2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

Surgical Instrument Segmentation for Endoscopic Vision with Data Fusion of CNN Prediction and Kinematic Pose

Fangbo Qin, Yangming Li, Yun-Hsuan Su, De Xu, *Senior Member, IEEE*, Blake Hannaford*, *Fellow, IEEE*

***Abstract*—The real-time and robust surgical instrument segmentation is an important issue for endoscopic vision. We propose an instrument segmentation method fusing the convolutional neural networks (CNN) prediction and the kinematic pose information. First, the CNN model ToolNet-C is designed, which cascades a convolutional feature extractor trained over numerous unlabeled images and a pixel-wise segmentor trained on few labeled images. Second, the silhouette projection of the instrument body onto the endoscopic image is implemented based on the measured kinematic pose. Third, the particle filter with the shape matching likelihood and the weight suppression is proposed for data fusion, whose estimate refines the kinematic pose. The refined pose determines an accurate silhouette mask, which is the final segmentation output. The experiments are conducted with a surgical navigation system, several animal-tissue backgrounds, and a debrider instrument.**

I. Introduction

Endoscopic vision plays an important role in surgical robots and computer-assisted surgical systems. Visual perception tasks, such as attribute labeling, pose estimation, image-based navigation, and three-dimensional (3D) reconstruction [1-4], are useful for the robotic surgery guidance and the real-time assistance to surgeons. Instrument segmentation is to separate the instrument foreground apart from the organ background, offering the location, orientation and presence status of instrument. The segmentation mask can also act as preliminary information in other perception tasks.

The image based instrument segmentation directly categorizes each pixel into background or foreground based on image features like color, shape and texture. The segmentation is challenging in the endoscopic images having weak contour, changing illumination, instrument-organ contact, mirror reflection and few textures. Another strategy of instrument segmentation is exploiting the kinematic information. Given the instrument pose and the camera model, the instrument's 3D shape can be projected onto the image as a 2D foreground mask. The kinematic pose of the instrument is available when measured by surgical robots or navigation systems. For surgical robots, such as the Raven II and da Vinci robots [5,6], the end effector pose is calculated by the robot kinematic model. For the navigation system, which is widely applied in the stereotactic surgery, the poses of the instrument and the endoscope are tracked in Cartesian space [7,8]. Thus, with the real time kinematic pose, the foreground mask can be simultaneously updated in the endoscopic image.

A. Related Works

1) Image Segmentation: The traditional image segmentation is based on a hand-crafted feature extraction module, which outputs the features like HSV color, gradient orientation, Gabor filter response, *etc*. [9-10]. The feature vector of each pixel is input into a trainable classifier, often based on the machine learning algorithms, such as random forest [9] and gradient boosting decision tree (GBDT) [10]. However, the hand-crafted features have limited richness and hierarchy.

Convolutional neural networks (CNNs) have achieved the advantage of automatic feature learning, the convenience of end-to-end training manner, and the high computation capability of graphics processing unit (GPU). Laina *et al*. proposed the CNN based concurrent segmentation and localization method [11]. Fully convolutional networks (FCN) were combined with the fast optical flow tracking to realize real time surgical instrument segmentation [12]. In [13], the convolutional auto-encoder was embedded with the recurrent neural networks layers to model the dependencies between pixels. The two lightweight models, ToolNet-MS and ToolNet-H, were the first two CNN architectures that could be used for real time instrument segmentation [14]. The ToolNet-H presented better accuracy, enabled by its holistically-nested structure and multi-scale loss. Although CNNs have large model capabilities, their supervised training relies on a large dataset that is laborious to label. If the number of training samples is too small relative to the model complexity, the model will be overfitted and have poor generalization performance. Moreover, CNN cannot guarantee the correct predictions for input patterns unseen during training.

2) CNN Based Feature Learning: An alternative approach to utilize the advantage of CNN but avoid the dependency on a large labeled dataset is the CNN based feature learning. In [15], the CNN was pre-trained offline on a large dataset to recognize object. The learned rich feature hierarchies were transferred to an online tracking task in which there was only

F. Qin and D. Xu are with Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 101408, China. Y. Li is with Department of Electrical Computer and Telecommunications Engineering Technology, Rochester Institute of Technology, Rochester, NY 14623, USA. Y. H. Su, and B. Hannaford are with Department of Electrical Engineering, University of Washington, Seattle, WA 98195-2500, USA. {blake@uw.edu}

978-1-5386-6027-0/19/$31.00 ©2019 IEEE

one labeled example in the first frame. The EndoNet used CNN as the feature extractor, whose output was passed to the machine learning models to estimate the surgical phase [16]. The unsupervised feature learning can leverage the unlabeled but numerous images to obtain the generalized feature representation. Radford *et al.* proposed the deep convolutional generative adversarial networks, whose discriminator could be reused as feature extractors for supervised tasks [17]. A fully convolutional auto-encoder was proposed for unsupervised feature learning, which was successfully trained in the end-to-end manner [18].

