Enhanced U-Net Tool Segmentation using Hybrid Coordinate Representations of Endoscopic Images

Kevin Huang, Digesh Chitrakar, Wenfan Jiang, and Yun-Hsuan Su

2021 International Symposium on Medical Robotics (ISMR) | 978-1-6654-0622-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISMR48346.2021.9661519

***Abstract*— This paper presents an approach to enhanced endoscopic tool segmentation combining separate pathways utilizing input images in two different coordinate representations. The proposed method examines U-Net convolutional neural networks with input endoscopic images represented via (1) the original rectangular coordinate format alongside (2) a morphological polar coordinate transformation. To maximize information and the breadth of the endoscope frustrum, imaging sensors are oftentimes larger than the image circle. This results in unused border regions. Ideally, the region of interest is proximal to the image center. The above two observations formed the basis for the morphological polar transformation pathway as an augmentation to typical rectangular input image representations. Results indicate that neither of the two investigated coordinate representations consistently yielded better segmentation performance as compared to the other. Improved segmentation can be achieved with a hybrid approach that carefully selects which of the two pathways to be used for individual input images. Towards that end, two binary classifiers were trained to identify, given an input endoscopic image, which of the two coordinate representation segmentation pathways (rectangular or polar), would result in better segmentation performance. Results are promising and suggest marked improvements using a hybrid pathway selection approach compared to either alone. The experiment used to evaluate the proposed hybrid method utilized a dataset consisting of 8360 endoscopic images from real surgery and evaluated segmentation performance with Dice coefficient and Intersection over Union. The results suggest that on-the-fly polar transformation for tool segmentation is useful when paired with the proposed hybrid tool-segmentation approach.**

***Index Terms*— robot-assisted minimally invasive surgery; telesurgery; surgical tool segmentation; U-Net**

I. Introduction

Robot-assisted minimally invasive surgery (RMIS) exhibits several salient patient-side benefits over open surgery, including reduction in pain, recovery time and medication. While benefits also exist for the surgical operator, perception and situational awareness can be improved. Oftentimes, visual feedback is available through the use of endoscopes. However, limited field of view, occlusions, and lack of realistic force feedback while operating in a dynamic environment reduce scene and task understanding. Computer vision is a promising pathway for remedying several of these areas.

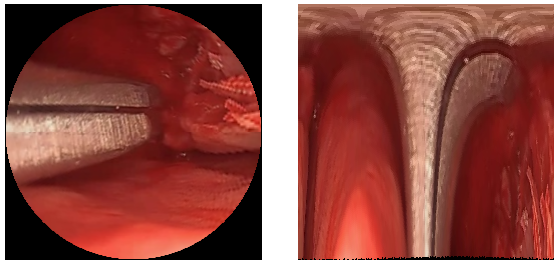
Kevin Huang and Digesh Chitrakar are with Trinity College, Department of Engineering, 300 Summit Street, Hartford, CT 06106 USA {kevin.huang,digesh.chitrakar}@trincoll.edu

Wenfan Jiang and Yun-Hsuan Su are with Mount Holyoke College, Department of Computer Science, 50 College Street, South Hadley, MA 01075 USA {jiang24w, msu}@mtholyoke.edu

This material is based upon work supported by the National Science Foundation under Grant 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 the National Science Foundation.

A. Background and Related Work

1) Endoscopic Tool Segmentation: Intraoperative endoscopy is often used in concert with laparoscopic procedures and has been demonstrated to reduce complications [1], [2]. When it comes to RMIS, endoscopy is often the assumed and most baseline form of visual feedback [3]–[5]. With that said, several approaches consider the use of multicamera systems in RMIS [6]–[9]. Accurate tool segmentation can be used to assist in tool tracking and guidance [10] and be used for vision-based force estimation [11]–[14]. Myriad machine learning techniques have been investigated for segmenting tool pixels from tissue pixels in endoscopic images [15]–[18]. The use of U-Net for image segmentation is of particular interest [5]. Improving tool segmentation accuracy is an important and popular field of research.



(a) (b)

Fig. 1: (a) Typical rectangular endoscopic image from an RMIS procedure with content presented in a circular shape, called the image circle (b) the polar transformed version of the circular image.

The work presented here aims to expand machine learning-based segmentation approaches by leveraging observations related to the morphological structure of ideal endoscopic imaging from RMIS procedures. Firstly, the content from endoscopic cameras is limited to a circular area since the image sensor is designed to typically be larger than the image circle of the endoscope [19]. The remainder of the image consists of zero padding to retain a rectangular shape. Secondly, ideally the region of interest (ROI) of a given surgical operation remains in the center of the field of view, and tool-tissue interactions occur at the ROI.

With the above two considerations, the use of a preprocessing morphological step transforming circular endoscopic images into polar coordinates centered about the image circle center is proposed as a means to provide potentially more amenable image representation for segmentation. Figure 1 shows (a) a sample endoscopic image from the University of Washington Sinus Surgery Cadaver/Live Dataset [20] and (b) its polar representation.

Figure 2 illustrates that with perspective projection and straight surgical tools that polar transformed endoscopic images may be more suitable for rectangular kernels used in most image segmentation networks.

