Min-Max Similarity: A Contrastive Learning Based Semi-Supervised Learning Network for Surgical Tools Segmentation

Ange Lou¹, Xing Yao², Ziteng Liu², Jack Noble¹ ²
¹ Department of Electrical and Computer Engineering, Vanderbilt University
² Department of Computer Science, Vanderbilt University
{ange.lou, xing.yao, ziteng.liu, jack.noble}@vanderbilt.edu

Abstract: Segmentation of images is a popular topic in medical AI. This is mainly due to the difficulty to obtain a significant number of pixel-level annotated data to train a neural network. To address this issue, we proposed a semi-supervised segmentation network based on contrastive learning. In contrast to the previous state-of-the-art, we introduce a contrastive learning form of dual-view training by employing classifiers and projectors to build all-negative, and positive and negative feature pairs respectively to formulate the learning problem as solving min-max similarity problem. The all-negative pairs are used to supervise the networks learning from different views and make sure to capture general features, and the consistency of unlabeled predictions is measured by pixel-wise contrastive loss between positive and negative pairs. To quantitative and qualitative evaluate our proposed method, we test it on two public endoscopy surgical tool segmentation datasets and one cochlear implant surgery dataset which we manually annotate the cochlear implant in surgical videos. The segmentation performance (dice coefficients) indicates that our proposed method outperforms state-of-the-art semi-supervised and fully supervised segmentation algorithms consistently. The code is publicly available at: https://github.com/AngeLouCN/Min_Max_Similarity

1 Introduction

Semantic tool segmentation in the surgical video is important for surgical scene understanding and computer-assisted interventions as well as for the development of robotic automation [1]. However, the robustness and performance of algorithms are highly affected by different illumination conditions, bleeding, smoke and occlusions. Although a large amount of annotated data can alleviate this problem to some extent, they are very expensive and time-consuming to be collected in the medical area, especially for the surgical video data. Besides, accurate semantic segmentation for the surgical tool is meaningful for the minimally invasive surgery (MIS), which causes less pain, reduced time of the hospital stay, fast recovery, reduced blood loss, and less scaring process. The operation is complex, and the surgeons have to precisely tackle hand-eye coordination, which may lead to restricted mobility and a narrow field of view [2].

Recently, the most effective semantic segmentation approach is using deep neural networks. Various architectures have been developed including novel encoder-decoder-based networks [3,4,5], pre-trained feature extractors-based networks [6], and attention-based networks [7,8]. Most of the existing algorithms are breaching through supervised settings when sizeable, annotated data available. However, annotating medical images requires expert knowledge and cannot be crow-source. Thus, the scale of datasets is limited in the medical field, especially in the surgical video datasets. And most of the existing networks perform not good on those small-scale medical datasets. To solve the limitation of annotated data, some purely unsupervised learning approaches [9,10] have been developed. Generally, those unsupervised learning approaches are not acceptable in medical image segmentation due to their low accuracy. As such, semi-supervised learning (SSL) become a more promising method, which can use the limited number of pixel-level annotated images available along with the large quantities of broader.

There are lots of semi-supervised segmentation techniques like mean teacher [14] and Duo-SegNet [15] that have been approved workable on medical image segmentation. Mean teacher is a two-network system in which the student network is trained in a stochastic manner and the teacher network is updated from student’s parameters. But the mean teacher cannot correct errors during training, because the teacher network is totally dependent on the updates of the student. Duo-SegNet solves this problem because of the multi-view learning strategy [16]. Both networks are simultaneously train and models can learn from each other during the training. However, the critic in Duo-SegNet only updated from the loss between labeled predictions and their ground truth which wastes large amounts of features from unlabeled data.