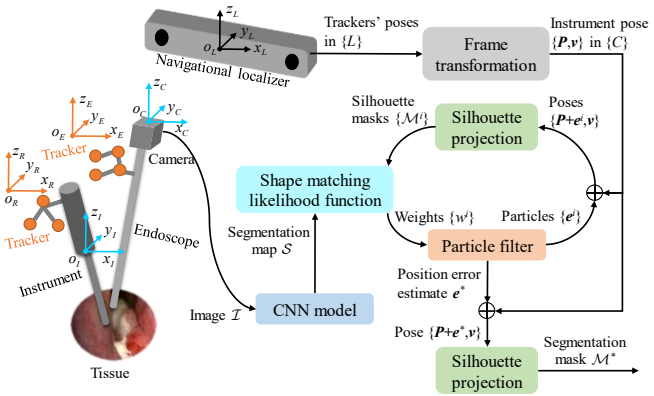
3) Data Fusion with Kinematic Information: The kinematic information is robust but suffers from errors in practical application. In [19], the extended Kalman filter was used to fuse the robot kinematics with the endoscopic stereo vision to track the manipulator joints, which helps to reject outliers and fill in gaps of detection failure. The brute-force joint search matching was used to correct the raw kinematics to match the virtual rendering template with the endoscopic image [20]. Su *et al*. leveraged the kinematic information to provide a shape prior mask, which was fused with the image color filter based on the template matching in frequency domain, to compensate for the shape prior's offsets in scale, translation and rotation [21]. As is reported in [19-21], the kinematic pose errors were caused by the model error, elastic deformation and time misalignment.

B. Motivation

This work aims to realize real time and robust surgical instrument segmentation in endoscopic image, for surgical systems that have both endoscopic camera and kinematic sensing. 1) It is meaningful to enable the CNN learning from a few labeled images, which avoids laborious labeling work and long training time. The lightweight CNN model ToolNet-C is proposed, which cascades a convolutional feature extractor trained over a large unlabeled dataset, and a pixel-wise segmentor trained over a tiny labeled dataset (e.g. 30 training samples). 2) Second, the instrument silhouette projection is implemented, so that a kinematic pose determines a segmentation mask in the endoscopic image. 3) It is advantageous to fuse the two kinds of information to obtain the robust and accurate segmentation. A particle filter based data fusion method is proposed, in which the shape matching likelihood is designed to weight the particles, and the weight suppression is used to enable the efficient particle resampling.

II. System Overview

The surgical system contains an endoscopic imaging system and a navigation system. The navigation system has an optical localizer, which can track the 6-degree-of-freedom poses of the optical trackers, by observing the reflective sphere markers on the tracker. Assuming the tracker is fixed on a rigid instrument and the relative pose is calibrated, the kinematic pose of instrument can be obtained. The coordinate frames are built as shown in Fig. 1. For the instrument segmentation, the CNN model predicts a map $\mathcal{S}$ on which each pixel has a confidence of belonging to the foreground. The silhouette projection determines a segmentation mask $\mathcal{M}$ based on the instrument position $\boldsymbol{P}$ and orientation $\boldsymbol{v}$. The



{L} Localizer frame; {E} Endoscope tracker frame; {R} Instrument tracker frame; {C} Camera frame; {I} Instrument frame

Fig. 1. System overview.

particle filter is utilized to fuse the two kinds of information and estimate the latent position error as $\boldsymbol{e}^*$. Then the instrument position is refined as $\boldsymbol{P}+\boldsymbol{e}^*$, determining an accurate segmentation mask $\mathcal{M}^*$. The orientation error has an ignorable influence when the instrument tip length in view is small.

III. CNN Based Instrument Segmentation

A. ToolNet-C Segmentation Model

As is shown in Fig. 2(a), the ToolNet-C model is designed by cascading a feature extractor $\mathbb{F}$ and a pixel-wise segmentor $\mathbb{S}$. The basic idea is that if $\mathbb{F}$ learns a set of rich and reusable features from numerous unlabeled images, $\mathbb{S}$ can be a lightweight pixel-wise classifier and requires just a few labeled training images.

The input image $\mathcal{I}$ with the pixel value range [0,255] is linearly scaled and shifted into the value range [-1,1] and then fed to $\mathbb{F}$ to extract the rich hierarchical features. The kernels of the four convolutional layers in $\mathbb{F}$ are all with the 5×5 size and 2×2 stride. Batch normalization (BN) is applied to improve the stability of deep model training. The leaky rectified linear unit (LReLU) is used as the activation function. The feature maps $\mathcal{H}_i$ (i=1,2,3,4) are output by the first four convolutional layers respectively.

After fed into the segmentor $\mathbb{S}$, $\mathcal{H}_i$ is converted to $\mathcal{H}_{i\_1}$ in a cross-channel manner, by the convolutional layer with the 1×1 kernel-size and 1×1 stride, and the feature channel number is not changed. Secondly, $\mathcal{H}_{i\_1}$ is converted to $\mathcal{H}_{i\_2}$ by the convolutional layer with the 3×3 kernel-size and 1×1 stride, and the feature channel number is reduced to 32. Rectified linear unit (ReLU) is used as the activation function for these two convolutional layers. Thirdly, the four adapted feature maps $\mathcal{H}_{i\_2}$ (i=1,2,3,4) are resized to the same size and then aggregated as $\mathcal{H}_a$ by concatenating all the channels. Finally, the aggregated feature map $\mathcal{H}_a$ is processed by the convolutional layer with the 3×3 kernel-size and 1×1 stride, and then by the softmax function, which outputs the segmentation map $\mathcal{S}$. The size of $\mathcal{S}$ is half of the size of $\mathcal{I}$, so that the smaller map size is better for the real-time performance of the following data fusion procedure.

