Co-generation and Segmentation for Generalized Surgical Instrument Segmentation on Unlabelled Data

Megha Kalia[1]([✉]) [iD], Tajwar Abrar Aleef[2] [iD], Nassir Navab[3], Peter Black[4], and Septimiu E. Salcudean[1]

[1] Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada
`{mkalia,tims}@ece.ubc.ca`
[2] School of Biomedical Engineering, University of British Columbia, Vancouver, Canada
`tajwaraleef@ece.ubc.ca`
[3] Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany
`nassir.navab@tum.de`
[4] Vancouver Prostate Centre, Vancouver, Canada
`peter.black@ubc.ca`

Abstract. Surgical instrument segmentation for robot-assisted surgery is needed for accurate instrument tracking and augmented reality overlays. Therefore, the topic has been the subject of a number of recent papers in the CAI community. Deep learning-based methods have shown state-of-the-art performance for surgical instrument segmentation, but their results depend on labelled data. However, labelled surgical data is of limited availability and is a bottleneck in surgical translation of these methods. In this paper, we demonstrate the limited generalizability of these methods on different datasets, including robot-assisted surgeries on human subjects. We then propose a novel joint generation and segmentation strategy to learn a segmentation model with better generalization capability to domains that have no labelled data. The method leverages the availability of labelled data in a different domain. The generator does the domain translation from the labelled domain to the unlabelled domain and simultaneously, the segmentation model learns using the generated data while regularizing the generative model. We compared our method with state-of-the-art methods and showed its generalizability on publicly available datasets and on our own recorded video frames from robot-assisted prostatectomies. Our method shows consistently high mean Dice scores on both labelled and unlabelled domains when data is available only for one of the domains.

Keywords: Surgical instrument segmentation · Unpaired image to image translation · Generative adversarial learning

---

M. Kalia and T. A. Aleef—Contributed equally to the manuscript

1   Introduction

Surgical instrument segmentation is fundamental to Augmented Reality (AR) in image-guided robot-assisted surgery (RAS) [11] and has been an active topic of research, with convolutional neural network (CNN)-based methods surpassing prior methods by a significant margin [2,5,10]. CNN-based methods depend on the availability of annotated surgical data, which may be difficult to obtain [9]. Their performance has been reported for publicly available *ex-vivo* and porcine *in-vivo* RAS surgeries, but not in human RAS.

Recently, many generative approaches have been proposed to mitigate the problem of limited clinical labelled data [4,7,13]. For laparoscopic instrument segmentation, [13] proposed a generative adversarial network (GAN)-based method to use a small amount of labelled data. In [7], labelled data from cadaver surgery was transferred to *in-vivo* surgery. Then a separate segmentation model was trained using either the translated cadaver data or translated *in-vivo* data to the cadaver domain. In [4], an image-to-image (I2I) mapping of simulated to real surgical instruments was proposed, with blending into the camera background. In the above methods, the translated data was used to train a segmentation model. Finding validated quantitative metrics for the quality of translated data is difficult and is the topic of on-going research [6,15]. In many cases, the generative models change the surgical instruments' shape and introduce artefacts while the overall accuracy decreases [Fig. 1]; this is undesirable for clinical application. Hence, a segmentation strategy leveraging the power of generative models to alleviate the problem of unlabelled clinical data while addressing the predominant current challenges of generative models is imperative.

Therefore, in the current paper, we present a joint unpaired I2I mapping and segmentation strategy for better generalizability of a surgical instrument segmentation model to a domain with no labelled data. The generative and segmentation models are trained together and reach convergence in a synergistic manner. The generative model maps from a source domain with labelled data to a target domain with unlabelled data with constant feedback from the segmentation model. The segmentation model trains in parallel on the generated target images and on the labelled source images. The convergence criterion of this joint-system is the segmentation quality. The segmentation model also regularizes the generative model that can otherwise change the shape of the surgical instruments during the I2I mapping. We call our method coSegGAN. The closest method to it is presented in [7]. However, unlike in [7], our segmentation model is not pre-trained. It provides feedback to the generators as it learns using the generated data, thus seeing much more varied data. Unlike prior work, we provide an explicit shape constraint on latent space to provide intermediate supervision during generative training. Through evaluation on real surgical sequences and publicly available datasets, we show that coSegGAN has better generalizability than existing methods. The main contribution of the paper is presenting a joint generation and segmentation framework that provides state of the art (SOTA) results to segment surgical instruments on unlabelled data. The method performs better than using a generative model for data augmentation as a separate