Contrastive learning plays a dominant role in lots of downstream computer vision tasks which lack annotated data. The intuition of this approach is that different transformations of an image should have similar feature representations and that these representations should be dissimilar from those of a
different image [11]. And then, a suitable contrastive loss [12,13] is formulated to measure this intuition and a neural network is training with unlabeled data to minimize this loss. Through contrastive learning, a large amount of unlabeled can contributes to training. There are lots of contrastive learning works that have achieved fully-supervised-level performance, like MoCo [25] and SimCLR [26]. The architecture of MoCo is similar to mean teacher, only one of the networks is updated from the training. SimCLR applies different augmentation for the same images from minibatch to build two different views and learn general features. Although these two state-of-the-art achieve compatible performance with fully supervised ResNet-50 [27] on ImageNet, both of them use ResNet-50 with a 4 × hidden layer widths which make networks have 375 million parameters.

In this paper, we propose a novel contrastive learning-based semi-supervised technique – Min-Max Similarity – to segment the surgical tools. Unlike the existing contrastive learning state-of-the-art methods which build different views [16] by applying augmentation to original images. The Min-Max Similarity directly separate the labeled datasets into two totally different subsets (X_{1} \cap X_{2} = \emptyset) and then to predict the representation for an unlabeled image by using the segmentation networks which are trained from these two subsets. The contributions of our work are highlighted as follows:

- We are the first to use contrastive-learning-based semi-supervised technique to solve the problem of segmenting cochlear implants from cochlear implant surgery videos.
- We build a cochlear implant segmentation dataset from 5 different surgical videos with 40 frames each and fill the gap in this field.
- We proposed a novelty contrastive-learning-based semi-supervised segmentation framework that outperforms the most current state-of-the-art on two public surgical tool segmentation datasets and our cochlear implant segmentation dataset.

![Fig. 1](image)

**Fig. 1.** The architecture of Min-Max Similarity. \( \mathcal{F}_{1}, \mathcal{F}_{2}, \mathcal{C}_{1}, \mathcal{C}_{2} \) and \( \mathcal{P}_{1}, \mathcal{P}_{2} \) denote segmentation networks, classifiers and projectors, respectively. Here, classifiers extract features to build all-negative pairs, and projectors project unlabeled predictions to high dimension features for pixel-wise consistency measurement.

### 2. Methods

Here, we introduce the **Min-Max Similarity (MMS) algorithm** by providing an overview of the network architecture as shown in Fig. 1. Let \( \mathcal{X} = \{(X_{1}, Y_{1}), \ldots, (X_{m}, Y_{m})\} \) be a labeled set, where each pair \((X_{i}, Y_{i})\) consists of an image \( X_{i} \in \mathbb{R}^{C \times H \times W} \) and its ground truth \( Y_{i} \in \{0,1\}^{H \times W} \). The unlabeled images are assigned as \( \mathcal{U} = (U_{i})_{i=1}^{n} \), and \( U_{i} \in \mathbb{R}^{C \times H \times W} \) be a set of \( n \) unlabeled images which \( n \gg m \).

The MMS contains three modules, a dual view segmentation network, classifier and projector, shown by \( \mathcal{F}_{1}, \mathcal{F}_{2}, \mathcal{C}_{1}, \mathcal{C}_{2} \) and \( \mathcal{P}_{1}, \mathcal{P}_{2} \) respectively in Fig. 1. The segmentation networks \( \mathcal{F}_{1}, \mathcal{F}_{2} \) share the same encoder-decoder architecture. We leverage Res2Net [17] as an encoder which is pre-trained on ImageNet [18]. The decoder is similar to the U-Net, we applied four combinations of convolutional and up-sampling layers to predict the output with original resolution. Classifier and projector are
combinations of convolution and max-pooling layers. Classifier and projectors are the combinations of convolutional and max pooling layer, and we choose three and two max pooling respectively. The labeled set is equally separated to two subsets \( \{X_i\}_{i=1}^{m} = X_1 \cup X_2 \) and \( X_1 \cap X_2 = \emptyset \) , and to make sure segmentation networks learning from different views [16]. The predictions of labeled data \([\mathcal{F}_1(X_1), \mathcal{F}_2(X_2)]\) are sent to classifiers \([\mathcal{C}_1, \mathcal{C}_2]\) to generate classification feature vectors \([\mathcal{C}_1(\mathcal{F}_1(X_1)), \mathcal{C}_2(\mathcal{F}_2(X_2))]\) . The dissimilarity of \([\mathcal{C}_1(\mathcal{F}_1(X_1)), \mathcal{C}_2(\mathcal{F}_2(X_2))]\) is measured by \(\mathcal{L}_{\text{InfoNCE-sup}}\), which is defined as:

\[
\mathcal{L}_{\text{InfoNCE-sup}} = -\log \frac{\exp \left( \frac{q \cdot k_+}{\tau} \right)}{\sum_{i=0}^{K} \exp \left( \frac{q \cdot k_i}{\tau} \right)} - \log \frac{1}{\sum_{i=0}^{K} \exp \left( \frac{q \cdot k_i}{\tau} \right)}
\]

(Note: \(X_1 \cap X_2 = \emptyset\) means there are all-negative pairs, thus \(q \cdot k_+ = 0\)). And then labeled prediction \([\mathcal{F}_1(X_1), \mathcal{F}_2(X_2)]\) compared with its ground truth by supervised loss \(\mathcal{L}_{\text{sup}}\). The supervised loss is defined as:

\[
\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{IoU}}^w + \mathcal{L}_{\text{BCE}}^w
\]

where \(\mathcal{L}_{\text{IoU}}^w\) and \(\mathcal{L}_{\text{BCE}}^w\) represent the weighted IoU loss and binary cross-entropy (BCE) loss. Due to the size and blurred boundary of surgical tools, the weighted loss pays more attention to hard pixels rather than assigning all pixels equal weights. The unlabeled images are applied heavily augmentation and sent to two segmentation networks after the labeled images. In the low level, we compare the similarity between two unlabeled predictions with similarity loss which is defined as:

\[
\mathcal{L}_{\text{similarity}} = \mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{IoU}}^w + \mathcal{L}_{\text{BCE}}^w
\]

Then the unlabeled predictions \([\mathcal{F}_1(U), \mathcal{F}_2(U)]\) go through their projectors to generate the high-level feature maps \([\mathcal{P}_1(\mathcal{F}_1(U)), \mathcal{P}_2(\mathcal{F}_2(U))]\). We use pixel-wise contrastive loss to measure the similarity of two high-level feature maps. The elements at the same spatial space are considered to be positive pair and others are negative pair, and the contrastive loss are measured by:

\[
\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp \left( \frac{q \cdot k_+}{\tau} \right)}{\sum_{i=0}^{K} \exp \left( \frac{q \cdot k_i}{\tau} \right)}
\]

Therefore, the two segmentation networks are updated by the sum of these four losses:

\[
\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{sup}} + \lambda_2 \mathcal{L}_{\text{InfoNCE-sup}} + \lambda_3 \mathcal{L}_{\text{similarity}} + \lambda_4 \mathcal{L}_{\text{InfoNCE}}
\]

The overall MMS optimization process is summarized in Algorithm 1.

**Algorithm 1: Min-Max Similarity (training)**

**Input:** Define segmentation networks \([\mathcal{F}_i()]_{i=1}^{2}\), classifiers \([\mathcal{C}_i()]_{i=1}^{2}\), projectors \([\mathcal{P}_i()]_{i=1}^{2}\), batch size \(B\) and maximum epoch \(E_{\text{max}}\). Labeled images \(X = \{(X_1, Y_1), ..., (X_m, Y_m)\}\). Unlabeled images \(U = \{U_1, ..., U_n\}\) and two labeled sets \(X^1; X^2 \subset X\) and \(X^3 \cap X^2 = \emptyset\).