Note that the resizing method in $\mathbb{S}$ is the nearest neighbor interpolation method, which reserves the originally extracted feature values. In contrast, if the linear or cubic interpolation is used, the interpolation between a positive and a negative

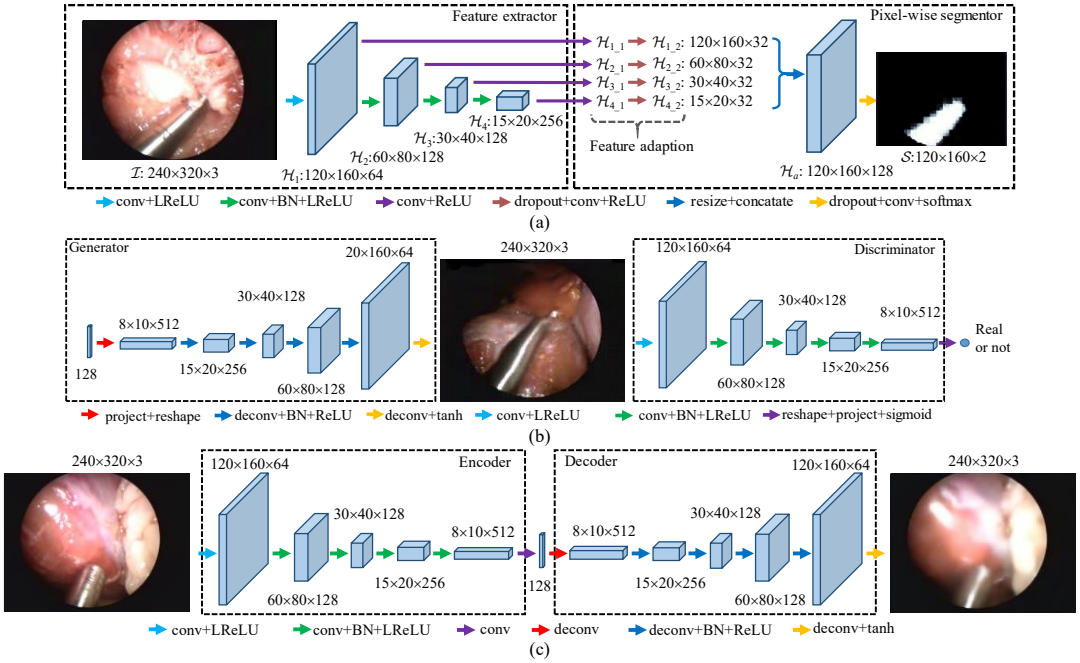


Fig. 2. CNN model architectures. (a) TooNet-C segmentation model, (b) DCGAN and (c) FCAE models are used for unsupervised feature learning.

sample might give ambiguous samples in training. The maximum receptive field size for $\mathcal{H}_{i\_2}$ is 93×93, therefore each pixel in $\mathcal{H}_a$ contains the features mapped from a 93×93 patch in the input image.

B. *Unsupervised Feature Learning*

Firstly, $\mathbb{F}$ is embedded into an unsupervised feature learning model, which is trained over unlabeled images. After the model is trained, the parameters of $\mathbb{F}$ are obtained. The two representative methods, deep convolutional generative adversarial nets (DCGAN) [17] and fully convolutional auto-encoder (FCAE) [18], are utilized for feature learning. Their structures are customized so that $\mathbb{F}$ can be embedded into the discriminator of DCGAN and the encoder of FCAE as the four convolutional layers, as is shown in Fig. 2(b,c).

DCGAN concurrently learns a generator $G(\cdot)$ that maps a random noise vector $\boldsymbol{z}$ to a fake image and a discriminator $D(\cdot)$ that distinguishes the generated fake image from the real image $\mathcal{I}$. The discriminator output indicates the probability of being a real image. D and G are both formed by convolutional layers. The learning objective is given by

$$\min_G \max_D \left(\mathbb{E}_{\mathcal{I}} \left[\log D(\mathcal{I}) \right] + \mathbb{E}_{\boldsymbol{z}} \left[\log \left(1 - D\left[G(\boldsymbol{z}) \right] \right) \right] \right) \quad (1)$$

The global optimum of this min-max game is achieved when the generated fake images have the same distribution with the real images. As is shown in Fig. 2(b), the generated image looks real, but the instrument in it has shape distortion.

FCAE concurrently learns a encoder $E(\cdot)$ and a decoder $E'(\cdot)$. The encoder maps an image into the code vector $\boldsymbol{z}$, which is a latent image expression in the low dimensional space. The decoder reconstructs an image $\boldsymbol{z}$. The learning objective is to minimize the squared $L2$ distance between the input images and reconstructed images, namely,

$$\min_{E,E'} \sum_{m,i,j,k} \left(\mathcal{I}_{m,i,j,k} - E'\left[E(\mathcal{I}) \right]_{m,i,j,k} \right)^2 \quad (2)$$

where m, i, j, and k are the indices of sample, row, column and channels, respectively. As is shown in Fig 2(c), the image reconstructed from the 128-dimensional code vector is similar to the input but blurrier.

According to the learning objectives, the discriminator of DCGAN learns the discriminative features which the real images own. The encoder of FCAE learns the representative features to model the image's latent structure.