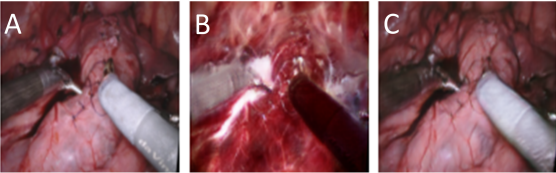| Method	Test / Train	Endovis	UCL	Surgery
Ternaus	Endovis	94.9	58.2	86.0
Ternaus	UCL	63.6	93.2	43.5
Rasnet	Endovis	89.2	84.2	80.5
Rasnet	UCL	56.1	92.7	36.8
U-Net	Endovis	95.0	33.5	58.8
U-Net	UCL	72.2	87.7	45.1

Fig. 1. (Left) Table showing the limited generalizability of state of the art (SOTA) methods across domains. Mean Dice scores are shown for different methods on datasets *Endovis*, *UCL*, and *Surgery*. (Right) Figures showing the problem with cycleGAN where the intermediate generated output can be unrealistic while overall cycle consistency loss is low. Figures A, B, and C show the original, translated, and reconstructed domain.

step. To the best of our knowledge, this is the first method that segments surgical instruments with no labelled data by jointly training the generative and segmentation model as a joint feedback system to perform an I2I mapping between the labelled and the unlabelled domain.

2 Methods

2.1 Network Details

The generative part of coSegGAN uses a cycleGAN like architecture with two generators and two discriminators [16]. Let x_{ai} and x_{bi} denote the two i^{th} images in the domains ψ_A and ψ_B, respectively, and x_a and x_b denote the set of all images in domains a and b, respectively. y_{ai} denotes the corresponding individual label for the ith x_{ai} image and y_a is the set of all such labels. G_A and G_B are the two generators estimating the mappings, with $G_A : x_{b \to a}$ and $G_B : x_{a \to b}$, respectively. The discriminator D_A is responsible to discriminate between given true images in domain ψ_A and generated images $G_A(x_b)$. Similarly, D_B is responsible for discriminating between the true domain ψ_B and generated images $G_B(x_a)$. Both G_A and G_B have a U-Net-like architecture [12] with a contracting and expanding path. The contracting path consists of four 4×4 convolutional layer with stride 2 + Leaky ReLu + Instance normalization [14] blocks where in each subsequent block the output is halved and the channel numbers are doubled. The expanding path consists of three blocks with each block having an up-sampling layer + 4×4 convolution with a stride of 1 + ReLu activation + Instance normalization. The output of each block was concatenated with the low-level features from the contracting path by skip connections and then passed as an input to the next block. The output of the final block was passed though a convolutional layer followed by a tanh activation. For the discriminator, we used a patchGAN similar to [16]. For the segmentation model (S) in coSegGAN we used the original U-Net architecture but with 16 base filters to prevent over-fitting and to reduce computation. This did not decrease the performance of segmentation when compared to the original U-Net, as determined empirically.
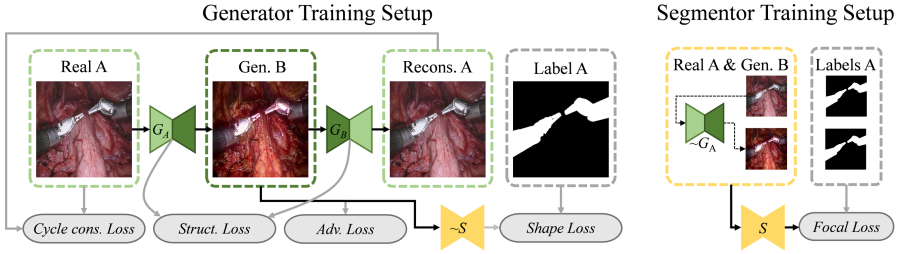
Fig. 2. Overview of the training setup for the generation and segmentation side. The diagram on the left shows the A to B mapping side of the cycleGANs that is modified to incorporate shape loss and structural loss during training. The grey blocks indicate the different losses used for updating the weights of the generator. On the right side, the input and loss used for the segmentation model training are shown. Any network indicated by '~' in this figure indicates it has frozen weights.