**Output:** parameters \([\theta_i]_{i=1}^{2}\) of \([\mathcal{F}_i()]_{i=1}^{2}\);

**Initialization:** Initialize network, classifier and projector parameters \([\theta_i]_{i=1}^{2}\), \([\mu_i]_{i=1}^{2}\) and \([\nu_i]_{i=1}^{2}\);

for epoch = 1, ..., \(E_{\text{max}}\) do

for each batch \(B\) do

Generate predictions for labeled data \(\mathcal{F}_1(x)\) for all \(X_1 \in X^1, \mathcal{F}_2(x)\) for all \(X_1 \in X^2\) and then for unlabeled data \(\mathcal{F}_1(u)\) and \(\mathcal{F}_2(u)\) for all \(U_i \in U\);

Generate classifier and projector feature maps \(\mathcal{C}_1(\mathcal{F}_1(x)), \mathcal{C}_2(\mathcal{F}_2(x)), \mathcal{P}_1(\mathcal{F}_1(u))\), and \(\mathcal{P}_2(\mathcal{F}_2(u))\);

Let \(\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{sup}} + \lambda_2 \mathcal{L}_{\text{InfoNCE-sup}} + \lambda_3 \mathcal{L}_{\text{similarity}} + \lambda_4 \mathcal{L}_{\text{InfoNCE}}\), as defined in Equations (1) – (4);

Update \([\theta_i]_{i=1}^{2}\) by descending its gradient on \(\mathcal{L}_{\text{total}}\);

end

end

The overall MMS optimization process is summarized in Algorithm 1.
3. Related Work

We begin by briefly discussing the most recent semi-supervised learning method, namely mean teacher model [14], deep co-training [19], cross-pseudo supervision, and Duo-SegNet [15]. Our goal here is to discuss the differences and novelties between our proposed approach and some of the approaches that have been adopted for semi-supervised image segmentation.

In mean teacher, the model contains a student and a teacher network. During the training, the student network is supervised by two losses: 1). Segmentation loss between labeled predictions and their ground truth; 2). Consistency loss between unlabeled predictions with and without adding noise. And the teacher uses the exponential moving average (EMA) weights of the student model. The performance of the mean teacher is highly dependent on the learning ability of the student network. Deep co-training is inspired by the co-training framework to train multiple deep neural networks to be the different views and exploits adversarial examples to encourage view difference. Cross pseudo supervision (CPS) is a consistency regularization approach that is used to encourage high similarity between the predictions of two perturbed networks for the same input image and expand training data by using the unlabeled data with pseudo labels. The Duo-SegNet is also based on multi-view learning, the labeled and unlabeled predictions from the segmentation network are fed to critics to analyze the distribution. And the system is updated by minimizing the sum of segmentation loss and adversarial loss.

Unlike the mean teacher, the backbone of our MMS is a dual-view segmentation network, both segmentation networks are participated in training and share the information with each other. Compared with the only adding noise in the mean teacher method, MMS adds more heavy perturbations to unlabeled images to make sure the general features can be learned. Compared with deep co-training and Duo-SegNet, MMS does not spend more computational resource on predicted labels analysis. We focus on building positive and negative feature pairs to compute the consistency between two feature representations of predications. Avoid misleading the networks during the training, we fully use the properties of our labeled subsets (X_1 \cap X_2 = \emptyset) and contrastive learning to build the all-negative pairs, and to make sure the networks learn from different views. For the CPS, pseudo labels participate in supervision are not suitable for the small-scale medical datasets. Also, the errors in pseudo labels potentially fool the networks when there are not enough training data.

4. Experiments

**Implementation Details:** The Min-Max Similarity model is developed in PyTorch [21]. For training the segmentation networks, we use Adam optimizer.

