This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Learning Representations by Maximizing Mutual Information Across Views for Medical Image Segmentation

Weihao Weng and Xin Zhu*

The University of Aizu, Japan
zhuxin7703@gmail.com

Abstract. We propose a method that leverages multiple identical network structures to generate and process diverse augmented views of the same medical image sample. By employing contrastive learning, we maximize mutual information among features extracted from different views, ensuring the networks learn robust and high-level semantic representations. Results from testing on four public and one private endoscopic surgical tool segmentation datasets indicate that the proposed method outperformed state-of-the-art semi-supervised and fully supervised segmentation methods. After trained by 5% labeled training data, the proposed method achieved an improvement of 11.5%, 8.4%, 6.5%, and 5.8% on RoboTool, Kvasir-instrument, ART-NET, and FEES, respectively. Ablation studies were also performed to measure the effectiveness of each proposed module. Code is available at Mutual-Exemplar.

Keywords: Contrastive learning · Con-training · Medical image classification · Semi-supervised learning.

1 Introduction

Recent research indicates that many states in the United States are experiencing doctor shortages [20]. Computer-aided diagnosis (CAD) may be performed on a par with experienced doctors. Accurate segmentation of surgical tools is of great value in developing CAD systems to automate a range of clinical procedures. The dominant approach in medical image segmentation involves utilizing an encoder-decoder architecture (e.g., UNet [17] and its variants, which incorporate multi-scale processing[21] and attention mechanisms [6]), and categorize each pixel into specific classes. In a fully-supervised training context where substantial labeled data are available, the majority of current algorithms exhibit satisfying performance. However, the requirement for annotated data imposes additional burdens on medical experts, which contradicts our initial intention of reducing their workload. To address this challenge, several purely unsupervised learning [3] methods have been developed. Typically, the segmentation accuracy of unsupervised learning methods has no standard in a wide range of image applications. Therefore, semi-supervised learning (SSL) is a preferred approach

[15] through training by combining a small amount of labeled images and a large quantity of unlabeled images.

The state-of-the-art semi-supervised learning methods include cross pseudo [7], mean teacher [18], deep co-training [16], and contrastive learning [4]. Cross pseudo trains two perturbed networks on the same data, using pseudo-label to learn the unlabeled data. Mean teacher trains two networks on the same data with independent augmentation or noise. Teacher network updates with the exponential moving average weights of the student networks, making its predictions less sensitive to noise. These stable predictions are then used to train the student network. Deep co-training trains multiple networks on different subsets of the data. It prevents the networks from getting stuck in a local minimum. However, cross pseudo heavily relies on the accuracy of pseudo labels; mean teacher significantly depends on the learning ability of the student network; deep co-training strongly depends on each subset providing complementary information not available in the others. Therefore, contrastive learning is preferred in practice. By pulling positive pairs together and pushing negative pairs apart in an embedding space, contrastive learning excels at learning efficient representations. The state-of-the-art of contrastive learning for medical image segmentation is Pseudo-CL, which combines contrastive learning with pseudo labels. Our method is proposed as a semi-supervised technique to improve co-training with contrastive learning. Mutual role models are inspired by Min-Max Similarity [13] by training two identical networks on different data. It measures the similarity between the outputs of the decoders in the two networks using both Intersection over Union (IoU) loss and binary cross-entropy (BCE) loss, and measures the similarity between the outputs of the projectors in the two networks with contrastive loss.

Our contributions lie in addressing two challenges in co-training:

1. *Data Waste*: traditional co-training involves splitting the dataset into several subsets and training multiple networks concurrently (hence the term "co"). Each network utilizes only one subset. Since each network trains on a different subset, the features extracted by them are distinct. The issue arises because, in pursuing different features, each network does not have access to the entire dataset. Typically, when two networks are used, either training on half of the labeled data, either network wastes half of the labeled data. In contrast, our proposed method uses data augmentation to obtain different features, allowing each network to utilize entire labeled data, thereby eliminating the problem of data waste.
2. *Wrong Exemplar*: contrastive learning-based co-training trains two networks simultaneously. If one network makes high confidence but incorrect prediction, using contrastive learning to make the features produced by the other network more similar to those of the erroneous network can lead us further away from correct answers. Traditionally, using two networks is an essential compromise because the labeled data is divided into subsets. If divided into too many subsets, each network would receive insufficient data. Our approach employs more networks to mitigate the impact of individual network errors on overall training.