2.2   Training Strategy

We trained the generators, discriminators & the segmentation model in an alternative fashion. In the first run, the weights through the generators, G_A and G_B were back-propagated while freezing the weights of the discriminators and the segmentation model. Then in the next run, the discriminators as well as the segmentation model, D_A, D_B, & S were trained and updated. For training S, both x_a and $G_B(x_a)$ were fed as the input. Since the generated images are translated versions of the real image, the corresponding labels for $G_B(x_a)$ are the same as x_a. Note that S is seeing different variations of the generated target domain images in every epoch because the generators and S are learning in parallel. While the quality of the I2I mapping from the generators increases, the quality of the images seen by S also increases. Details can be seen in Fig. 2.

2.3   Loss Functions

Segmentation Model. In order not to overwhelm the loss with the higher number of background pixels, we used an α-balanced variant of focal loss, $\mathcal{L}_{foc}$ [8], a modification of cross-entropy, where the γ factor controls the contribution of high-probability samples in the loss calculation. We used the hyper-parameters γ and α as 2.0 and 0.25, respectively. The total segmentation loss, $\mathcal{L}_{seg}$, is

$$\mathcal{L}_{seg} = \mathcal{L}_{foc}(x_a, y_a) + \mathcal{L}_{foc}(G_B(x_a), y_a) \tag{1}$$

Generative Model. For cycleGAN we used an adversarial loss, $\mathcal{L}_{GAN}$, and a pixel-level cycle consistency loss $\mathcal{L}_{cyc}$ proposed in [16]. Although $\mathcal{L}_{cyc}$ reduces the number of possibilities when mapping across domains and regularizes the cycleGAN, it does not suffice to preserve the higher-level semantics in the image. This can change the shapes of surgical instruments during the translation, which is not desirable. Therefore we included feedback from the segmentation model

in the total generative loss. This penalizes the generation of unrealistic surgical instrument shapes in $G_B(x_a)$. Since we are interested in the mapping from x_a to x_b, which later is fed as an input to the segmentation model, we included this constraint only on the generator G_B. This shape preservation loss, $\mathcal{L}_{shape}$, is

$$\mathcal{L}_{shape} = \mathcal{L}_{foc}(G_B(x_a,)y_a) \tag{2}$$

In cycleGAN models, $\mathcal{L}_{cycTotal}$ is the sum of two cycle consistency losses such that, $\mathcal{L}_{cycTotal} = \mathcal{L}_{cyc}(x_a, G_A(G_B(x_a))) + \mathcal{L}_{cyc}(x_b, G_B(G_A(x_b)))$. These losses enforce pixel level constraints between the original inputs x_a and x_b and reconstructed outputs $G_A(G_B(x_a))$ and $G_A(G_B(x_a))$, where the two GANs are optimized together. There is no intermediate supervision after each generative step $G_A : x_{b \to a}$ and $G_B : x_{a \to b}$. Thus G_A and G_B can produce unrealistic images while the total $\mathcal{L}_{cyc}$ is reduced (Shown in Fig. 1, (right)). In particular, the mapping across domains should change only the 'appearance' of the scene while retaining the domain-invariant structural elements. To preserve the structural properties of the scene across domains, we introduce an explicit, intermediate, feature level, latent space loss. This latent space loss, and the total generated loss, are:

$$\mathcal{L}_{structure} = \mathbb{E}\left[\|e_A(x_a) - e_B(G_B(x_a))\|_1\right] + \mathbb{E}\left[\|e_B(x_b) - e_A(G_A(x_b))\|_1\right] \tag{3}$$

$$\mathcal{L}_{generator} = \lambda_1 \mathcal{L}_{GANTotal} + \lambda_2 \mathcal{L}_{cycTotal} + \lambda_3 \mathcal{L}_{shape} + \lambda_4 \mathcal{L}_{structure} + \lambda_5 \mathcal{L}_I \tag{4}$$

where, e_A and e_B are encoders in G_B and G_A, respectively, $\mathcal{L}_{GANTotal} = \mathcal{L}_{GAN}(G_B, D_B, x_a, x_b) + \mathcal{L}_{GAN}(G_A, D_A, x_b, x_a)$ and $\mathcal{L}_I$ is the identity mapping loss as given in [16]. Values of λ_1, λ_2, λ_3, λ_4, and λ_5 are 1, 10, 1, 5, and 1, respectively. These values were tuned during the hyper-parameter tuning phase.