C. *Supervised Segmentor Training*

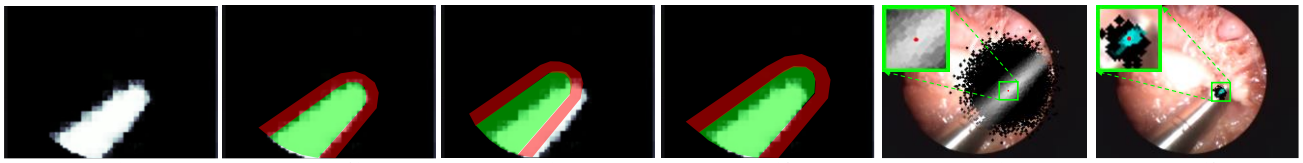
After the unsupervised feature learning, the parameters of $\mathbb{F}$ are obtained and fixed. With the N_t labeled images $\mathcal{I}_n$ (n=1,2,...,N_t), only the parameters of $\mathbb{S}$ is optimized by supervised training. The N_t ground-truth label maps $\mathcal{G}_n$ are binary maps whose pixel value belongs to {0,1}. The learning objective is to minimize the squared $L2$ distance between the output maps $\mathcal{S}$ and ground-truth label maps $\mathcal{G}$, as given by

$$\min_{\boldsymbol{\theta}} \sum_{m,i,j,k} \left[\mathcal{S}_{m,i,j,k} - \mathcal{G}_{m,i,j,k} \right]^2, \ s.t. \left\| \boldsymbol{w}_{\boldsymbol{\theta}} \right\|_2 \le c \quad (3)$$

where the indices m, i, j and k indicate the m^{th} sample, i^{th} row, j^{th} column, and k^{th} class. The 1st class is the background and 2nd class is the foreground. Inspired by [22], the max-norm constraint and dropout technique are applied to prevent the overfitting problem, which is essential to the learning with small dataset. $\boldsymbol{w}_{\boldsymbol{\theta}}$ is the last convolutional layer's weight matrix. The max-norm constraint is to limit the $L2$ norm of the weight matrix within the upper-bound c. Because the training samples are few, the training method is stochastic gradient descent instead of mini-batch gradient descent.

IV. Kinematic Pose Based Silhouette Projection

Suppose the 3D shape of the surgical instrument is known, a grid of 3D points is sampled on the instrument's outer surface. The 3D points are transformed into the endoscopic camera frame and projected onto the image as a grid of 2D



(a) (b) (c) (d) (e) (f)

Fig. 3. Illustration of shape matching likelihood and weight suppression. (a) Foreground confidence map predicted by CNN, (b) match with high likelihood, (c,d) mismatch with translation and scale offsets. (e) Global distribution of particle weight, (f) local distribtion of particle weight after weight suppression. The brighter dot indicates the particle with higher weight.

points. Then the outer envelope of the 2D point grid is found to form the counters of the silhouette mask. Finally, the pixels overlapping with the back border are eliminated from the silhouette mask. The frame transformation and calibration method are briefly introduced here.

A point ${}^{I}\boldsymbol{P}$ on the instrument surface is transferred into the camera frame $\{C\}$ by the transformation chain ${}^{C}\boldsymbol{P}={}^{C}\boldsymbol{T}_{E}\left({}^{L}\boldsymbol{T}_{E}\right)^{-1}{}^{L}\boldsymbol{T}_{R}{}^{R}\boldsymbol{T}_{I}{}^{I}\boldsymbol{P}$. The transformations ${}^{R}\boldsymbol{T}_{I}$, ${}^{L}\boldsymbol{T}_{R}$ and ${}^{L}\boldsymbol{T}_{E}$ are directly provided by the commercial surgical navigation system. The intrinsic camera parameters are calibrated with the classical chessboard based method. Using the sphere marker on the optical tracker as the object, the sphere center's 3D position is given by the localizer and the sphere center's 2D position is extracted from the image, which form a 3D-2D point correspondence. After $n\geq 4$ point correspondences are collected, the Perspective-n-Point (PnP) method [23] is utilized to solve the extrinsic parameters, which determines the transformation ${}^{C}\boldsymbol{T}_{E}$. Note that the silhouette projection can also be used with the surgical robot by simply modifying the frame transformation based on the robot kinematic model.

V. Particle Filter Based Data Fusion

The CNN prediction is directly based on the endoscopic image, but outliers and incompleteness might occur. The silhouette projection generates segmentation indirectly and gives the regular segmentation shape, but suffers from the kinematic pose error that is not ignorable compared to the instrument tip size. Thus, it is appealing to fuse the two kinds of information. Assuming the kinematic position error $\boldsymbol{e}$ is the latent state. The observations include the endoscopic image $\mathcal{I}$ and the raw kinematic pose $\{\boldsymbol{P},\boldsymbol{v}\}$. The particle filter is utilized to estimate the posterior distribution of the latent state with a set of random particles $\{\boldsymbol{e}^i\}$ and the associated weights $\{\boldsymbol{w}^i\}$ (i=1,2,…,N_p) [24], as given in Algorithm 1.