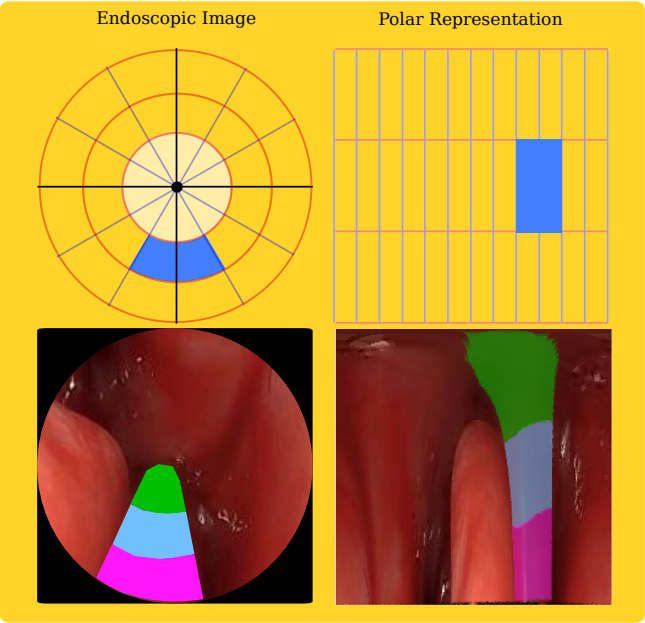


Fig. 2: A rectangular kernel implemented on a polar-transformed endoscopic image corresponds to a geometry suitable for representing the tool in the original image.

2) U-Net Image Semantic Segmentation: Semantic image segmentation is often approached with deep neural networks, including the use of feature-enhanced convolutional neural networks [21]. One such approach utilizes a deeply-supervised symmetric encoder-decoder architecture, called the U-Net [22]. In the medical imaging field, the U-Net structure has been used extensively, particularly for organ or tumor volumetric segmentation from tomography slices [23]–[25]. Because of prevalence and familiarity in the medical imaging community, the U-Net architecture is the chosen method for image segmentation in this work.

3) Polar Medical Imaging: Several sensors, including ultrasonic sensing and radar systems, gather information radially and thus in polar spatial coordinates. Some medical imaging systems also vary the angle between successive measurements, including computer aided tomography using an X-Ray scanner, and often implement a polar morphological transformation for visualization [26]. In particular, these data are visualized rectangularly with original sensor data interpolated or resampled using algorithms based on the Fourier slice theorem, e.g. the Jakowatz and O'Sullivan gridding methods [27].

Whereas these approaches are developed since data are spatially sampled in polar coordinates, the imaging data from endoscopes are presented in rectangular form - the useful image content is just restricted within a shape of a circle. In this work, a simple polar coordinate transform is performed where radial lines in Cartesian space are mapped to vertical lines in the polar cortical plane representation, and concentric circles in the Cartesian space are mapped to horizontal lines in the polar cortical plane representation. Lossless methods for this purpose exist [28] yet impart a greater computational burden.

a) Log-Polar Transform: Both polar and log-polar transformations can be used for affine image registration. Polar registration can accommodate for arbitrary size of rotations and at different scales. Scaling factors are represented as simple phase shifts in the log-polar transformed coordinates. The affine image registration problem was demonstrated to improve with a log-polar non-linear least squares hybrid approach [29]. The rotation and scaling effects in log-polar coordinates is well-studied, as well as implications with regard to optical flow and translational motions [30].

b) Adaptive Polar Transform: Matungka et al. observed that, while the log-polar transform is useful for robustness to rotation and scale variations, registration becomes an issue when occlusions or other alterations occur. This is due to the non-uniform spatial sampling of the polar transformation. Thus, a new adaptive sampling scheme was proposed that increases angular sampling frequency with increased radius. The resultant information in the cortical plane representation is non-rectangular, however [31].

B. Contributions

To the best of the authors' knowledge, this work is the first to simultaneously

- implement morphological polar transformation of endoscopic imaging, thus removing zero padding and rearranging data spatially;
- implement dual U-Net tool segmentation frameworks via two pathways
  - i) $\mathcal{R}$: using typical rectangular representation;
  - ii) $\mathcal{P}$: using polar coordinate counterparts;
- present a neural-network based hybrid approach that improves segmentation Dice score using a pathway selector $\mathcal{S}$ to determine the more probable prediction result of the two coordinate representation pathways for each input image. Two selector options - $\mathcal{S}_{\mathcal{M}}$ and $\mathcal{S}_{\mathcal{A}}$ - are presented in Section II-E.

The results of this work suggest that endoscopic images with certain features are better suited for tool segmentation using images arranged spatially using the described polar transformation as compared to unaltered rectangular coordinate image representations. The hybrid approach determines which images are better suited for either segmentation pathway.

II. Methods

Images were obtained from the University of Washington Sinus Surgery Cadaver/Live Dataset [20], [32]. The endoscopic videos were recorded using the Stryker 1088 HD camera and the Karl Storz Hopkins Ø 4mm 0° endoscope at 30fps. The data set includes manual annotations of tool pixels, with a variety of visual obstacles present, including: motion blur, blood, smoke, shadows and specular reflections. The overall workflow of the hybrid coordinate representation segmentation approach is depicted in Fig. 3 and Fig. 4.

