, with binary masks of tool pixels manually annotated by an expert. Images are 256 × 256 pixels with 8-bit depth in three color channels, i.e., full color. A total of 7404 images comprise the datatset.

C. Polar and Cartesian/Rectangular Models

All the training data were split randomly into two groups to initialize the process. These groups will be denoted:

1) G_{PD_i}, the set of images for training the model to segment polar formatted images in iteration i;
2) G_{CD_i}, the images for training the model to segment rectangular/Cartesian formatted images in iteration i.

The size of the initial split is 3702, which is half of the size of the entire training set. Thus, to start

$$|G_{PD_0}| = |G_{CD_0}| = 3702$$

As iterations progress, the set cardinalities may begin to vary with images potentially switching training groups.

1) Group Based K-Folds: Polar models are trained on images and labels in polar representations, as described in [2]; Cartesian models are trained on images and labels in the original rectangular format. After each round of training (elaborated in Sec.II-C.3), the input groups G_{PD} and G_{CD} are reassessed based on the implemented migration framework (Sec.II-E).

To quantify the performance of images in their own category (G_{PD_i} or G_{CD_i}) while reducing bias, a 5-fold cross-validation method was implemented. The five folds were assigned randomly. Therefore, the polar (G_{PD}) and Cartesian (G_{CD}) image-label pairs were each used to train five models. Together, 10 independent models per iteration were trained. Each image appears as part of the training set in four out of five models within its own category (in-group.

2) U-Net and Hyperparameters: The U-Net image segmentation networks implemented in this experiment were trained using the dice coefficient loss function, D_L. Suppose Y_T and Y_P are the tool pixel counts in ground truth and segmentation prediction images respectively. The dice coefficient loss is computed as:

$$D_L = 1 - \frac{2[Y_T \cap Y_P] + S}{Y_T + Y_P + S} \tag{1}$$

, where S is a smoothing component = 1 to prevent dividing by zero [2]. The learning rate was set to 1×10^{-5}, and an Adam optimizer was used.

The models with the same input type were trained using the same hyperparameters and evaluated against the same labels. Different augmentation parameters were used in the Keras API depending on the model group type (G_{PD} or G_{CD}). For models G_{PD} type models,occasional horizontal and vertical flips are allowed; whereas G_{CD} models enabled an additional random rotation up to 360 degrees.

3) Checkpoint-based Training: The authors implemented a dynamic training method in response to challenges achieving consistent convergence using traditional methods with a fixed number of epochs. This method involves training for five epochs per round (each with 100 steps) and monitoring the loss to save the best-performing epoch as a checkpoint. If a checkpoint is saved, training continues; otherwise, the last checkpoint's weights are used. Training restarts with new weights if there is no improvement after three rounds and 50 attempts. The process ends after 20 rounds.

D. Evaluation

Each image was used to train four out of 10 models, making it a valid test image for the remaining six models. Consequently, each image generates six dice scores when evaluated against these models. For instance, an image from G_{CD} will be tested against the single in-group model not trained with that image, and the five out-of-group models from G_{PD}. This results in six separate predictions per image: one in-group and five out-of-group. As illustrated in Fig.3, the in-group score is denoted D_{in} and the five out-of-group scores are lumped into the variable D_{out}.

Unlike during training, G_{PD} predictions in the evaluation phase are transformed back to rectangular-representation space [2] (this process is not lossless) to compare against the Cartesian ground truth. Furthermore, since endoscopic images have circular shaped image content, the black border pixels of the rectangular images are not considered. Therefore, a modified dice score D_{Mod} is generated only taking into account the relevant pixels:

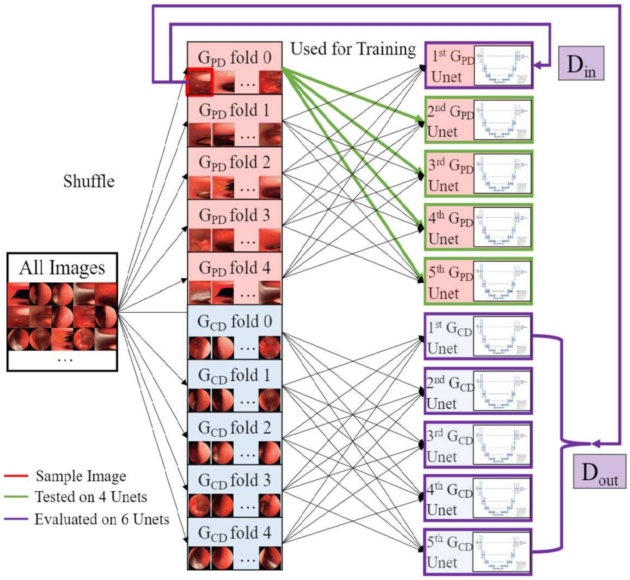

Fig. 3. The generation of in-group scores and out-of-group scores for a sample image (marked in red) in fold 0 of, for example, G_{PD}. Green arrows indicate that the sample image is used as a training image for the indicated model.