Algorithm 1: Particle Filter based Data Fusion

Parameters: Particle number N_p, covariance matrix $\boldsymbol{C}$, likelihood threshold τ, discount factor σ.

t=0, flag_init=0;
while task is not terminated:
  Read the raw image $\mathcal{I}_t$ and the kinematic pose $\{\boldsymbol{P}_t,\boldsymbol{v}_t\}$;
  if t=0 ***or*** *flag_init*=0: # *Initialize particles*
    Draw particles $\boldsymbol{e}_t^i \sim \mathcal{N}(\boldsymbol{0};\boldsymbol{C})$ (i=1,2,…, N_p);
    flag_init←1;
  else: # *Update particles*
    Draw particles $\boldsymbol{e}_t^i \sim \mathcal{N}(\boldsymbol{e}_{t-1}^i;\boldsymbol{C})$;
  Weight the particles by $w_t^i \leftarrow p(\boldsymbol{e}_t|\mathcal{I}_t,\boldsymbol{P}_t,\boldsymbol{v}_t)$ with (4);
  if $\max(w_t^i)<\tau$: # *Fail to get confident observation*
    flag_init←0 and output $\boldsymbol{e}_t^* \leftarrow \sigma\boldsymbol{e}_{t-1}^i$;
  else: # *Estimate the posterior expectation*
    Weight suppression by $w_t^i \leftarrow w_t^{*i}$ with (5);
    Weight normalization by $w_t^i \leftarrow w_t^i / \sum_i w_t^i$;
    Output the expectation by $\boldsymbol{e}_t^* \leftarrow \sum_i w_t^i \boldsymbol{e}_t^i$;
    Resampling [24]; t←t+1.

Based on the generic PF, firstly, the shape matching likelihood is designed to weight the particles. Assuming the kinematic pose is compensated for by the position error $\boldsymbol{e}$, the silhouette mask is projected as $\mathcal{M}(\boldsymbol{P}+\boldsymbol{e},\boldsymbol{v})\in\{0,1\}$, which is abbreviated as $\mathcal{M}_{P,e,v}$. At the same time, the endoscopic image $\mathcal{I}$ is fed to CNN and mapped as a segmentation map $S(\mathcal{I})\in[0,1]$, whose second channel $S(\mathcal{I})_2$ is the foreground confidence map. The function measuring the shape matching likelihood of the observations conditioned on the latent state $\boldsymbol{e}$ is given by,

$$p(\mathcal{I},\boldsymbol{P},\boldsymbol{v}|\boldsymbol{e})=\frac{\sum_{i,j}\left(S(\mathcal{I})_2\circ\mathcal{M}_{P,e,v}\right)_{i,j}}{\sum_{i,j}\left(\mathcal{M}_{P,e,v}\right)_{i,j}}-\frac{\sum_{i,j}\left[S(\mathcal{I})_2\circ\mathcal{M}_c\circ\left(\mathcal{M}_{P,e,v}\oplus\kappa-\mathcal{M}_{P,e,v}\right)\right]_{i,j}}{\sum_{i,j}\left[\mathcal{M}_c\circ\left(\mathcal{M}_{P,e,v}\oplus\kappa-\mathcal{M}_{P,e,v}\right)\right]_{i,j}} \tag{4}$$

where $\circ$ is the entry-wise product. $\oplus$ is the morphological dilation operator with the 25×25 structuring element κ. $\mathcal{M}_c$ is the circular mask used to eliminate the dilated pixels on the black border. As illustrated in Fig. 3 (a-d), the likelihood is the average confidence on $S(\mathcal{I})_2$ within the silhouette region (green) minus the average confidence within the silhouette region's outer neighborhood (red). If $\boldsymbol{P}+\boldsymbol{e}$ deviates from the actual position, the silhouette mask will have offsets in position and scale so that the likelihood is lower.

Second, the weight suppression is used before the standard weight normalization step. In real time application, a small number of particles are employed and are distributed around the optimal latent state. However, as illustrated in Fig. 3(e), the particle weight distribution around the maximum point is flat, which is not good for particle resampling. Because if the weights vary in a small range comparing to their mean, the resampling cannot efficiently polish the particles. Therefore, the weights are suppressed by the mean weight, namely,

$$w_t^{*i}=\max\left(0,w_t^i-\frac{1}{N_p}\sum_i w_t^i\right) \tag{5}$$

After the weight suppression, only the particles close to the optimal state still have positive weights, as shown in Fig. 3(f).

Third, if the maximum weight is below the threshold τ, the observation is considered unreliable. In this case, the position error state is updated by reducing its previous value with the discount factor σ, and the particles will be initialized again at

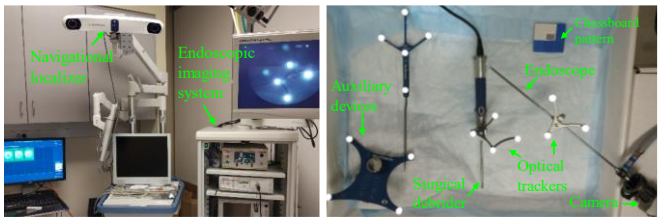


Fig. 4. Experimental platform and devices.

the next time step. In the data fusion process, the size of the confidence map and the silhouette mask is the half of the input image size, which reduces the computation burden. After the position error estimate is updated, the final silhouette mask $\mathcal{M}^*$ has the same size with the input image.