2 Methodology

This section introduces the proposed method, named Mutual Exemplar, illustrated conceptually in Fig. 1. Let $\mathcal{D} = \mathcal{X} \cup \mathcal{U}$ be a dataset for training, where $\mathcal{X} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{m}$ and $\mathcal{U} = \{\mathbf{U}_i\}_{i=1}^{n}$ denotes the labeled and unlabelled datasets, respectively. Mutual Exemplar trains three networks $\{\mathcal{F}_1(\cdot), \mathcal{F}_2(\cdot), \mathcal{F}_3(\cdot)\}$ simultaneously. Mutual Exemplar applies different augmentations for the same images from a mini-batch to build three different views and learn general features. Specifically, applying weak augmentation (flips) for $\mathcal{F}_1$, applying moderate augmentation (weak augmentation plus affine transformations, random grayscale noise, gaussian blur, color jitter) for $\mathcal{F}_2$, and applying strong augmentation (moderate augmentation plus rotations and GridMask) for $\mathcal{F}_3$.

In each epoch, we ensure that different data (whether labeled or unlabeled data) are input into different networks. For the labeled data, we use the widely used [17] combination of Intersection over Union (IoU) loss and the binary cross-entropy (BCE) loss $\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{IoU}} + \mathcal{L}_{\text{BCE}}$,

$$\mathcal{L}_{\text{sup}} = -\left(X \log(Y) + (1 - X) \log(1 - Y)\right) + \left(1 - \frac{X \cdot Y}{X + Y - X \cdot Y}\right). \quad (1)$$

For the unlabeled data, the network outputs a feature map from the projector after the first layer of the encoder $E(Aug(U))$ and a prediction from the classifier after the decoder $P(Aug(U))$. The feature map is saved as pseudo-labeled feature map Y^f and prediction as pseudo-labeled prediction Y^p. In later epochs, if the network outputs a feature map and prediction for the unlabeled data with a lower supervised loss, the pseudo-labels are updated with the new feature map and prediction. The unsupervised loss for the unlabeled data $\mathcal{L}_{\text{unsup}}$ consists of a soft supervised loss $\mathcal{L}_{\text{soft}}$ and a contrastive loss $\mathcal{L}_{\text{cont}}$. $\mathcal{L}_{\text{soft}} = \mathcal{L}_{\text{IoU}} + \mathcal{L}_{\text{BCE}}$, where the saved prediction Y^p is utilized as the pseudo-label,

$$\mathcal{L}_{\text{soft}} = -\left(X \log(Y^p) + (1 - X) \log(1 - Y^p)\right) + \left(1 - \frac{X \cdot Y^p}{X + Y - X \cdot Y^p}\right). \quad (2)$$

Let q^{f_i} denotes the feature of a network's feature map $E_i(\text{Aug}_i(U))$, $k_{\neq i}$ denotes the feature of another network's feature map $E_{\neq i}(\text{Aug}_{\neq i}(U))$, and q^{Y^f} denotes the feature of the pseudo-labeled feature map Y^f. Different networks have different input data, so features from other networks are used to build negative pairs $\{q^{f_i} \cdot k_{\neq i}\}$. Due to structural similarities across medical images, segmentation results from the same locations but different images are likely to be similar. Therefore, we avoid using features from the same location of different images to build the negative pair.

We use q^{Y^f} to build the positive pair $\{q^{f_i} \cdot q^{Y^f}\}$, and use $k_{\neq i}$ to build the negative pairs $\{q^{f_i} \cdot k_{\neq i}\}$. The contrastive loss $\mathcal{L}_{\text{cont}}$ is defined as

$$\mathcal{L}_{\text{cont}} = -\log \frac{\exp\left(q^{f_i} \cdot q^{Y^f} / \tau\right)}{\exp\left(q^{f_i} \cdot q^{Y^f} / \tau\right) + \sum_{k_{\neq i}} \exp\left(q^{f_i} \cdot k_{\neq i} / \tau\right)}, \quad (3)$$

where the temperature constant $\tau = 0.07$.