$$D_{Mod} = \frac{2[Y_T \cap Y_P] + S - 2 \times P_{IR}}{Y_T + Y_P + S - 2 \times P_{IR}} \tag{2}$$

, where $P_{IR} = 14616$ is the number of irrelevant pixels. The final dice score is calculated as the average dice score of foreground and background.

E. Migration Frameworks

After each training round, the goal is to determine which, if any images should switch from G_{PD} to G_{CD}, or vice-versa, for the next training round. First, a switching score S_T will be calculated for each image based on the six evaluation dice scores for that image. Three algorithmic variations of S_T were explored:

$$S_{T_{Max}} = \max(D_{out}) - D_{in} \tag{3}$$

$$S_{T_{Mea}} = \text{mean}(D_{out}) - D_{in} \tag{4}$$

$$S_{T_{Med}} = \text{median}(D_{out}) - D_{in} \tag{5}$$

The intention is that an image is more likely to switch groups if S_T is high.

To achieve this, two migration frameworks were investigated, each imposing a slightly different constraint on the image transfers between G_{PD} and G_{CD}. Note that for both frameworks, each of the three S_T variations (Eq.3-5) were tested.

1) Dynamic Flow ($\mathbb{M}_D$)*:* A heuristically tuned threshold $T_{\text{switch}} \in [0, 1]$ is applied to S_T for each image. Since there is no constraint on the number of candidate images pending migration from G_{PD} to G_{CD} or vice versa, $|G_{PD}|$ and $|G_{CD}|$ can vary.

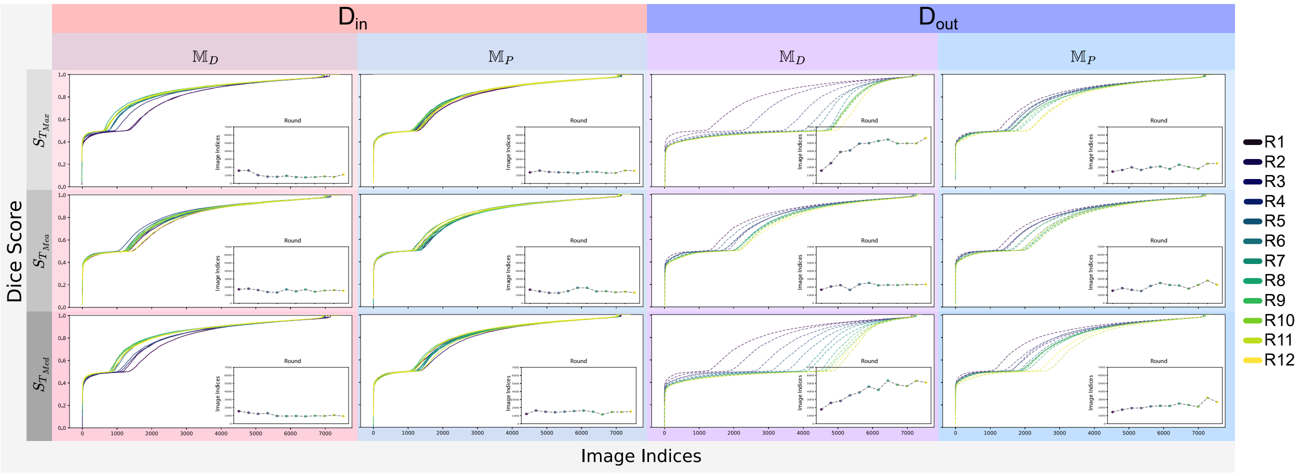

Fig. 4. The sorted in-group (D_{in}) and out-of-group (D_{out}) Dice scores for each proposed migration framework across the 12 total training rounds. The inserted plots in bottom right of each subgraph indicate the image indices corresponding to a large dice performance jump in each training round.