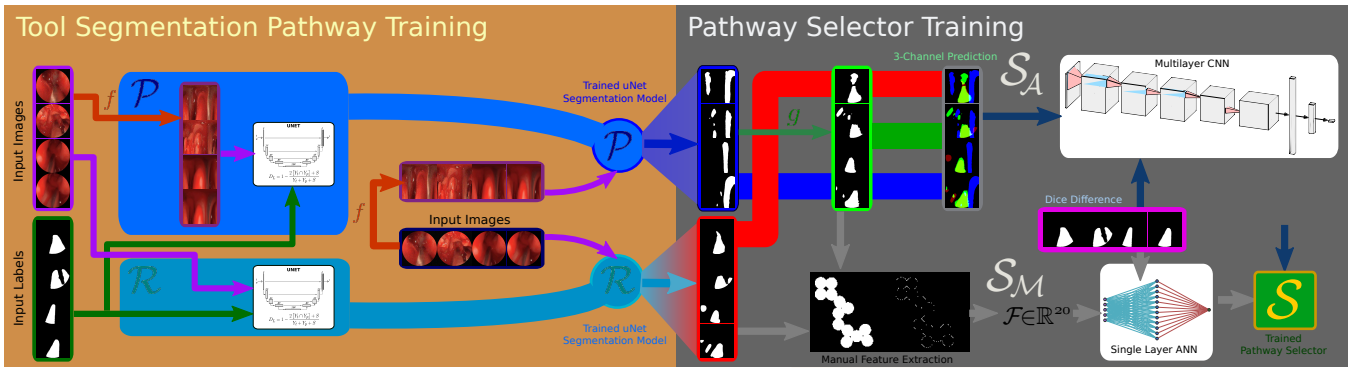


Fig. 3: Flowchart diagram depicting the hybrid coordinate representation network training method. The training image set is first used to train two separate U-Net segmentation networks, (one for each of the image coordinate representation pathways $\mathcal{P}$ and $\mathcal{R}$). The U-Net in $\mathcal{R}$ takes as input rectangular for mated endoscopic images i, while $\mathcal{P}$ takes polar images, i.e. $f(i)$. The training prediction masks and resultant Dice coefficients are used to train binary classifiers, $\mathcal{S}_\mathcal{M}$ and $\mathcal{S}_\mathcal{A}$, to learn which pathway yields better segmentation. $\mathcal{S}_\mathcal{M}$ uses manually selected binary image processing features, and uses only final rectangular format predictions. $\mathcal{S}_\mathcal{A}$, on the other hand, uses filter-based features and utilizes both types of masks generated in $\mathcal{P}$ as well as the mask type generated in $\mathcal{R}$. The pathway selector, $\mathcal{S}$, is thus trained and may consist of either $\mathcal{S}_\mathcal{M}$ or $\mathcal{S}_\mathcal{A}$.

A. *Image Pre-Processing*

All endoscopic images in the data set were resized to 256 $\times$ 256 pixels via gridded linear interpolation. The largest circle (with radius 128 pixels) from the center of the original image was extracted from the rectangular image to isolate useful endoscopic image information - note that less consistent endoscopic imaging may employ a circle detection method prior to resizing in case the circle image appears in a different location or is of a different size. The training set was further expanded by convolving an additive Gaussian white noise kernel across all training images and then normalizing pixel values to between 0 and 1.

B. *Polar Transformation, f*

Let $\mathbb{N}_p$ be the set of natural numbers $\{1, 2, ..., 256\}$. A pre-processed endoscopic image can be represented as a set of 3-tuples, call it $C = \{(x, y, v)\}$, where $(x, y) \in \mathbb{N}_p \times \mathbb{N}_p$, representing pixel coordinates, and v as the associated value of pixel (x, y). The goal is to represent radially with origin at the center of the image circle. The proposed method is computationally efficient, but does not sample space uniformly.

For an original endoscopic image C, an accompanying polar image is most easily expressed as a matrix whose entries correspond to spatial pixel location, call it $\mathcal{Z} \in \mathbb{N}_p^{256\times256}$. Let P represent a set of 3-tuples where an element $(\chi, \psi, s) \in P$ is generated by taking the entry in $\mathcal{Z}$ from the χ^{th} row and ψ^{th} column, and $\chi, \psi \in \{1, 2, ..., 256\}$ and assigning as s. The polar transformation is a surjective mapping $f : C \longrightarrow P$. To that end, select arbitrarily an element in P, suppose it is (χ, ψ, s). Then

$$
\begin{aligned}
(\chi, \psi, s) &= f\big((x, y, s)\big) \\
&\text{where} \\
x &= \left\lfloor \frac{\psi}{2} \cos\left(2\pi \frac{\chi}{256}\right) + 128 \right\rfloor \qquad (1) \\
y &= \left\lfloor \frac{\psi}{2} \sin\left(2\pi \frac{\chi}{256}\right) + 128 \right\rfloor \qquad (2)
\end{aligned}
$$

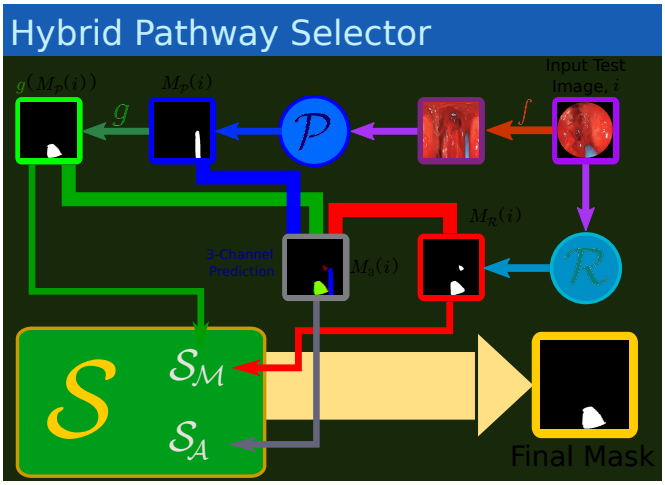


Fig. 4: Given an input image i, pathway $\mathcal{R}$ generates binary mask $M_\mathcal{R}(i)$. Pathway $\mathcal{P}$ takes $f(i)$ and generates both $M_\mathcal{P}(i)$ and $g\left(M_\mathcal{P}(i)\right)$. $\mathcal{S}_\mathcal{M}$ extracts 10 features from each of $M_\mathcal{R}(i)$ and $g\left(M_\mathcal{P}(i)\right)$ to predict the best mask, while $\mathcal{S}_\mathcal{A}$ uses a 3-channel image $M_3(i)$ composed of all three masks.