VI. Experiments and Results

A. Hardware Configuration

The Medtronic Stealth Station S7 surgical navigation system was used to provide the kinematic pose of the surgical instrument, as shown in Fig. 4. The endoscopic imaging is based on the Stryker 1088 HD camera system and the Karl Storz Hopkins ∅4mm 0° endoscope, which provided the 320×240 resolution images at 30fps. The surgical instrument was a ∅4mm Medtronic debrider, which is widely utilized in nasal surgeries. In this research, the data was collected on these two commercial systems and then processed offline. The computer for the offline experiment had an Intel i7-6700K 4.0GHz CPU and a Nvidia GTX1070 GPU. The computation of CNN was on the GPU, and the other computation was on CPU.

B. Data Collection and Evaluation Metrics

With the same surgical debrider and different tissue samples, the two endoscopic videos were recorded. To make the dataset challenging, in the videos exist the contact between instrument and tissue, the slow and fast motions of instrument, the significant brightness change, the mirror reflection and highlights, and the significant scale change of instrument in image. In addition, the instrument orientation is with limited change. The video 1 offered 18780 endoscopic image frames, which formed the unlabeled dataset $\mathbb{D}_U$ and were used for the unsupervised feature learning. 50 images in video 1 and their labels formed the segmentor training dataset $\mathbb{D}_S$, among which 30 images were used for training and the rest for validation. The video 2 was nine minutes long and had the synchronized kinematic pose trajectories of the instrument. For the evaluation, 500 images in the video 2 and the corresponding labels formed the test dataset $\mathbb{D}_T$.

Dice similarity coefficient (DSC) and *intersection of union* (IOU) were used as the metrics to evaluate the segmentation,

$$\mathrm{DSC}(X,Y)=\frac{1}{K}\sum_{k=1}^{K}\frac{2|X_k\cap Y_k|}{|X_k|+|Y_k|},\ \mathrm{IOU}(X,Y)=\frac{1}{K}\sum_{k=1}^{K}\frac{|X_k\cap Y_k|}{|X_k\cup Y_k|}$$

where X_k and Y_k represent the pixels belonging to the k^{th} class on the output map X and the ground truth map Y, respectively. K=2 is the class number. $|\cdot|$ is the number counting operation.

Considering that a false positive outlier occurs in the CNN output, it will decrease the DSC and IOU values. However, if the outlier is far from the actual instrument region, it will have no influence on the final result after the data fusion, because the shape matching likelihood just involves the pixels near the actual foreground region. Therefore, the metrics *with far outliers removed* are deigned as $\mathrm{DSC}^*=\mathrm{DSC}(X^*,Y)$ and $\mathrm{IOU}^*=\mathrm{IOU}(X^*,Y)$, where $X^*=X\cap(Y\oplus\kappa)$. κ is the 25×25 structuring element for the morphological dilation operation $\oplus$. DSC^* and IOU^* are appropriate for evaluating whether the prediction map is good for data fusion.

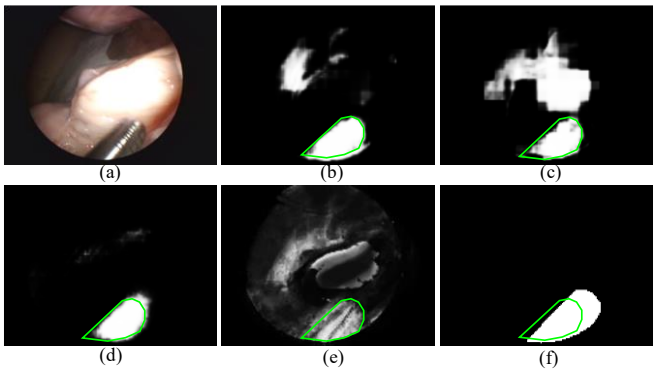


Fig. 5. Segmentation results without data fusion. (a) Raw image, (b) ToolNet-C with DCGAN, (c) ToolNet-C with FCAE, (d) ToolNet-H, (e) GBDT, (f) silhouette projection. The ground-truth is marked by green line.

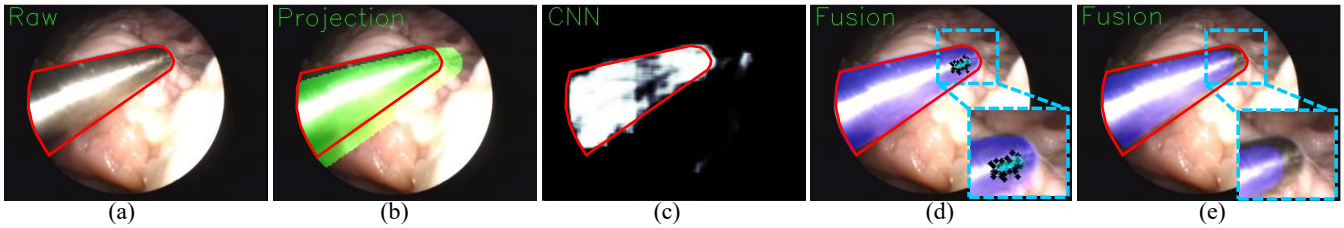
C. Training Configuration

The DCGAN and FCAE models were trained on $\mathbb{D}_U$ with data augmentation including the random flipping, scaling and rotation. The learning rate, training epoch and batch-size for mini-batch gradient descent were α=0.0005, 8 and 16, respectively. Afterwards, the pixel-wise segmentor was learned with $\mathbb{D}_S$, without any data augmentation The learning rate and training epoch were α=0.001 and 50, respectively, for stochastic gradient descent. The Adam optimizer was applied [25], whose exponential decay rates of the 1st and 2nd order moment estimates were β_1=0.5 and β_2=0.99, respectively. The keep probability of dropout was 0.8. The upper bound c of the weight norm was 4.