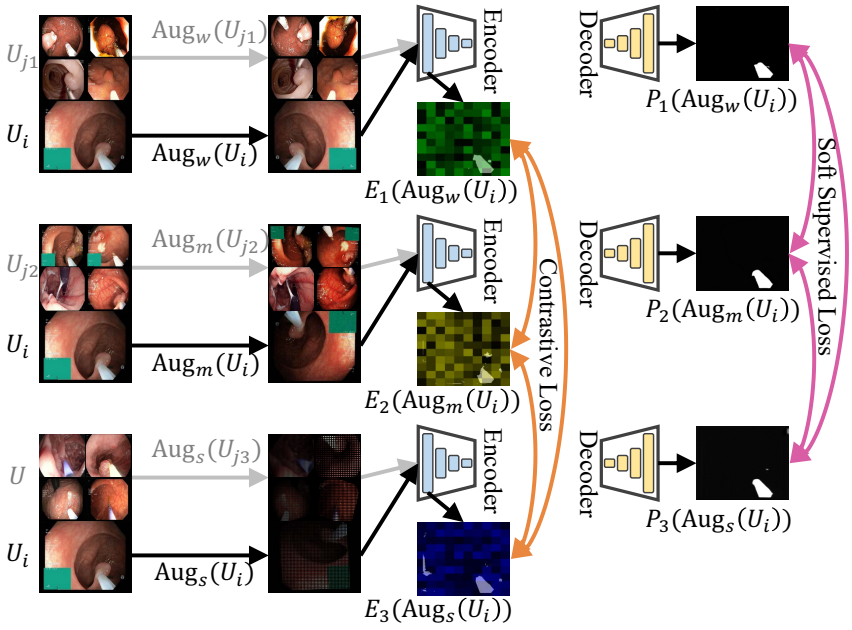

Fig. 1. Example of the proposed method applied to unlabeled data.

3 Experiments and Results

3.1 Experiment Settings

To ensure fairness in the comparison, we use the same network backbone developed in PyTorch [14] and implementation details as the state-of-the-art Co-training method Min-Max Similarity. The networks $\{\mathcal{F}_1(\cdot), \mathcal{F}_2(\cdot), \mathcal{F}_3(\cdot)\}$ share the same UNet type model with an ImageNet-pretrained Res2Net [8] encoder. The projector and classifier contains 3 and 2 stages, respectively. Each stage contains a convolution layer and a max pooling layer. The optimizer is Adam [12] (learning rate = 0.0001, $\beta_1 = 0.99$, $\beta_2 = 0.999$).

3.2 Datasets

Table 1 lists surgical tool datasets in this study. We use 4 public and 1 private datasets for this study. Private Flexible Endoscopic Evaluation of Swallowing (FEES) dataset [19] contains FEES videos from 100 patients. We use this dataset to evaluate the performance of each method on the multiclass segmentation task.

3.3 Competing Methods and Evaluation Metrics

Competing methods include the current state-of-the-art methods mentioned in the Section 1, including fully supervised UNet, UNet++, TransUNet, Cross

Table 1. List of surgical tool datasets, ordered by total number of images.

Dataset	Brief description	training set	test set
RoboTool [9]	514 images from 20 robotic surgical videos.	412 images.	102 images.
Kvasir-instrument [11]	590 gastrointestinal endoscopic images with surgical tools.	472 images.	118 images.
ART-NET [10]	816 images from 29 laparoscopic videos.	662 images.	154 images.
EndoVis'17 [1]	1,800 images from 8 robotic surgical videos.	1,575 images.	225 images.
FEES (private)	10,000 images from 100 FEES videos	8,000 images.	2,000 images.

Pseudo, Mean Teacher, Deep Co-training, Pseudo-CL and Min-Max Similarity. The quantitative performance is evaluated by Dice Sørensen coefficient (DSC), mean absolute error (MAE) and F1-score.

3.4 Experimental Results

The total numbers of images in 4 public datasets are ordered as - RoboTool <Kvasir-instrument <ART-NET <EndoVis'17. It is observed from Tables 2 and 3 that Mutual Exemplar demonstrates superior performance over other methods, particularly as the number of images available for training decreases. On the RoboTool and Kvasir-instrument datasets, we can observe a significant improvement. In the RoboTool and Kvasir-instrument datasets, the DSC of Mutual Exemplar over the second-best method gradually increases with a decrease in labeled data. On the RoboTool dataset, the DSC increases 2.1-11.5%. On the Kvasir-instrument dataset, the DSC increases 0.9-8.4%.

On the ART-NET dataset, Mutual Exemplar achieves the best performance when the networks are trained with 5 and 20% labeled training sets. For the 50% labeled training sets, the performance of Mutual Exemplar and the second-best method are closely matched. On the EndoVis'17 dataset, which contains a relatively abundant supply of labeled images for training, the performance of Mutual Exemplar is similar to that of the best competing method, lagging slightly by 0.6 and 1.1% for the 5 and 20% labeled training sets, respectively.