and $(x, y, s) \in C$. Then the corresponding element in $\mathcal{P}$ is $p_{\chi\psi} = s$. Visually $\mathcal{Z}$ is a square image, where pixel values in the pre-processed image circle are reparameterized by angle and radial distance from the image circle center.

C. *Back-Transformation, g*

After a polar represented endoscopic image is segmented, the image must be back-converted to the original image circle form in order to evaluate performance with rectangular coordinate endoscopic images. The back-transformation, g, is then composed of simply reverse operations of (1) and (2). Since sign ambiguity arises with inverse trigonometric operations, the regions of the polar image $\mathcal{Z}$ corresponding to different quadrants of the original image centered at the image circle center are set as constraints. Then for an arbitrary element (a, b, s) in the back-transformed image, the value is obtained from the polar image pixel (x, y) by

$$
\begin{aligned}
(a, b, s) &= g\big((x, y, s)\big) \\
\text{where} & \\
x &= \left\lfloor \sqrt{a^2 + b^2} \right\rceil \quad (3) \\
y &= \left\lfloor \tan^{-1}\left(\frac{y}{x}\right) \right\rceil \quad (4)
\end{aligned}
$$

D. Coordinate Representation Pathways

1) Rectangular Coordinate Representation Pathway, $\mathcal{R}$: No additional processing was performed on either training or testing images (beyond those already described in Section II-A) for the rectangular coordinate representation pathway - this represents the standard endoscopic image format used in most segmentation approaches. Given an input endoscopic image i, the $\mathcal{R}$ pathway simply passes i as an input to the trained U-Net tool segmentation network and generates a single predicted binary tool mask, $M_{\mathcal{R}}(i)$, as depicted in Fig.4.

$$i \longrightarrow \mathcal{R} \longrightarrow M_{\mathcal{R}}(i) \quad (5)$$

2) Polar Coordinate Representation Pathway, $\mathcal{P}$: In this pathway, endoscopic input images are first converted to a radial spatial coordinate representation of the endoscopic image data, as described in Section II-B. Training, segmentation and testing are all performed using image data in this coordinate representation before being back-transformed, as described in Section II-C, to evaluate segmentation performance. Given an input endoscopic image i, the $\mathcal{P}$ pathway first computes the polar coordinate representation of i, i.e. $f(i)$. $f(i)$ is an input to a trained U-Net polar tool segmentation network and generates a predicted polar binary tool mask, $M_{\mathcal{P}}(i)$. The final output of $\mathcal{P}$ for input image i consists of two masks: $M_{\mathcal{P}}(i)$ and the back-transformed polar tool mask, i.e. $g(M_{\mathcal{P}}(i))$. This is depicted in Fig.4.

$$f(i) \longrightarrow \mathcal{P} \longrightarrow \begin{cases} M_{\mathcal{P}}(i), \\ g\left(M_{\mathcal{P}}(i)\right) \end{cases} \quad (6)$$

3) U-Net Training: The segmentation model in both pathways was trained using the U-Net architecture with dice coefficient loss function, D_L. Suppose that Y_t is the ground truth segmentation and Y_p is the generated segementation prediction for a given input image. Then the dice loss is computed as the following

$$D_L = 1 - \frac{2\left[Y_t \cap Y_p\right] + S}{Y_t + Y_p + S} \quad (7)$$

where S was set to 1 in order to avoid dividing by 0.

a) Training Splits: The training-testing split was heuristically determined as 90-10 with a total of 7404 training images and 956 testing images. Images were selected randomly for training. A batch size of 2 for 50 epochs was used to train, with 100 batches used per epoch. The Adam optimizer with a learning rate of 1×10^{-4} was used for training, with the exponential decay rate of the first moment (β_1) set to 0.2 and the second moment (β_2) set to 0.8. A value of 1×10^{-7} was selected as ϵ.

b) Augmentations: Even though there were an abundance of training images (7404 images), over-fitting was observed in initial experiments with both coordinate representations. To improve the model, augmentations such as rotation, vertical and horizontal shifting, zooming, horizontal flip, shearing and rescaling were used for both pathways.

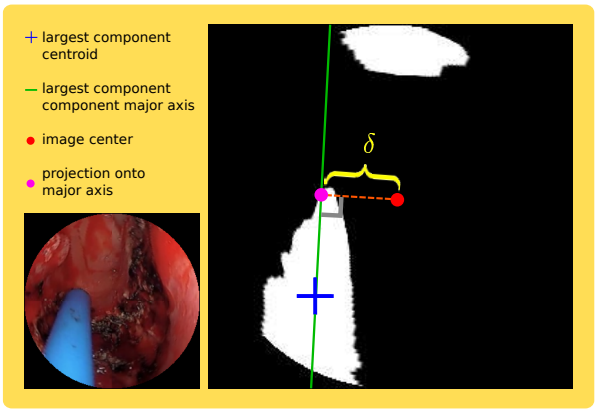


Fig. 5: Pictoral representation of feature variable δ. A line oriented as the major axis of the largest component and intersecting the component centroid is drawn. δ is then determined as the distance between the image center and its orthogonal projection on that line.