**Datasets:** We use two public surgical tool segmentation datasets and one of our cochlear implant datasets: 1). Kvasir-instrument [22] consists of 590 annotated frames containing gastrointestinal (GI) procedure tools such as snares, balloons, biopsy forceps, etc. The training set contains 472 images, and the testing set contains 118 images. 2). EndoVis’17 [23] contains 8 \times 255-frame robotic surgical videos. We select the eighth video as testing set and the other 7 as training sets. It contains more than one type of surgical tools; we showed the details in Table 1. 3). Cochlear implant (CI) is considered the standard-of-care treatment for profound sensory-based hearing loss. With over 700,000 recipients worldwide, CIs are arguably the most successful neural prostheses to date. Since electrode insertion is a crucial factor affects the hear performance of CI recipients, we generate CI dataset which contains 30 frames-per-second (fps) surgical videos. Cochlear Implant dataset contains 5 \times 30 frames-per-second (fps) surgical videos. Our training set consists of 183 labeled images and 7497 unlabeled images. And the testing set contains 40 images with ground truth. For two public datasets, we test our method with 5%, 20% and 50% of labeled training sets. And we resize all images to a resolution of 512 \times 288.

| Video Number | Types of Surgical Tools          |
|--------------|----------------------------------|
| # 1          | Forceps                          |
| # 2          | Forceps                          |
| # 3          | Needle drivers                   |
| # 4          | Needle drivers and forceps       |
| # 5          | Forceps, retractors and vessel sealer |
| # 6          | Needle driver, scissors and forceps |
| # 7          | Forceps and vessel sealer        |
| # 8 (testing set) | Forceps, retractor, scissors and retractors |
Data Augmentation: To enhance the networks’ learning ability of general representation features, we heavily apply data augmentation techniques: rotation, flip, affine, random grayscale, gaussian blur, color jitter, and GridMask [24].

Competing Methods and Evaluation Metrics: We compare our proposed MMS with the current SOTA methods that include fully supervised UNet, mean teacher, deep co-training, cross pseudo supervision and Duo SegNet. All approaches are evaluated using the Dice Sørensen coefficient (DSC). The results of Kvasir-instrument and EndoVis’17 are shown in Table 2 and results of cochlear implant are shown in Table 3.

Performance Comparison: The qualitative and quantitative results comparison of the proposed method to four state-of-the-art methods are shown in Fig. 2, Table 2, and Table 3. The results reveal that the proposed method (MMS) outperforms the other selected SOTA methods on the three datasets. Especially for the small-scale dataset – Kvasir-instrument, which only contains 472 labeled training samples. The two segmentation networks of MMS are only supervised by 12 labeled images each (5% of Kvasir) but get 7%-37% improvements compared with other four methods. With the increment of labeled data, MMS always has advantages on segmentation performance. On Kvasir-instrument with 50% labeled images, MMS can reach 92.5% dice accuracy which is 2.4% and 3.1% higher than fully supervised UNet and most recent state-of-the-art method – Cross Pseudo Supervision. Also, it is easy to find that MMS are more robust than other methods from the second column in Fig. 2. MMS has less affected by the bubbles, which make the prediction contour (blue line) more fit the surgical tools. The advantage of our MMS is much more obvious on the results of EndoVis’17 datasets. In the experiment only with 5% and 20% labeled training data, it only contains one type of surgical tool – forceps but three types: retractors, forceps and scissors in the testing set (video #8). That is the reason why on 5% and 20% have similar performance, but our MMS also outperforms other methods. When the labeled data increase to 50%, labeled training set has one more type of surgical – needle drivers but networks still do not see the retractors and scissors. However, MMS have strong ability to capture the general features to recognize those tools never seen before. It is also easy to understand why the improvements between MMS and other methods are larger on EndoVis’17 dataset. On our cochlear implant dataset, there is only one type of tool – cochlear implant, so all of those semi-supervised method can reach a good result. However, only MMS could predict a complete and accurate contour for our cochlear implant. From Fig. 2, it is easy to find more details that MMS predicts the most accurate contour of surgical tools regardless of their size, type and number.