Performance Comparison on FEES dataset. Results from Tables 2 and 3 show that the performance of fully supervised UNet, UNet++, and TransUNet do not significantly differ. We argue that the four public segmentation tasks are relatively easy in extracting meaningful features. On the FEES dataset, fully supervised UNet++ and TransUNet significantly outperform UNet. Furthermore, increasing the label ratio from 20% to 50% does not enhance UNet's performance to the same extent as UNet++ and TransUNet, suggesting that UNet's performance is constrained by its simpler network structure. Results from Table 4 show that Mutual Exemplar and Pseudo-CL consistently outper-

Table 2. Segmentation results of RoboTool and Kvasir-instrument datasets. For networks trained on various label-ratio training sets, the best result is in **bold** and the second-best result is underlined, with evaluations performed on the test set. Performance differences between Mutual Exemplar and the best competing methods are shown under the Mutual Exemplar's results (↑ better, ↓ worse.)

Dataset	Method	DSC			MAE		
RoboTool	Fully Supervised UNet	0.786			0.088		
	Fully Supervised UNet++	0.807			0.068		
	Fully Supervised TransUNet	0.808			0.063		
	Label ratio l_a	5%	20%	50%	5%	20%	50%
	UNet	0.516	0.661	0.730	0.075	0.133	0.105
	UNet++	0.500	0.691	0.734	0.152	0.098	0.087
	TransUNet	0.516	0.718	0.732	0.123	0.087	0.090
	Mean Teacher	0.575	0.742	0.784	0.137	0.074	0.061
	Deep Co-training	0.519	0.714	0.752	0.143	0.080	0.068
	Cross Pseudo	0.559	0.711	0.758	0.147	0.083	0.069
	Min-Max Similarity	0.646	<u>0.781</u>	<u>0.831</u>	0.104	0.058	0.046
	Pseudo-CL	<u>0.650</u>	0.771	0.801	<u>0.098</u>	<u>0.056</u>	<u>0.045</u>
	Mutual Exemplar	**0.725**	**0.807**	**0.848**	**0.079**	**0.051**	**0.042**
		↑ 11.5%	↑ 3.3%	↑ 2.1%	↑ 19.4%	↑ 8.9%	↑ 6.7%
Kvasir-instrument	Fully Supervised UNet	0.901			0.027		
	Fully Supervised UNet++	0.893			0.023		
	Fully Supervised TransUNet	0.905			0.015		
	Label ratio l_a	5%	20%	50%	5%	20%	50%
	UNet	0.706	0.730	0.799	0.075	0.055	0.043
	UNet++	0.567	0.736	0.823	0.085	0.041	0.028
	TransUNet	0.541	0.753	0.867	0.093	0.029	0.015
	Mean Teacher	0.605	0.788	0.892	0.065	0.031	0.020
	Deep Co-training	0.489	0.764	0.866	0.084	0.045	0.027
	Cross Pseudo	0.709	0.824	0.894	0.051	0.037	0.020
	Min-Max Similarity	<u>0.776</u>	<u>0.874</u>	<u>0.925</u>	<u>0.043</u>	<u>0.024</u>	<u>0.013</u>
	Pseudo-CL	0.720	0.819	0.910	0.062	0.041	0.023
	Mutual Exemplar	**0.841**	**0.899**	**0.933**	**0.039**	**0.020**	**0.012**
		↑ 8.4%	↑ 2.9%	↑ 0.9%	↑ 9.3%	↑ 16.7%	↑ 7.7%

form the fully supervised UNet, indicating they enable the network to learn better representations, thereby augmenting UNet's capabilities.

Ablation study. Mutual Exemplar introduces four main modifications over the state-of-the-art Co-training method. The first modification is ensuring that each network learns from the entire training dataset (Entire training), the scond is using three networks (Tri-view), the third is applying three augmentations (Tri-Aug), and the fourth is employing a method similar to pseudo-label (P-label). Using Tri-view without Tri-Aug is achieved by having two out of the three networks use the same strong augmentation. Employing Tri-Aug without Tri-view is also achieved by three networks. One network uses strong augmentation, and another uses moderate augmentation. The weights of these two networks are determined by their average. Results in Table 4 indicate that each proposed

Table 3. Segmentation results of ART-NET and EndoVis'17 datasets.