Training Details and Hyper-Parameters. For training and testing our models, we use Tensorflow & Keras API on a NVIDIA Tesla V100 GPU (16 GB). For training the proposed models, we used a batch size of 8, and Adam optimizer with β_1 and β_2 of 0.9 and 0.999, respectively, with a learning rate of 10^{-3}. We trained our models for 100 epochs (approximately 12 h) and saved weights of the segmentation model with the highest validation Dice score [17]. Code is available at: https://github.com/tajwarabraraleef/coSegGAN.

3   Experiments

Datasets: Endovis Challenge, 2017, *in-vivo* **Dataset** [1]**:** It is a porcine surgery procedure with a training set consisting of 8 videos of 225 frames each and a test set consisting of 8 videos of 75 frames and 2 videos of 300 frames each. We used 6 videos for training and 2 videos for validation from the training set. We used 8 videos from the test set for testing; these were not used for validation. In the paper, we refer to this dataset as *Endovis*. In Table 1 *Endovis* is abbreviated as *Endo*.

UCL *ex-vivo* Dataset [3]**:** The dataset consists of 14 videos with different animal tissues as background. Similar to [3], we used 8, 2 and 4 videos for training, validation and testing, respectively.

Prostatectomy Dataset. We prepared the training dataset from 5 videos of robot-assisted radical prostatectomy procedures with the da Vinci Si surgical system from Vancouver General Hospital, Vancouver, Canada. We manually selected 1327 frames to isolate surgical instruments from other visible objects in the surgical field of view. These frames do not have corresponding labels. To evaluate the performance of the various methods on actual surgical data, we prepared a test set of 182 frames taken from 4 different surgeries independent from the training set. The test data represents approximately 12% of the entire surgical data used. We manually labelled surgical instruments in these frames only for the purpose of testing coSegGAN and existing methods. All the frames were center cropped to give a final size of 721×503 pixels. We will refer to this dataset as *Surgery* in the rest of the paper. Ethics to collect data was obtained from the Institutional Clinical Research Ethics Board. For all three datasets, we resized the frames to 256×256 to accelerate the computation.

Evaluation. We compared coSegGAN with Ternausnet, the best performing method in the Endovis Challenge [1] for binary segmentation and RASnet, reporting a mean 94.65% Dice coefficient on Endovis. For a fair comparison to coSegGAN, we performed data augmentation with the cycleGAN architecture given in Sect. 2. The cycleGAN model was run for 50 epochs in all cases as it converged in 50 epochs. After cycleGAN I2I translation from source (with labels) to target domain, the SOTA segmentation models were trained with both the translated and original domain data. We also performed an ablation experiment comparing coSegGAN with and without the proposed $\mathbf{L}_{structure}$ loss. We refer to RASnet, Ternausnet, and our U-Net variant with focal loss, trained using the augmented data generated from a separate cycleGAN (unlike our joint strategy) as $RASnet+$, $Ternausnet+$ and $U\text{-}Net_{FL}+$ respectively. The coSeg-GAN network without $\mathbf{L}_{structure}$ is called $coSegGAN-$. We performed evaluation of four combinations of datasets for labelled and unlabelled domains. For ease of reporting, we refer to *Endovis* (labelled) + *Surgery* (Unlabelled), *UCL* (labelled) + *Surgery* (Unlabelled), *Endovis* (labelled) + *UCL* (Unlabelled), and *UCL* (labelled) + *Endovis* (Unlabelled) data combinations as case 1, case 2, case 3, and case 4, respectively. Since, we want to quantify the generalizability of our method across labelled and unlabelled domains, for a particular dataset combination, we also calculated an absolute difference in the Dice scores, $\Delta\,Dice$, and absolute difference in Intersection over Union (IoU), $\Delta\,IoU$, between labelled domain A and unlabelled domain B. The lower the $\Delta\,Dice$ and $\Delta\,IoU$, the higher is the generalizability between domains (refer to Table 1).