E. Intelligent Pathway Selector, $\mathcal{S}$

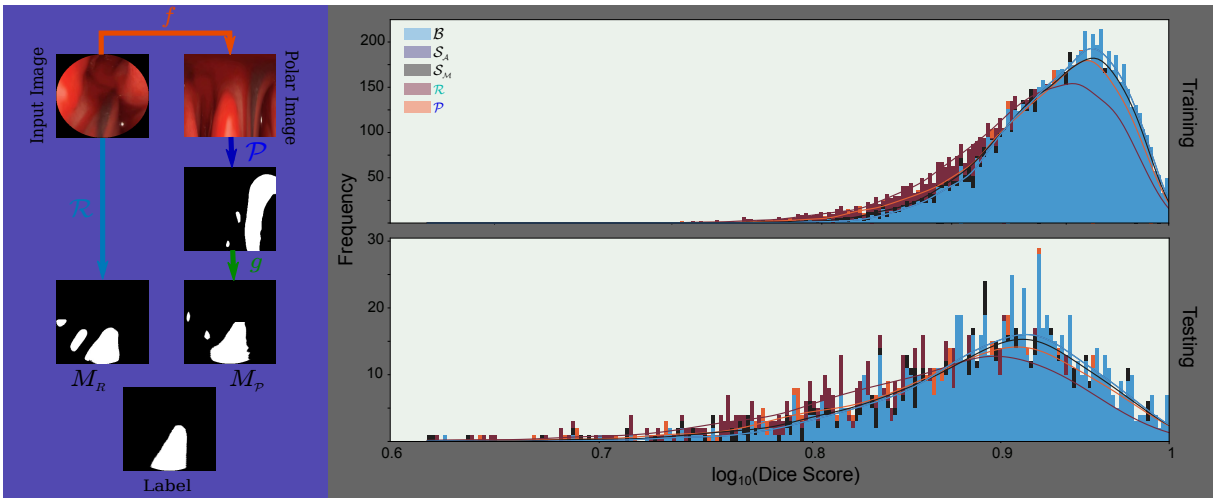
The U-Net segmentation networks trained for the rectangular and polar coordinate representation pathways yielded varying results on the training input images. The Dice coefficient was calculated for segmentation mask predictions generated from each of the two pathways for each training image, thus creating data output labels. Based on the Dice coefficient metric, 4395 input training endoscopic images were better segmented using the polar coordinate representation pathway network as compared to 3009 segmented better with the rectangular counterpart. By decoupling the training processes of the two segmentation models, each network utilized strong feature traits within each spatial representation across the entire dataset. These results were used to train binary classifiers for pathway selection of test images. Two binary selector classifiers were examined:

i) $\mathcal{S}_{\mathcal{M}}$ - with *manually* picked features that heuristically present strong binary distinguishing power.
ii) $\mathcal{S}_{\mathcal{A}}$ - with *automatically* generated filter-based features.

1) Manual Feature Classifier, $\mathcal{S}_{\mathcal{M}}$: As an input, $\mathcal{S}_{\mathcal{M}}$ takes for a single image i one predicted tool segmentation binary mask from each of the two pathways, namely $M_{\mathcal{R}}(i)$ and $g(M_{\mathcal{P}}(i))$. The selector is tasked to determine which of the two masks will yield better tool segmentation as measured by the Dice coefficient.

a) Feature Selection: A total of ten single variable features were generated for an input binary tool segmentation mask. These features were generated using basic binary image processing algorithms and connected component analysis (in the 4-connected sense). For a mask M, feature extraction is performed via F, such that

$$F(M) = \begin{bmatrix} \alpha & \gamma & \delta & \zeta & o & \lambda & \mu & \xi & \rho & \psi \end{bmatrix}^{\mathrm{T}} \quad (8)$$



(a) (b)

Fig. 6: (a) Sample segmentation results from representation pathways $\mathcal{R}$ and $\mathcal{P}$ (b) Training and Testing Dice Score Histogram

where α is the number of pixels in the largest component, γ the total number of connected components, δ the tool direction distance from the image center, ζ the total number of tool labeled pixels, o the circularity of the largest connected component, λ the pixel length of the major axis of the largest component, μ the Euler number of the largest component, ξ the eccentricity of the ellipse that has the same second-moments as the largest component, ρ the pixel perimeter of the largest component, and ψ the solidity of the largest component. Each feature was Z-normalized to that parameter's distribution within the training data. The parameter δ calculation is depicted in Fig. 5.

For each input to $\mathcal{S}_\mathcal{M}$, i.e. $(M_\mathcal{R}(i), g(M_\mathcal{P}(i)))$, a feature vector $\vec{\mathcal{F}}\,(M_\mathcal{R}(i), g(M_\mathcal{P}(i))) \in \mathbb{R}^{20}$ was generated as

$$\vec{\mathcal{F}}(M_\mathcal{R}(i), g(M_\mathcal{P}(i))) = \begin{pmatrix} F(M_\mathcal{R}(i)) \\ F(g(M_\mathcal{P}(i))) \end{pmatrix} \tag{9}$$

Feature vectors from binary mask prediction pairs using pathways $\mathcal{R}, \mathcal{P}$ on all 7404 training images were used as training inputs for a single layer artificial neural network. A single variable label output, the difference between the Dice coefficients using $M_\mathcal{R}$ versus $M_\mathcal{P}$ denoted $\mathcal{DD}_{gt}$, was used for the supervised training. The hidden layer consisted of 100 nodes and was trained using the quasi-Newton Broyden-Fletcher-Goldfarb-Shanno numerical optimization algorithm.