Table 2: Comparison with state-of-the-art methods on Kvasir-instrument and EndoVis’17.

| Dataset      | Method               | DSC       |
|--------------|----------------------|-----------|
| Kvasir-instrument | Fully supervised | 0.901 |
|              | l_0 = 5% l_0 = 20% l_0 = 50% |           |
|              | UNet                 | 0.706     | 0.730 | 0.799 |
|              | Mean Teacher         | 0.605     | 0.788 | 0.892 |
|              | Deep Co-training     | 0.489     | 0.764 | 0.866 |
|              | Cross Pseudo         | 0.709     | 0.824 | 0.894 |
|              | Duo-SegNet           | 0.403     | 0.834 | 0.861 |
|              | Min-Max Similarity (ours) | 0.776 | 0.874 | 0.925 |
|              | Fully supervised     |           |       |       |
|              | l_0 = 5% l_0 = 20% l_0 = 50% |           |
|              | EndoVis’17           |           |       |       |
|              | UNet                 | 0.761     | 0.710 | 0.822 |
|              | Mean Teacher         | 0.782     | 0.790 | 0.878 |
|              | Deep Co-training     | 0.737     | 0.734 | 0.843 |
|              | Cross Pseudo         | 0.768     | 0.777 | 0.832 |
|              | Duo-SegNet           | 0.814     | 0.788 | 0.879 |
|              | Min-Max Similarity (ours) | 0.837 | 0.838 | 0.921 |

Table 3: Comparison with state-of-the-art methods on Cochlear implant.

| Dataset      | Method               | DSC       |
|--------------|----------------------|-----------|
| Cochlear implant | Fully supervised | 0.863 |
|              | l_0 = 2.4% (183/7497) |           |
|              | Mean Teacher         | 0.908     |       |       |
|              | Deep Co-training     | 0.847     |       |       |
|              | Cross Pseudo         | 0.910     |       |       |
|              | Duo-SegNet           | 0.869     |       |       |
|              | Min-Max Similarity (ours) | 0.920 |       |       |
Ablation Study: We also perform ablation studies to show the effectiveness of adding classifiers and projectors. In our algorithm, we benefit from labeled and unlabeled data via 1). classifiers build negative pairs to encourage networks learning from different views of labeled data, 2). build positive and negative from representation of unlabeled images to analyze the pixel-level consistency which in essence minimizes errors. Also, we choose best values of weights $\lambda_1, \lambda_2, \lambda_3$ and $\lambda_4$. All experiments in Table 4 are conducted for Kvasir-instrument dataset with 50% of annotated data.

Table 4. Ablation study

(a). Network Structure Analysis.

| Experiment                          | DSC  |
|-------------------------------------|------|
| Min-Max Similarity w/o classifiers  | 0.925|
| w/o projectors & w/o classifiers    | 0.814|

(b). Hyper-parameter Analysis for loss weights.

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | DSC  |
|-------------|-------------|-------------|-------------|------|
| 0.2         | 0.2         | 0.3         | 0.3         | 0.919|
| 0.3         | 0.2         | 0.3         | 0.2         | 0.917|
| 0.5         | 0.1         | 0.3         | 0.1         | 0.915|
| 0.25        | 0.25        | 0.25        | 0.25        | 0.925|

5. Conclusion

We proposed a contrastive-learning-based algorithm for semi-supervised surgical tools segmentation and demonstrated its effectiveness on publicly available datasets and our own dataset. Compared with the
existing semi-supervised and contrastive learning method, which focus on how to measure the consistency between unlabeled predictions, we proposed the all-negative pair concept to encourage the networks leaning from different view and capture real general features. Also, we use pixel-wise contrastive loss to measure the pixel-level consistency to enhance the segmentation performance. From the experiments, MMS has been approved could accurately predict the tools which have never seen before but other compared methods failed. Also, MMS could achieve fully supervised performance by only using small amounts of labeled training set, especially in our cochlear implant dataset.

Furthermore, we find a lack of research and work in the field of cochlear implant segmentation area, which is crucial to the CI insert surgery. Therefore, we build a segmentation dataset from real world cochlear implant inserting surgical videos to fill this gap. Also, we proposed robust and accurate semi-supervised segmentation method to solve the limitations of annotated images from surgical videos and prove it performance on different datasets.