Table 1. Comparison of Mean Dice and IoU scores of coSegGAN with SOTA methods

Method		Dom A	DomB (Unlabelled)	Domain A		Domain B (Unlabelled)		Δ Values	
				Dice	IoU	Dice	IoU	Dice	IoU
coSegGAN	Case 1	Endo	Surgery	**93.7%**	88.4%	**92.8%**	84.7%	**0.9%**	3.7%
	Case 2	UCL	Surgery	91.1%	84.2%	74.3%	59.8%	16.8%	24.4%
	Case 3	Endo	UCL	**93.2%**	88.3%	**90.0%**	82.2%	**3.2%**	6.1%
	Case 4	UCL	Endo	93.5%	91.1%	79.4%	66.8%	14.1%	24.3%
RASnet+	Case 1	Endo	Surgery	88.3%	79.9%	78.1%	64.7%	10.2%	15.2%
	Case 2	UCL	Surgery	92.3%	85.8%	47.8%	33.0%	44.5%	52.8%
	Case 3	Endo	UCL	88.4%	80.0%	83.3%	71.9%	5.1%	8.1%
	Case 4	UCL	Endo	92.4%	85.9%	66.8%	52.9%	25.6%	33.0%
Ternaus+	Case 1	Endo	Surgery	94.2%	89.9%	88.7%	80.4%	5.5%	9.5%
	Case 2	UCL	Surgery	95.8%	92.1%	46.0%	31.3%	49.8%	60.8%
	Case 3	Endo	UCL	93.3%	89.2%	41.7%	29.0%	51.6%	60.2%
	Case 4	UCL	Endo	93.4%	87.8%	55.0%	41.2%	38.4%	46.6%
U-Net_{FL+}	Case 1	Endo	Surgery	91.8%	85.7%	58.0%	42.5%	33.8%	43.5%
	Case 2	UCL	Surgery	93.2%	87.4%	36.0%	22.9%	57.2%	64.5%
	Case 3	Endo	UCL	83.9%	73.5%	23.3%	13.6%	60.7%	31.5%
	Case 4	UCL	Endo	74.6%	61.2%	56.5%	42.0%	18.1%	19.2%
coSegGAN–	case 1	Endo	Surgery	94.1%	89.6%	92.3%	86.0%	**1.8%**	3.6%
	Case 2	UCL	Surgery	93.5%	93.0%	74.5%	69.7%	19.0%	23.3%
	Case 3	Endo	UCL	93.3%	88.2%	90.8%	83.0%	**2.5%**	5.2%
	Case 4	UCL	Endo	94.2%	89.2%	74.4%	64.8%	19.8%	24.4%

4   Results and Discussion

For case 1, the proposed coSegGAN network gave significantly higher Dice (92.8%) and IoU scores (84.7%) on unlabelled domain B (*Surgery*) when compared to *RASnet+*, *Ternausnet+* and U-$Net_{FL}+$ which have Dice scores of 78.1% (IoU = 64.7%), 88.7% (IoU = 80.4%), and 84.1% (IoU = 42.5%), respectively. For case 2 as well, the Dice score for coSegGAN on unlabelled domain (*Surgery*) is 74.3% (IoU = 59.8%) while *RASnet+*, *Ternausnet+*, and U-$Net_{FL}+$ have lower Dice scores of 47.8% (IoU =33.0%), 46.0% (IoU = 31.3%), and 45.6% (IoU = 22.9%), respectively. Similarly, for case 3, the Dice (IoU = 82.2%) score for coSegGAN on unlabelled data (*UCL*) is 90%, which is higher than *RASnet+*, *Ternausnet+* and U-$Net_{FL}+$ with Dice scores of 83.3% (IoU = 71.9%), 41.7% (IoU = 29.0%), and 81.8% (IoU = 13.6%), respectively. For case 4, the Dice score for coSegGAN on unlabelled *Endovis* data is 79.4% (IoU = 66.8%), which, similar to other cases, is higher than the rest of the methods; Dice scores of *RASnet+*, *Ternausnet+* and U-$Net_{FL}+$ being 66.8% (IoU = 52.9%), 55.0% (IoU = 52.9%) and 56.5% (IoU = 42.0%), respectively.