For the comparison, the ToolNet-H [14] and gradient boosting decision tree (GBDT) [10] were also trained with $\mathbb{D}_S$, also without any data augmentation. As to ToolNet-H, the learning rate and training epoch were α=0.001 and 50, respectively, for stochastic gradient descent. The same Adam optimizer was utilized. As to GBDT, the hand-crafted features included the RGB color, HSV color and 32 Gabor filter responses. The offline GBDT classifier had the exponential loss, the weak-classifier number of 32, the maximum depth of tree of 5, and the learning rate of 0.1.

D. Segmentation Results without Data Fusion

The above four image segmentation models along with the silhouette projection were tested with the dataset $\mathbb{D}_T$. ToolNet-C's output was resized to the input image size with the nearest neighbor interpolation. As is shown by the examples in Fig. 5, ToolNet-C model with DCGAN presented the sharp segmentation contour and some far outliers. ToolNet-C with FCAE was not as accurate as that with DCGAN and had more outliers. The ToolNet-H gave the weaker outlier response but the less accurate contour. The GBDT gave the worst results because its hand-crafted fea-



(a) (b) (c) (d) (e)

Fig. 6. Segmentation results with data fusion. (a) Raw image, (b) silhouette projection (green), (c) ToolNet-C with DCGAN, (d) ToolNet-C and particle filter (blue), (e) ToolNet-C and template matching (blue). The ground-truth is marked by red line.

TABLE I
SEGMENTATION PERFORMANCES WITHOUT DATA FUSION

Segmentation method	mDSC (%)	mIOU (%)	mDSC* (%)	mIOU* (%)	Time cost (ms)
ToolNet-C with DCGAN	82.4	76.2	**90.8**	**84.9**	7
ToolNet-C with FCAE	81.2	74.4	89.2	82.7	7
ToolNet-H [14]	**84.3**	**77.6**	87.5	80.8	6
GBDT [10]	65.3	57.7	72.8	65.0	29
Silhouette projection	85.6	77.9	89.7	83.3	<0.5

tures were not rich and hierarchical. The silhouette projection offered regular shape mask but suffered from offsets. The outputs of these methods were evaluated with the metrics in Section VI.*C*. The prediction maps were firstly binarized with the threshold 0.7. The mean DSC (mDSC), mean IOU (mIOU), mean DSC* (mDSC*), and mean IOU (mIOU*) over the 500 images in $\mathbb{D}_T$ are shown in Table I. With the mDSC and mIOU metrics, ToolNet-H provided the highest scores. With the DSC* and IOU* metrics, ToolNet-C with DCGAN provided the highest scores, which had the best potential for data fusion, because it showed the best accuracy after removing the far outliers.

E. Segmentation Results with Data Fusion

In this experiment, the above four image segmentation models were fused with the silhouette projection. The particle filter was configured with the particle number N_p=50, the discount factor σ=0.7, and the likelihood threshold τ=0.3. The covariance matrix $\boldsymbol{C}$ was a diagonal matrix with the diagonal elements of 0.1. For comparison, the template matching based data fusion was also investigated [21]. The template matching corrected the translation and scale offsets based on the cross correlation metric. Multi-thread programming techniques were utilized in the particle filter and the template matching to reduce the execution time.

The segmentation with data fusion was continuously run over the entire video 2 with the synchronized kinematic pose trajectory. During running, if the image frame was included in $\mathbb{D}_T$, its segmentation result was compared to the ground truth in $\mathbb{D}_T$ with the metrics DSC and IOU. Finally, the mDSC and mIOU were calculated, as shown in Table II. By comparing Table I and II, it is found that the segmentation accuracy was improved after the data fusion. The particle filter behaved better than the template matching. The best accuracy was provided by the ToolNet-C with DCGAN and the particle filter. Compared to the best performances without data fusion, the mDSC and mIOU with data fusion were 12.2% and 14.7% higher, respectively. The average time cost was 33ms which satisfied the real-time application requirement. An example is shown in Fig. 6. Although the prediction of ToolNet-C had outliers and defections, the final segmentation mask was outlier-free and shape-complete. As to the template matching, the existence of the inconsistency between the prior shape and the actual instrument shape was a major disadvantage. More examples were shown in the online video at https://youtu.be/2iDUs7ppz9Y.

TABLE II
SEGMENTATION PERFORMANCES WITH DATA FUSION

Data Fusion Method	Segmentation method	mDSC (%)	mIOU (%)	Time cost (ms)
Particle filter	ToolNet-C with DCGAN	**94.6**	**90.9**	31
	ToolNet-C with FCAE	93.7	89.4	31
	ToolNet-H [14]	91.9	86.8	29
	GBDT [10]	87.4	80.6	53
Template matching [21]	ToolNet-C with DCGAN	90.3	84.8	10
	ToolNet-C with FCAE	89.4	84.0	10
	ToolNet-H [14]	90.0	84.6	9
	GBDT [10]	84.9	79.8	32

VII. CONCLUSION

This work addresses the surgical instrument segmentation issue for endoscopic vision, considering the data fusion of CNN prediction and kinematic pose. The proposed CNN model ToolNet-C learns features from numerous unlabeled images and learns segmentation from few labeled images, making its application more convenient. The particle filter based data fusion leverages the CNN prediction, instrument 3D shape and kinematic pose, so that the segmentation result is more robust and has regular shape. The weight suppression and shape matching likelihood are proposed to effect the resampling and weighting of particles in the particle filter. The proposed segmentation pipeline is suitable for endoscopic vision of local surgeries with challenging imaging condition but limited scene variation.