F. Automatic Filter Feature Classifier, $\mathcal{S}_\mathcal{A}$

For the $\mathcal{S}_\mathcal{A}$ pathway selector, the inputs are RGB images of size 256 x 256. Each layer corresponds to each of the types of masks generated: the rectangular coordinate binary mask generated by the $\mathcal{R}$ pathway $M_\mathcal{R}(i)$, and the polar and backconverted masks from the $\mathcal{P}$ pathway, $M_\mathcal{P}(i)$ and $g(M_\mathcal{R}(i))$. This 3-channel prediction is denoted $M_3(i)$ as depicted in Fig.4. The target outputs for $\mathcal{S}_\mathcal{A}$ are again $\mathcal{DD}_{gt}$. M_3 masks were generated for all 7404 training images and used to train a convolutional network consisting of three Conv2D layers with filter count and kernel sizes being $(32, 3 \text{ x } 3), (32, 3 \text{ x } 3), (64, 3 \text{ x } 3)$ followed by 2 x 2 max-pooling. Two Dense layers with 64 and 1 node(s) are added. The network uses ReLU and sigmoid activation functions for hidden and output layer(s) respectively.

G. Testing

A total of 956 testing endoscopic images were used to evaluate the hybrid coordinate representation approach. The testing images were segmented with four approaches: (1) U-Net trained on unaltered rectangular images (2) U-Net trained using polar representations, and hybrid approaches (3) $\mathcal{S}_\mathcal{M}$ and (4) $\mathcal{S}_\mathcal{A}$. Mean Dice and Intersection-over-Union (IoU) [33] scores were calculated for each method.

III. Results

Figure 6 shows (a) segementation results and (b) the Dice score distribution of the training and testing images using the four strategies $\mathcal{P}$, $\mathcal{R}$, $\mathcal{S}_\mathcal{M}$, $\mathcal{S}_\mathcal{A}$ and optimal selection $\mathcal{B}$, the best possible Dice and IoU scores from perfect selection of predicted masks for test images calculated post hoc. Table I shows the pathway selection classification performance results on the 956 testing images using $\mathcal{S}_\mathcal{M}$ and $\mathcal{S}_\mathcal{A}$.

Strategy	Accuracy	Precision		Recall	
		$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$
$\mathcal{S}_\mathcal{M}$	73.5%	61.2%	83.7%	75.3%	72.5%
$\mathcal{S}_\mathcal{A}$	99.7%	99.1%	100%	100%	99.5%

TABLE I: Testing Accuracy, Precision and Recall

Figure 7 depicts feature importance and confusion statistics through misclassification histograms of the 20 manually selected features using pathway selection approach $\mathcal{S}_\mathcal{M}$. Table II compares test image segmentation results for the four segmentation approaches. Measured as percent of potential improvement attained over the baseline pathway $\mathcal{R}$, the percentage (%) of maximum metric score achieved is calculated as the following:

$$\%\ \text{Max Score}(\mathcal{T}) = \frac{\text{Score}(\mathcal{T}) - \text{Score}(\mathcal{R})}{\text{Score}(\mathcal{B}) - \text{Score}(\mathcal{R})} \tag{10}$$