2) Probability Based Flow ($\mathbb{M}_P$): In this framework, each image has a migration probability P_i, where i denotes the round index, and $P_0 = 0.5$. After each training round, P_i is updated to P_{i+1} based on equation (6).

$$P_{i+1} = P_i \times (1 + S_T) \tag{6}$$

At the end of the i^{th} training round, the image has a P_i probability of migrating groups. Note that as opposed to $\mathbb{M}_D$, which utilizes a hard threshold for switching groups, $\mathbb{M}_P$ simply updates the transition probability of each image. This highlights the stochastic nature of this framework.

III. Results and Discussions

A. Aggregate Performance

The overall performance of the training can be summarized by the Area Under the Curve (AUC) of D_{in} over the 12 training rounds examining the plots of Dice Score vs. Image Index, as shown in Fig. 4. The in-group analysis of maximum normalized AUC is shown in Table I, where the AUC is calculated by summing all 7404 in-group scores and dividing by the total area 7404, as shown in (7). These scores correspond to the D_{in} plots shown in Fig. 4.

$$AUC = \frac{\Sigma D_{in}}{N \times 1} \tag{7}$$

TABLE I

Overall Performance

Maximum AUC of in-group scores

		S_T	$S_{T_{Max}}$	$S_{T_{Mea}}$	$S_{T_{Med}}$
Framework	Method	$\mathbb{M}_D$	0.851630460	0.816038652	0.844494048
		$\mathbb{M}_P$	0.817996301	0.812301844	0.813608189

Table I shows that $\mathbb{M}_D$ using switching score $S_{T_{Max}}$ generates the best overall grouping performance after 12 roundsm while and the $\mathbb{M}_P$ using switching score $S_{T_{Mea}}$ results in the worst AUC. Within the same migration framework, $S_{T_{Mea}}$ overall produced the least favorable Dice scores, with this observation being most noticeable in framework $\mathbb{M}_D$.

B. Segmentation Performance Progression

Figure 4 shows the D_{in} (left two columns) and D_{out} (right two columns) of all images sorted in ascending order. The iteration rounds R1-R12 are color-coded from purple to yellow. Meanwhile, D_{in} improvements of G_{PD} and G_{CD} images between the start and end of the checkpoint-based training process are illustrated in Fig.5. Lastly, as shown in Fig.7, D_{in} and the spread/median of D_{out} is compared for each of the two migration frameworks.

1) Evolution of D_{in} and D_{out}: Proper sorting of image groups, G_{PD}, G_{CD}, is hypothesized to result in increased in-group score D_{in} and simultaneously decreased out-of-group performance, D_{out}. This is generally observed in Fig. 4, with the stark exception of framework $\mathbb{M}_D$ using switching score $S_{T_{Mea}}$. Framework $\mathbb{M}_D$ and score $S_{T_{Max}}$ exhibited this behavior the most. Generally, framework $\mathbb{M}_D$ demonstrated the desired D_{in} and D_{out} evolution more than $\mathbb{M}_P$. Finally, as illustrated in the inset graphs of Fig. 4, the best performed combinations, $\mathbb{M}_{D\ max/med}$, show the widest spread across training rounds at which the D_{out} performance jumps occur.

2) Group-Specific D_{in} Improvements: Figure 5 shows both initial (data points) as well as 12th round (trace) dice scores, and this delta is shown with grey bars. From this, it is observed that G_{PD} images showed general in-group improvement over time, as shown by the length and density of vertical lines. This contrasts with the sparseness of grey bars in the G_{CD} plots, Meanwhile, both G_{PD} and G_{CD} exhibit tremendous in-group improvement in the best 12th round performing images. Interestingly, framework $\mathbb{M}_D$ using switching score $S_{T_{Mea}}$ showed sparser D_{in} improvements compared to its G_{PD} counterparts.

3) Out-of-Group Score Variation: Figure 7 shows the variation of D_{out} in the 12th and final round. Framework