Task	Method	DSC			MAE		
ART-NET	Fully Supervised UNet		0.894			0.029	
	Fully Supervised UNet++		0.908			0.023	
	Fully Supervised TransUNet		0.904			0.019	
	Label ratio l_a	5%	20%	50%	5%	20%	50%
	UNet	0.660	0.713	0.812	0.072	0.062	0.038
	UNet++	0.717	0.761	0.866	0.053	0.051	0.030
	TransUNet	0.685	0.764	0.841	0.047	0.043	0.032
	Mean Teacher	0.747	0.835	0.889	0.051	0.033	0.021
	Deep Co-training	0.726	0.820	0.875	0.049	0.033	0.021
	Cross Pseudo	0.759	0.824	0.874	0.047	0.035	0.023
	Min-Max Similarity	0.784	0.869	0.917	0.045	0.029	**0.017**
	Pseudo-CL	0.756	0.835	0.890	0.050	0.032	0.029
	Mutual Exemplar	**0.835**	**0.890**	**0.919**	**0.032**	**0.021**	0.018
		↑ 6.5%	↑ 2.4%	↑ 0.2%	↑ 28.9%	↑ 27.6%	↓ 5.9%
EndoVis'17	Fully Supervised UNet		0.894			0.027	
	Fully Supervised UNet++		0.909			0.026	
	Fully Supervised TransUNet		0.904			0.029	
	Label ratio l_a	5%	20%	50%	5%	20%	50%
	UNet	0.823	0.869	0.885	0.057	0.040	0.029
	UNet++	0.825	0.882	0.890	0.058	0.044	0.041
	TransUNet	0.837	0.873	0.882	0.047	0.039	0.035
	Mean Teacher	0.875	0.901	0.910	0.037	0.028	0.024
	Deep Co-training	0.848	0.895	0.895	0.038	0.026	0.026
	Cross Pseudo	0.886	0.909	0.913	0.029	0.025	0.021
	Min-Max Similarity	0.909	**0.931**	**0.940**	0.023	**0.018**	**0.017**
	Pseudo-CL	**0.919**	0.922	0.929	0.024	0.023	0.021
	Mutual Exemplar	**0.919**	0.925	0.930	**0.021**	0.020	0.019
		↑ 0%	↓ 0.6%	↓ 1.1%	↑ 8.7%	↓ 11.1%	↓ 0.9%

method enhances segmentation performance to varying degrees. Notably, employing Tri-Aug and P-label without Tri-view essentially mirrors a mean teacher-like operation, where the expectation is that average weights make the network more robust. However, the outcome is lower performance. We argue that the averaging process dilutes the distinct perspectives introduced by augmentation.

Limitations and Future Works We conducted experiments of Lesion Segmentation using datasets from Fluorescence Microscopy [5], Heart MRI [2], and Spleen CT (details can be found in the *Supplementary Material*), finding that the proposed method did not show significant improvement over Pseudo-CL, which is the state-of-the-art contrastive learning method for lesion segmentation. Pseudo-CL can be considered as a combination of contrastive learning, pseudo-labeling, and mean teacher. The teacher network utilizes the mean weights of the student networks, making its pseudo-labels less sensitive to noise. We attempted to optimize the proposed Mutual Exemplar using the mean teacher concept; that is, we did not directly update the pseudo-labels with the new feature map and

Table 4. Segmentation results of FEES dataset and ablation study. ✓✓indicate that each network learns from the entire training dataset.

Task	Method			DSC			MAE		
FEES	Fully Supervised UNet			0.688			0.098		
	Fully Supervised UNet++			0.794			0.068		
	Fully Supervised TransUNet			0.803			0.063		
	Label ratio l_a			5%	20%	50%	5%	20%	50%
	UNet			0.655	0.679	0.682	0.102	0.099	0.098
	UNet++			0.652	0.713	0.776	0.101	0.083	0.068
	TransUNet			0.649	0.720	0.780	0.099	0.077	0.065
	Mean Teacher			0.651	0.657	0.704	0.097	0.087	0.079
	Deep Co-training			0.689	0.722	0.723	0.086	0.077	0.072
	Cross Pseudo			0.668	0.685	0.709	0.091	0.081	0.079
	Min-Max Similarity			0.658	0.677	0.691	0.101	0.094	0.091
	Pseudo-CL			0.736	0.748	0.792	0.081	0.065	0.056
	Mutual Exemplar			**0.779** ↑ 5.8%	**0.811** ↑ 8.4%	**0.837** ↑ 5.7%	**0.067** ↑ 17.3%	**0.053** ↑ 18.5%	**0.047** ↑ 16.1%
Ablation Study	Tri-view	Tri-Aug	P-label						
	✓		✓	0.622	0.732	0.763	0.107	0.077	0.070
	✓	✓		0.626	0.723	0.744	0.115	0.079	0.069
		✓	✓	0.649	0.761	0.806	0.113	0.069	0.059
	✓✓		✓✓	0.720	0.733	0.785	0.075	0.072	0.065
	✓✓	✓✓		0.694	0.731	0.771	0.080	0.070	0.066
		✓✓	✓✓	**0.753**	**0.799**	**0.828**	**0.070**	**0.055**	**0.050**