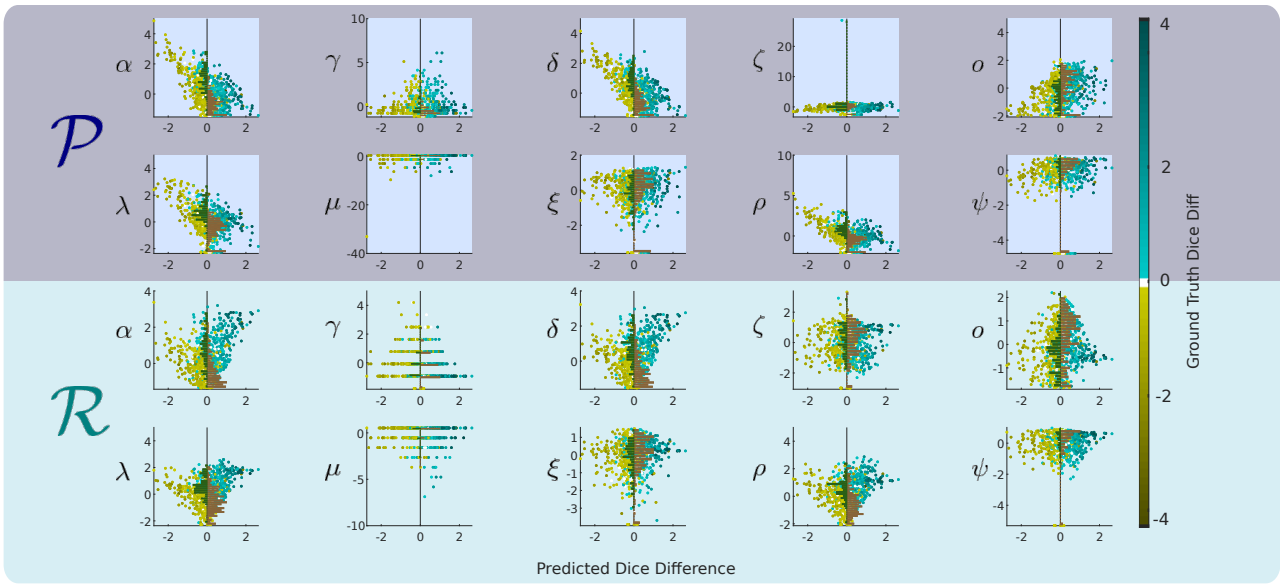


Fig. 7: Feature importance and confusion diagrams of the 956 test images using the manual feature pathway selector $\mathcal{S}_\mathcal{M}$. The vertical axes of the 20 subgraphs show the 20 features in $\vec{\mathcal{F}}$. The horizontal axis represents the predicted dice difference, $\mathcal{DD}_{pr}$. The color bar represents the ground truth dice difference labels, $\mathcal{DD}_{gt}$. A directional histogram along the center of each subgraph visualizes the number of incorrectly labeled data samples as a function of the feature values. Specifically, the false positives (yellow bars) are data samples with $\mathcal{DD}_{gt}<0$, $\mathcal{DD}_{pr}>0$, and the false negatives (green bars) are data samples with $\mathcal{DD}_{gt}>0$, $\mathcal{DD}_{pr}<0$.

, where Score(.) represents either the Dice or IoU metric score. $\mathcal{T}$ is the target strategy including the polar representation $\mathcal{P}$ or the two hybrid coordinate representation approaches $\mathcal{S}_\mathcal{M}$ and $\mathcal{S}_\mathcal{A}$.

Strategy	Dice				IoU			
	Mean		Median		Mean		Median	
	Raw	% Imp	Raw	% Imp	Raw	% Imp	Raw	% Imp
$\mathcal{R}$	0.867	0.00%	0.877	0.00%	0.771	0.00%	0.781	0.00%
$\mathcal{P}$	0.884	58.08%	0.893	59.91%	0.797	63.08%	0.807	64.15%
$\mathcal{S}_\mathcal{M}$	0.891	82.47%	0.899	83.85%	**0.808**	**84.51%**	**0.816**	**86.10%**
$\mathcal{S}_\mathcal{A}$	**0.896**	**99.97%**	**0.903**	**100.0%**	0.798	62.41%	0.809	69.27%
$\mathcal{B}$	0.896	100.0%	0.903	100.0%	0.815	100.0%	0.822	100.0%

TABLE II: Segmentation results (Dice and IoU) from 956 test images. The best performances are shown in **blue**. $\mathcal{B}$ shows the best segmentation performance by optimally selecting segmentation mask calculated manually post hoc.

Table II shows that $\mathcal{S}_\mathcal{M}$ achieves the best IoU score and $\mathcal{S}_\mathcal{A}$ demonstrates an impressive Dice score almost identical to the optimal selection result $\mathcal{B}$. Considering potential improvement over baseline pathway $\mathcal{R}$, $\mathcal{S}_\mathcal{A}$ achieves almost the max score in terms of Dice score while $\mathcal{S}_\mathcal{M}$ achieves about 85% of the maximum potential improvement in the metric of IoU.

Observing the results in Fig.6, $\mathcal{S}_\mathcal{A}$ is hardly visible as it almost fully overlaps with distribution $\mathcal{B}$ since the selector performs almost as well. Out of 956 test images, only three were misclassified by $\mathcal{S}_\mathcal{A}$. $\mathcal{S}_\mathcal{M}$ achieves a training distribution similar to $\mathcal{P}$ around its peak range, whereas its testing performance is more similar to $\mathcal{S}_\mathcal{A}$ and $\mathcal{B}$. This suggests that the explainable manual feature selection approach $\mathcal{S}_\mathcal{M}$ exhibits generalizability to unseen data samples.

Most subplots from Fig.7 appear symmetric, and thus making predictions by any single feature value is challenging. Using an ANN helped to extract more distinguishable features within the manually selected 20. Features α, δ, λ, and ρ from pathway $\mathcal{P}$ exhibit strong linear negative correlation with $\mathcal{DD}_{gt}$. Features α and δ from pathway $\mathcal{R}$ exhibit strong linear positive correlation with $\mathcal{DD}_{gt}$. These observations suggest that α, the number of connected components, and δ, the tool direction distance from the image center, of predicted tool segmentation masks have strong distinguishing power between $\mathcal{P}$ and $\mathcal{R}$ segmentation performance.

IV. Conclusion

This work demonstrated that a polar morphological transform of endoscopic images may sometimes result in better tool segmentation. Two different types of selectors were trained, one with manually determined features, $\mathcal{S}_\mathcal{M}$, and the other trained filter-based features, $\mathcal{S}_\mathcal{A}$. The proposed hybrid approaches intelligently selected coordinate representation pathway for each input image (either $\mathcal{P}$ or $\mathcal{R}$), and demonstrated results show improvements over U-Nets trained with either coordinate representation alone (polar or rectangular). While hybrid pathway selector $\mathcal{S}_\mathcal{A}$ exhibited almost perfect classification for optimizing Dice score, the resultant IoU score improvements were modest only. $\mathcal{S}_\mathcal{M}$, on the other hand, provided balanced improvements in both metrics over pathway $\mathcal{R}$ alone, and the manual feature designations provide more direct interpretable/explainable inferences. In particular, the analysis of each of the 20 features in Fig. 7 indicate that predicted masks with fewer connected components (α) and largest connected component directed towards the image center (δ) are amenable to better tool segmentation with pathway $\mathcal{P}$.