The Δ $Dice$, for coSegGAN for case 1 is much lower 0.9% (IoU = 3.7%) while for $RASnet+$, $Ternausnet+$, and $U\text{-}Net_{FL}+$ it is 10.2% (IoU = 15.2%), 5.5% (IoU = 9.5%), and 33.8% (IoU = 43.5%), respectively. For case 2, Δ $Dice$ for coSegGAN is 16.8% (IoU = 24.4%), while for $RASnet+$, $Ternausnet+$ and $U\text{-}Net_{FL}+$ it is 44.5% (IoU = 52.8%), 49.8% (IoU = 60.8%) and 57.2% (IoU = 64.5%), respectively. For case 3, Δ $Dice$, for coSegGAN is 3.2% (IoU = 6.1%), which is much lower than $RASnet+$, $Ternausnet+$ and $U\text{-}Net_{FL}+$ with Δ $Dice$ of 5.1% (IoU = 8.1%), 51.6% (IoU = 60.2%), and 60.7% (IoU = 31.5%), respectively. For case 4, similarly, the Δ $Dice$ for coSegGAN is 14.1% (IoU = 24.3%) when compared to $RASnet+$, $Ternausnet+$, and $U\text{-}Net_{FL}+$ with Δ $Dice$ of 25.6% (IoU = 33.0%), 38.4% (IoU = 46.6%), and 18.1% (IoU = 19.2%), respectively. Consistently higher Dice and IoU on unlabelled data and significantly lower Δ $Dice$ and Δ IoU of coSegGAN show its generalizability when compared to all other methods for all the cases.

For coSegGAN, in cases 2 and 4, when the mapping is from UCL (labelled) to either $Surgery$ or $Endovis$, the Δ $Dice$ is higher than cases 1 and 3, showing comparatively less generalizability. This could be because the UCL data is an $ex\text{-}vivo$ dataset where data distribution potentially differs from a real surgery, with remarkably different lighting and background. Also, there is only one type of surgical instrument visible in the UCL dataset, which might have hindered the mapping to multiple types of instruments.

In the ablation experiment, coSegGAN–, i.e., coSegGAN without the $\mathbf{L}_{structure}$, showed comparable performance with coSegGAN, except case 4, where the performance of coSegGAN is significantly higher (approximately 5%) on the unlabelled $Endovis$ dataset. coSegGAN– has higher Δ $Dice$ for all cases except case 4, showing that with the $\mathbf{L}_{structure}$ loss coSegGAN generalizes better to both labelled and unlabelled datasets.

A qualitative comparison of coSegGAN with other methods for different surgeries can be seen in Fig. 3. As can be seen (column 1), coSegGAN performs better in preserving overall tool structure, with finer details, when compared to other methods. In comparison to $Ternausnet+$ and $RASnet+$, the method also produces fewer false positives [Fig. 3 (column 2)]. Although coSegGAN performs better than SOTA methods in identifying tools, it occasionally fails to identify the tool in the presence of blood where surgical instrument blends in with the background. Usually this happens at the image periphery where the region is relatively dark compared to the well-lit image center. Figure 3 (column 4) shows one such failure case.
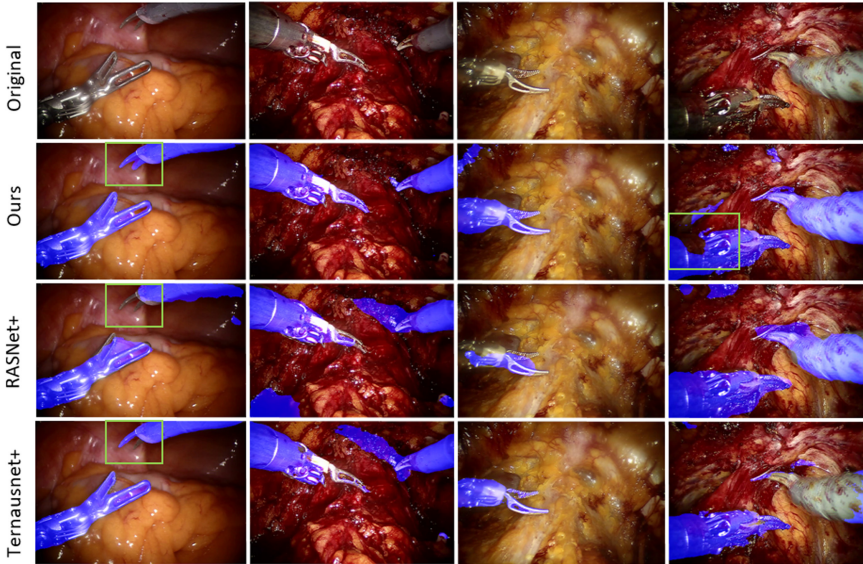
Fig. 3. Figure showing a qualitative comparison of our method with other methods. It can be seen that overall, our method preserves the shape of the instruments better with fewer false positives. (column 1) Inset showing preservation of instrument shape in our method. (column 4) Inset showing a failure case of our method.