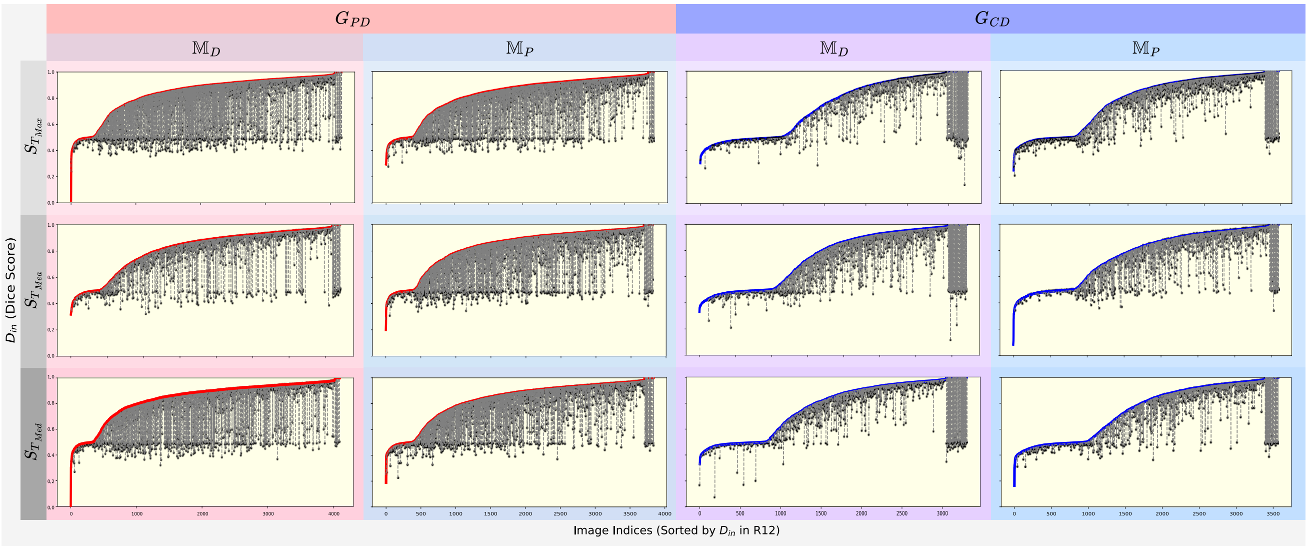

Fig. 5. D_{in} improvements between the first round (R1, marked as black dots) and the last round (R12, marked as colored traces). Note that results from the G_{PD} and G_{CD} images are respectively shown as red and blue.

$\mathbb{M}_D$ using switching scores $S_{T_{Max}}$ and $S_{T_{Med}}$ showed the least variation of D_{out} in the last round. These conditions also yielded both the least overall D_{out} score, yet the greatest difference between in-group and out-of-group performance.

C. Image Group Migration Dynamics

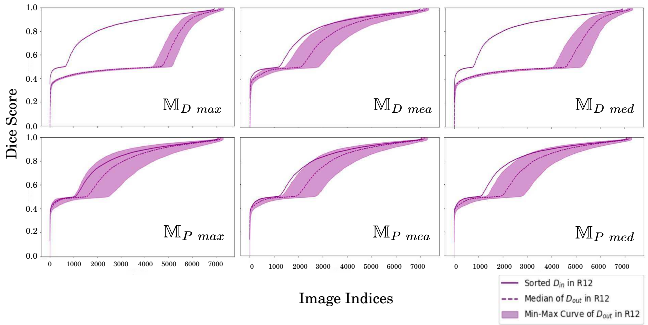

Fig. 7. Comparisons of the sorted (ascending) D_{in}, the sorted median of D_{out}, and its spread for each migration algorithm.

The group image count and D_{in} performances of each group (G_{PD} and G_{CD}) are visualized as a heatmap in Fig.6. Furthermore, a Sankey diagram, as shown in Fig.8, illustrates inter-group image transfers.

1) Group Heatmaps: Examining Fig. 6, for all algorithms except framework $\mathbb{M}_D$ using $S_{T_{Mea}}$, a decreasing trend in $|G_{CD}|$ and complementary increase in $|G_{PD}|$ was observed. Among them, $\mathbb{M}_D$ with $S_{T_{Max}}$ and $S_{T_{Med}}$ show a prodigious migration of images between rounds. In contrast, $\mathbb{M}_D$ with $S_{T_{Mea}}$ barely exhibits group size variation.

2) Image Migration: Considering the Sankey diagram, framework $\mathbb{M}_D$ with $S_{T_{Mea}}$ shows the least image transfer across all methods. The stochastic nature of $\mathbb{M}_P$ could account for the more prevalent image flow between groups as compared to $\mathbb{M}_D$. With that said, $\mathbb{M}_D$ exhibits a general decrease in image transfer as rounds progress (most obvious in $S_{T_{Max}}$ and $S_{T_{Med}}$), suggesting eventual convergence.