VIII. ACKNOWLEDGMENTS

We are pleased to acknowledge support from National Science Foundation (grant number: IIS-1637444).

REFERENCES

[1] S. Kumar, M. S. Narayanan, P. Singhal, *et al.*, "Surgical tool attributes from monocular video," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2014, pp.4887-4892.

[2] M. S. Nosrati, R. Abugharbieh, J. M. Peyrat, *et al.*, "Simultaneous multi-structure segmentation and 3D non-rigid pose estimation in image-guided robotic surgery," *IEEE Trans. Med. Imag.*, no. 35, vol. 1, pp. 1-12, 2016.

[3] S. Leonard, A. Sinha, A. Reiter, *et al.*, "Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in-vivo clinical data," *IEEE Trans. Med. Imag.*, online published, 2018.

[4] Q. Zhao, T. Price, S. Pizer, *et al.*, "The Endoscopogram: A 3D model reconstructed from endoscopic video frames," In *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.* (MICCAI), 2016, pp. 439-447.

[5] B. Hannaford, J. Rosen, D. W. Friedman, *et al.*, "Raven-II: an open platform for surgical robotics research," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 4, pp. 954–959, 2013.

[6] R. H. Taylor and D. Stoianovici, "Medical robotics in computer-integrated surgery. *IEEE Trans. Robot. Autom.*, vol. 19, no, 5, pp.765-781, 2003.

[7] G. Dagnino, I. Georgilas, S. Morad, et al., "RAFS: A computer-assisted robotic system for minimally invasive joint fracture surgery, based on pre- and intra-operative imaging," *IEEE Int. Conf. Robot. Autom.*, 2017, pp.1754-1759.

[8] Z. Fan, G. Chen, J. Wang, et al., "Spatial position measurement system for surgical navigation using 3-d image marker-based tracking tools with compact volume," *IEEE Trans. Biomed. Eng.,* vol. 65, no. 2, pp. 378-389, 2018.

[9] Y. Chen, M. M. Marinho, Y. Kurose, et al., "Towards robust needle segmentation and tracking in pediatric endoscopic surgery," in *Proc. SPIE Med. Imag.*, 2018, pp. 105762Y.

[10] J. Son, I. Jung, K. Park, *et al.*, "Tracking-by-segmentation with online gradient boosting decision tree," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3056-3064.

[11] I. Laina, N. Rieke, C. Rupprecht, *et al.*, "Concurrent segmentation and localization for tracking of surgical instruments," In *Proc. MICCAI*, 2017, pp. 664-672.

[12] L. C. Garcia-Peraza-Herrera, W. Li, C. Gruijthuijsen, *et al*., "Real-time segmentation of non-rigid surgical tools based on deep learning and tracking," Int. Workshop Comput. Assist. Robot. Endosc., 2016, pp. 84-95.

[13] M. Attia, M. Hossny, S. Nahavandi, *et al*., "Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 2017, pp.3373-3378.

[14] L. C. Garcia-Peraza-Herrera, W. Li, L. Fidon, *et al*., "ToolNet: holistically-nested real-time segmentation of robotic surgical tools," in *Proc. IEEE Int. Conf. Intell. Robot. Syst.*, 2017, pp. 5717-5722.

[15] N. Wang, S. Li, A. Gupta, *et al*., "Transferring rich feature hierarchies for robust visual tracking," *arXiv preprint arXiv:1501.04587*, 2015

[16] A. P. Twinanda, S. Shehata, D. Mutter, *et al*., "EndoNet: A deep architecture for recognition tasks on laparoscopic videos," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 86-97, 2016.

[17] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[18] F. Li, H. Qiao, and B. Zhang, "Discriminatively boosted image clustering with fully convolutional auto-encoders," *Pattern Recognition*, vol. 83, pp.161-173, 2018.

[19] A. Reiter, P. K. Allen, and T. Zhao, "Feature classification for tracking articulated surgical tools," Medical Image Computing and Computer-Assisted Intervention, 2012, pp. 592-600.

[20] A. Reiter, P. K. Allen, and T. Zhao, "Marker-less articulated surgical tool detection," in *Proc*. *Comput. Assist. Radiol. Surg.*, vol. 7, pp. 175-176 2012.

[21] Y. H. Su, K. Huang, and B. Hannaford, "Real-time vision-based surgical tool segmentation with robot kinematics prior," *IEEE Int. Symp. Med. Robot., 2018*.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, *et al*., "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp.1929-1958, 2014.

[23] V. Lepetit, M. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, 2009.

[24] M. Arulampalam, S. Maskell, N. Gordon, *et al*., "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Sign. Process.*, vol. 50, no. 2, pp. 174–188, 2002.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014