5   Conclusion

We presented a joint generative and segmentation strategy, coSegGAN, that outperforms SOTA methods in its generalization capability to unlabelled domain data. The evaluated SOTA methods use separate I2I mapped data augmentation and segmentation steps. The proposed losses helped to preserve finer tool structure. The method is easy to adapt to other deep learning segmentation methods and thus can significantly improve the existing methods. The method aims to utilize unlabelled surgical data, which is much easier to acquire than labelled data, to improve any instrument segmentation model in a simple yet effective manner. Therefore, coSegGAN has the potential to significantly facilitate surgical translation of current and future surgical tool segmentation methods because it effectively alleviates the problem of unlabelled data. Current testing of coSegGAN has been limited to footage from prostatectomy procedures. A thorough performance analysis for different types of RAS surgeries is part of future work.

References

1. Allan, M., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
2. Attia, M., Hossny, M., Nahavandi, S., Asadi, H.: Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder. In: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 3373–3378. IEEE (2017)

3. Colleoni, E., Edwards, P., Stoyanov, D.: Synthetic and real inputs for tool segmentation in robotic surgery. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 700–710. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_67

4. Colleoni, E., Stoyanov, D.: Robotic instrument segmentation with image-to-image translation. IEEE Robot. Autom. Lett. **6**(2), 935–942 (2021)

5. Iglovikov, V., Shvets, A.: TernausNet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv:1801.05746 (2018)

6. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. arXiv preprint arXiv:1904.06991 (2019)

7. Lin, S., Qin, F., Li, Y., Bly, R.A., Moe, K.S., Hannaford, B.: LC-GAN: image-to-image translation based on generative adversarial network for endoscopic images. arXiv preprint arXiv:2003.04949 (2020)

8. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)

9. Maier-Hein, L., et al.: Surgical data science for next-generation interventions. Nature Biomed. Eng. **1**(9), 691–696 (2017)

10. Pakhomov, D., Premachandran, V., Allan, M., Azizian, M., Navab, N.: Deep residual learning for instrument segmentation in robotic surgery. In: Suk, H.-I., Liu, M., Yan, P., Lian, C. (eds.) MLMI 2019. LNCS, vol. 11861, pp. 566–573. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32692-0_65

11. Pauly, O., Diotte, B., Habert, S., Weidert, S., Euler, E., Fallavollita, P., Navab, N.: Relevance-based visualization to improve surgeon perception. In: Stoyanov, D., Collins, D.L., Sakuma, I., Abolmaesumi, P., Jannin, P. (eds.) IPCAI 2014. LNCS, vol. 8498, pp. 178–185. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07521-1_19

12. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

13. Ross, T., et al.: Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. Int. J. Comput. Assist. Radiol. Surg. **13**(6), 925–933 (2018). https://doi.org/10.1007/s11548-018-1772-0

14. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: the missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)

15. Zhou, S., Gordon, M.L., Krishna, R., Narcomey, A., Fei-Fei, L., Bernstein, M.S.: HYPE: a benchmark for human eye perceptual evaluation of generative models. arXiv preprint arXiv:1904.01121 (2019)

16. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)

17. Zou, K.H., et al.: Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. Acad. Radiol. **11**(2), 178–189 (2004)