The MMS can still be improved by solving the bias between two labeled subsets, like one network sees a tool that another does not, which will be considered in our future work.
Reference

[1]. Colleoni, E., Edwards, P., & Stoyanov, D. (2020, October). Synthetic and real inputs for tool segmentation in robotic surgery. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 700-710). Springer, Cham.

[2]. Bernhardt, S., Nicolau, S. A., Soler, L., & Doignon, C. (2017). The status of augmented reality in laparoscopic surgery as of 2016. Medical image analysis, 37, 66-90.

[3]. Lou, A., Guan, S., & Loew, M. H. (2021, February). DC-UNet: rethinking the U-Net architecture with dual channel efficient CNN for medical image segmentation. In Medical Imaging 2021: Image Processing (Vol. 11596, p. 115962T). International Society for Optics and Photonics.

[4]. Ronneberger, O., Fischer, P., & Brox, T. (2015, October). Unet: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[5]. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In Deep learning in medical image analysis and multimodal learning for clinical decision support (pp. 3-11). Springer, Cham.

[6]. Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[7]. Fan, D. P., Ji, G. P., Zhou, T., Chen, G., Fu, H., Shen, J., & Shao, L. (2020, October). Pranet: Parallel reverse attention network for polyp segmentation. In International conference on medical image computing and computer-assisted intervention (pp. 263-273). Springer, Cham.

[8]. Lou, A., Guan, S., & Loew, M. (2021). CaraNet: Context Axial Reverse Attention Network for Segmentation of Small Medical Objects. arXiv preprint arXiv:2108.07368.

[9]. Sener, F., & Yao, A. (2018). Unsupervised learning and segmentation of complex activities from video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8368-8376).

[10].Zou, Y., Yu, Z., Kumar, B. V. K., & Wang, J. (2018). Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In Proceedings of the European conference on computer vision (ECCV) (pp. 289-305).

[11].Chaitanya, K., Erdil, E., Karani, N., & Konukoglu, E. (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. Advances in Neural Information Processing Systems, 33, 12546-12558.

[12].Hadsell, R., Chopra, S., & LeCun, Y. (2006, June). Dimensionality reduction by learning an invariant mapping. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 2, pp. 1735-1742). IEEE.

[13].Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

[14].Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems, 30.

[15].Peiris, H., Chen, Z., Egan, G., & Harandi, M. (2021, September). Duo-SegNet: Adversarial Dual-Views for Semi-Supervised Medical Image Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 428-438). Springer, Cham.

[16].Hussain, T., Muhammad, K., Ding, W., Lloret, J., Baik, S. W., & de Albuquerque, V. H. C. (2021). A comprehensive survey of multi-view video summarization. Pattern Recognition, 109, 107567.

[17].Gao, S. H., Cheng, M. M., Zhao, K., Zhang, X. Y., Yang, M. H., & Torr, P. (2019). Res2net: A new multi-scale backbone architecture. IEEE transactions on pattern analysis and machine intelligence, 43(2), 652-662.

[18].Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2020). Imagenet: A large-scale hierarchical image database, 2009. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 248-255).

[19].Qiao, S., Shen, W., Zhang, Z., Wang, B., & Yuille, A. (2018). Deep co-training for semi-
supervised image recognition. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 135-152).

[20] Chen, X., Yuan, Y., Zeng, G., & Wang, J. (2021). Semi-supervised semantic segmentation with cross pseudo supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2613-2622).

[21] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.

[22] Jha, D., Ali, S., Emanuelsen, K., Hicks, S. A., Thambawita, V., Garcia-Ceja, E., ... & Halvorsen, P. (2021, June). Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In International Conference on Multimedia Modeling (pp. 218-229). Springer, Cham.

[23] Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y. H., ... & Azizian, M. (2019). 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426.

[24] Chen, P., Liu, S., Zhao, H., & Jia, J. (2020). Gridmask data augmentation. arXiv preprint arXiv:2001.04086.

[25] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9729-9738).

[26] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

[27] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).