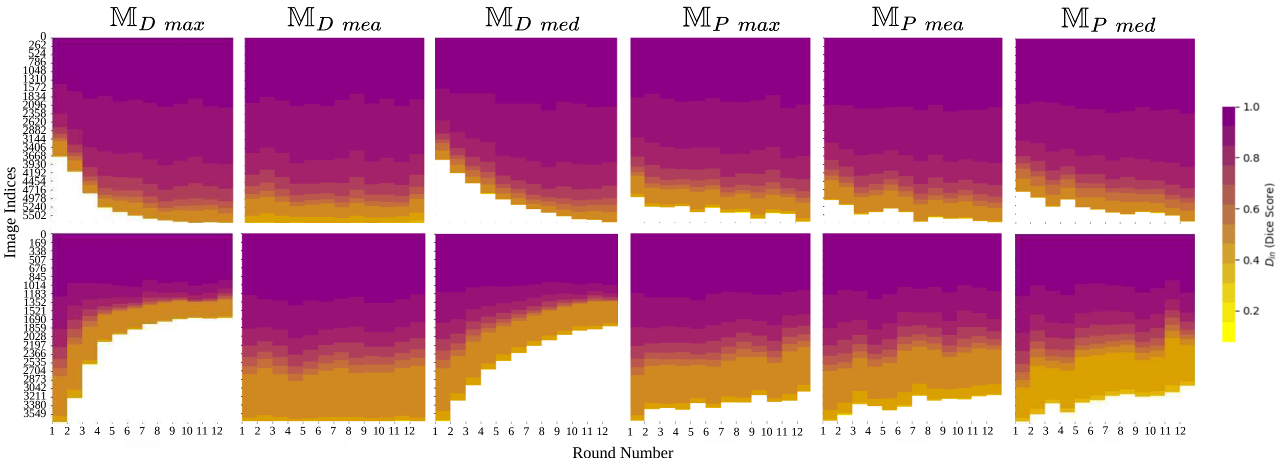

Fig. 6. Grouped D_{in} heatmaps across the 12 training rounds.

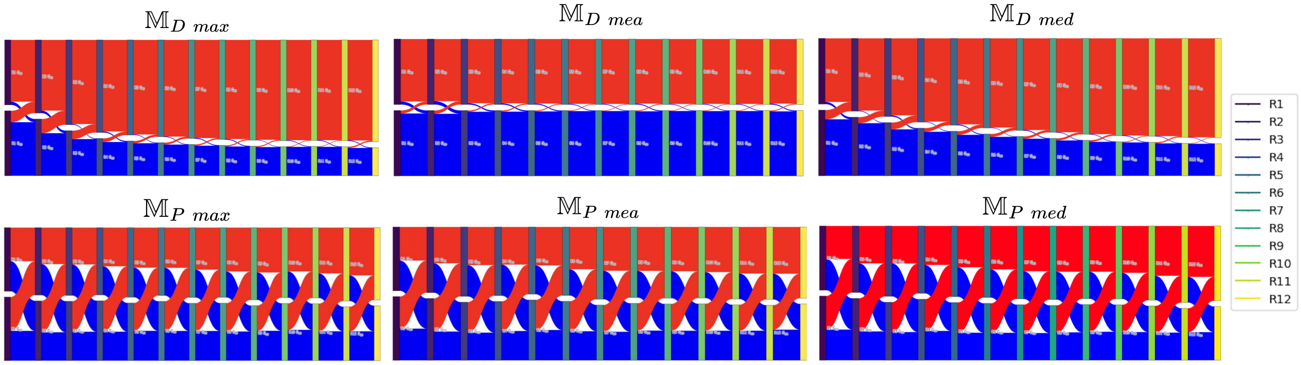

Fig. 8. The image flow between G_{PD} (red) and G_{CD} (blue) across training rounds. The round numbers R1-R12 are color-coded from purple to yellow.

IV. Conclusion

This paper proposes a novel method for decomposing endoscopic image datasets for tool segmentation to create specialized training sets for polar and rectangular image representations. It explores two migration frameworks, $\mathbb{M}_D$ (Dynamic Flow) and $\mathbb{M}_P$ (Probability-Based Flow), along with three variations of switching scores to determine the ideal method for transferring images between training groups.

Findings indicate that the combination of $\mathbb{M}_D$ with either the maximum, $S_{T_{Max}}$, or median, $S_{T_{Med}}$, switching score calculation method delivers superior performance. These configurations exhibited:

- increased aggregate performances in-group scores;
- substantial images migration between G_{PD} and G_{CD} groups, along with distinct differences between in-group and out-of-group scores;
- decreasing rates of image transfer over rounds, suggesting potential convergence.