prediction, but instead used the average of the new feature map and prediction with the old feature maps and predictions. Unfortunately, this did not significantly enhance the segmentation performance of Mutual Exemplar. However, we believe that future research could explore combining the mean teacher approach.

Another reason the proposed method has no significant improvement in Lesion Segmentation accuracy may be that it requires each network to generate distinct features from images with different augmentations. In Surgical Tools Segmentation, the greater difference between tools and the human body, amplified by augmentations, provides more beneficial information for contrastive learning. Future work will explore the impact of augmentations on feature generation to improve Mutual Exemplar's performance in Lesion Segmentation.

4 Conclusion

We propose Mutual Exemplar, a semi-supervised image segmentation method that employs contrastive learning to simultaneously train multiple networks with identical structures. Mutual Exemplar outperforms all competing semi-supervised methods across four public and one private surgical tools segmentation datasets.

Acknowledgment. We would like to thank Dr. Faouzi Alaya Cheikh for the insightful discussions and Dr. Mohib Ullah for helping with refining and improving the manuscript. We also extend our thanks to Dr. Mitsuyoshi Imaizumi for providing the FEES data. This study was partially supported by the Competitive Research Fund of The University of Aizu Grant Number P-7-2024, and JSPS KAKENHI Grant Number 24K12677.

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

References

1. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
2. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
3. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. Advances in neural information processing systems **32** (2019)
4. Basak, H., Yin, Z.: Pseudo-label guided contrastive learning for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19786–19797 (2023)
5. Caicedo, J.C., Goodman, A., Karhohs, K.W., Cimini, B.A., Ackerman, J., Haghighi, M., Heng, C., Becker, T., Doan, M., McQuin, C., et al.: Nucleus segmentation across imaging experiments: the 2018 data science bowl. Nature methods **16**(12), 1247–1253 (2019)
6. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
7. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2613–2622 (2021)
8. Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P.: Res2net: A new multi-scale backbone architecture. IEEE transactions on pattern analysis and machine intelligence **43**(2), 652–662 (2019)
9. Garcia-Peraza-Herrera, L.C., Fidon, L., D'Ettorre, C., Stoyanov, D., Vercauteren, T., Ourselin, S.: Image compositing for segmentation of surgical tools without manual annotations. IEEE transactions on medical imaging **40**(5), 1450–1460 (2021)
10. Hasan, M.K., Calvet, L., Rabbani, N., Bartoli, A.: Detection, segmentation, and 3d pose estimation of surgical tools using convolutional neural networks and algebraic geometry. Medical Image Analysis **70**, 101994 (2021)
11. Jha, D., Ali, S., Emanuelsen, K., Hicks, S.A., Thambawita, V., Garcia-Ceja, E., Riegler, M.A., de Lange, T., Schmidt, P.T., Johansen, H.D., et al.: Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In: MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27. pp. 218–229. Springer (2021)

12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Lou, A., Tawfik, K., Yao, X., Liu, Z., Noble, J.: Min-max similarity: A contrastive semi-supervised deep learning network for surgical tools segmentation. IEEE Transactions on Medical Imaging (2023)
14. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
15. Peiris, H., Chen, Z., Egan, G., Harandi, M.: Duo-segnet: adversarial dual-views for semi-supervised medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 428–438. Springer (2021)
16. Qiao, S., Shen, W., Zhang, Z., Wang, B., Yuille, A.: Deep co-training for semi-supervised image recognition. In: Proceedings of the european conference on computer vision (eccv). pp. 135–152 (2018)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
18. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
19. Weng, W., Zhu, X., Imaizumi, M., Murono, S.: Fees-is: Real-time instance segmentation of flexible endoscopic evaluation of swallowing. In: 2023 11th European Workshop on Visual Information Processing (EUVIP). pp. 1–6. IEEE (2023)
20. Zhang, X., Lin, D., Pforsich, H., Lin, V.W.: Physician workforce in the united states of america: forecasting nationwide shortages. Human resources for health **18**(1), 1–9 (2020)
21. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 3–11. Springer (2018)