REFERENCES

[1] M. A. Minhem, B. Y. Safadi, H. Tamim, A. Mailhac, and R. S. Alami, "Does intraoperative endoscopy decrease complications after bariatric surgery? analysis of american college of surgeons national surgical quality improvement program database," *Surgical endoscopy*, vol. 33, no. 11, pp. 3629–3634, 2019.

[2] S. Kawakatsu, M. Ohashi, N. Hiki, S. Nunobe, M. Nagino, and T. Sano, "Use of endoscopy to determine the resection margin during laparoscopic gastrectomy for cancer," *Journal of British Surgery*, vol. 104, no. 13, pp. 1829–1836, 2017.

[3] A. Zia and I. Essa, "Automated surgical skill assessment in rmis training," *International journal of computer assisted radiology and surgery*, vol. 13, no. 5, pp. 731–739, 2018.

[4] O. Özgüner, R. Hao, R. C. Jackson, T. Shkurti, W. Newman, and M. C. Cavusoglu, "Three-dimensional surgical needle localization and tracking using stereo endoscopic image streams," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 6617–6624.

[5] A. Attanasio, C. Alberti, B. Scaglioni, N. Marahrens, A. F. Frangi, M. Leonetti, C. S. Biyani, E. De Momi, and P. Valdastri, "A comparative study of spatio-temporal u-nets for tissue segmentation in surgical robotics," *IEEE Transactions on Medical Robotics and Bionics*, 2021.

[6] Y.-H. Su, K. Huang, and B. Hannaford, "Multicamera 3d reconstruction of dynamic surgical cavities: Autonomous optimal camera viewpoint adjustment," in *2020 International Symposium on Medical Robotics (ISMR)*. IEEE, 2020, pp. 103–110.

[7] ——, "Multicamera 3d reconstruction of dynamic surgical cavities: non-rigid registration and point classification," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 7911–7918.

[8] ——, "Multicamera 3d reconstruction of dynamic surgical cavities: Camera grouping and pair sequencing," in *2019 International Symposium on Medical Robotics (ISMR)*. IEEE, 2019, pp. 1–7.

[9] ——, "Multicamera 3d viewpoint adjustment for robotic surgery via deep reinforcement learning," *Journal of Medical Robotics Research*, p. 2140003, 2021.

[10] M. Allan, S. Ourselin, D. J. Hawkes, J. D. Kelly, and D. Stoyanov, "3-d pose estimation of articulated instruments in robotic minimally invasive surgery," *IEEE transactions on medical imaging*, vol. 37, no. 5, pp. 1204–1213, 2018.

[11] 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.

[12] F. Qin, Y. Li, Y.-H. Su, D. Xu, and B. Hannaford, "Surgical instrument segmentation for endoscopic vision with data fusion of cnn prediction and kinematic pose," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9821–9827.

[13] 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.

[14] 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.

[15] T. Kurmann, P. M. Neila, X. Du, P. Fua, D. Stoyanov, S. Wolf, and R. Sznitman, "Simultaneous recognition and pose estimation of instruments in minimally invasive surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 505–513.

[16] 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.

[17] I. Laina, N. Rieke, C. Rupprecht, J. P. Vizcaíno, A. Eslami, F. Tombari, and N. Navab, "Concurrent segmentation and localization for tracking of surgical instruments," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 664–672.

[18] Y.-H. Su, W. Jiang, D. Chitrakar, K. Huang, H. Peng, and B. Hannaford, "Local style preservation in improved gan-driven synthetic image generation for endoscopic tool segmentation," *Sensors*, vol. 21, no. 15, p. 5163, 2021.

[19] B. Münzer, K. Schoeffmann, and L. Böszörmenyi, "Detection of circular content area in endoscopic videos," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. IEEE, 2013, pp. 534–536.

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

[21] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, 2020.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[23] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[24] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2018, pp. 3–11.

[25] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[26] V. Naranjo, R. Lloréns, M. Alcañiz, and F. López-Mir, "Metal artifact reduction in dental ct images using polar mathematical morphology," *Computer methods and programs in biomedicine*, vol. 102, no. 1, pp. 64–74, 2011.

[27] L. A. Gorham, B. D. Rigling, and E. G. Zelnio, "A comparison between imaging radar and medical imaging polar format algorithm implementations," in *Algorithms for Synthetic Aperture Radar Imagery XIV*, vol. 6568. International Society for Optics and Photonics, 2007, p. 65680K.

[28] W. Park and G. S. Chirikjian, "Interconversion between truncated cartesian and polar expansions of images," *IEEE transactions on image processing*, vol. 16, no. 8, pp. 1946–1955, 2007.

[29] G. Wolberg and S. Zokai, "Robust image registration using log-polar transform," in *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, vol. 1. IEEE, 2000, pp. 493–496.

[30] H. Araujo and J. M. Dias, "An introduction to the log-polar mapping [image sampling]," in *Proceedings II Workshop on Cybernetic Vision*. IEEE, 1996, pp. 139–144.

[31] R. Matungka, Y. F. Zheng, and R. L. Ewing, "Image registration using adaptive polar transform," *IEEE Transactions on Image Processing*, vol. 18, no. 10, pp. 2340–2354, 2009.

[32] 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.

[33] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 658–666.