In contrast, switching score $S_{T_{Mea}}$ performed least favorably, particularly when combined with $\mathbb{M}_D$. This work highlights the potential of tailoring training sets to the distinct spatial representations within endoscopic images, and could enable improved segmentation performance for robotic-assisted endoscopic procedures.

References

[1] K. Huang, D. Chitrakar, W. Jiang, and Y.-H. Su, "Enhanced u-net tool segmentation using hybrid coordinate representations of endoscopic images," in *2021 International Symposium on Medical Robotics (ISMR)*. IEEE, 2021, pp. 1–7.

[2] K. Huang, D. Chitrakar, W. Jiang, I. Yung, and Y.-H. Su, "Surgical tool segmentation with pose-informed morphological polar transform of endoscopic images," *Journal of Medical Robotics Research*, vol. 7, no. 02n03, p. 2241003, 2022.

[3] N. Haouchine, W. Kuang, S. Cotin, and M. Yip, "Vision-based force feedback estimation for robot-assisted surgery using instrument-constrained biomechanical three-dimensional maps," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2160–2165, 2018.

[4] D. Pakhomov, V. Premachandran, M. Allan, M. Azizian, and N. Navab, "Deep residual learning for instrument segmentation in robotic surgery," in *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*. Springer, 2019, pp. 566–573.

[5] E. Colleoni, P. Edwards, and D. Stoyanov, "Synthetic and real inputs for tool segmentation in robotic surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 700–710.

[6] S. Nema and L. Vachhani, "Surgical instrument detection and tracking technologies: Automating dataset labeling for surgical skill assessment," *Frontiers in Robotics and AI*, vol. 9, p. 1030846, 2022.

[7] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele, and P. Jannin, "Detecting surgical tools by modelling local appearance and global shape," *IEEE transactions on medical imaging*, vol. 34, no. 12, pp. 2603–2617, 2015.

[8] K. Huang, D. Chitrakar, R. Mitra, D. Subedi, and Y.-H. Su, "Characterizing limits of vision-based force feedback in simulated surgical tool-tissue interaction," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 4903–4908.

[9] C. Freschi, V. Ferrari, F. Melfi, M. Ferrari, F. Mosca, and A. Cuschieri, "Technical review of the da vinci surgical telemanipulator," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 9, no. 4, pp. 396–406, 2013.

[10] L. C. Garcia-Peraza-Herrera, W. Li, L. Fidon, C. Gruijthuijsen, A. Devreker, G. Attilakos, J. Deprest, E. Vander Poorten, D. Stoyanov, T. Vercauteren, *et al.*, "Toolnet: holistically-nested real-time segmentation of robotic surgical tools," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5717–5722.

[11] Y.-H. Su, I. Huang, K. Huang, and B. Hannaford, "Comparison of 3d surgical tool segmentation procedures with robot kinematics prior," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4411–4418.

[12] Y.-H. Su, K. Huang, and B. Hannaford, "Real-time vision-based surgical tool segmentation with robot kinematics prior," in *2018 International Symposium on Medical Robotics (ISMR)*. IEEE, 2018, pp. 1–6.

[13] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: A review of theory and applications," *Ieee Access*, vol. 9, pp. 82 031–82 057, 2021.

[14] G. Du, X. Cao, J. Liang, X. Chen, and Y. Zhan, "Medical image segmentation based on u-net: A review," *Journal of Imaging Science and Technology*, 2020.

[15] O. Ronneberger, P. Fischer, and T. Brox, "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*. Springer, 2015, pp. 234–241.

[16] M. N. Zahoor and M. M. Fraz, "Fast optic disc segmentation in retina using polar transform," *IEEE Access*, vol. 5, pp. 12 293–12 300, 2017.

[17] ——, "A correction to the article "fast optic disc segmentation in retina using polar transform"," *IEEE Access*, vol. 6, pp. 4845–4849, 2018.

[18] S. Lin, F. Qin, R. A. Bly, K. S. Moe, and B. Hannaford, "Uw sinus surgery cadaver/live dataset (uw-sinus-surgery-c/l)," 2020.

[19] F. Qin, S. Lin, Y. Li, R. A. Bly, K. S. Moe, and B. Hannaford, "Towards better surgical instrument segmentation in endoscopic vision: Multi-angle feature aggregation and contour supervision," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6639–6646, 2020.