POLYCL: CONTEXT-AWARE CONTRASTIVE LEARNING FOR IMAGE SEGMENTATION

Aaron Moseley, Abdullah Al Zubaer Imran

University of Kentucky, Lexington, KY, USA

ABSTRACT

Medical image segmentation is one of the most important tasks in an imaging pipeline as it influences a number of image-guided decisions. To be effective, the standard fully-supervised segmentation approach requires a large amount of manually annotated training data. The expensive, time-consuming, and error-prone pixel-level annotation process hinders progress and makes it challenging to perform effective segmentations. It is, therefore, imperative that the models learn as efficiently as possible from the limited available data. Such limited labeled image segmentation can be facilitated by self-supervised learning (SSL), particularly contrastive learning via pre-training on unlabeled data and fine-tuning on limited annotations. To this end, we propose a *novel* self-supervised contrastive learning framework for medical image segmentation leveraging inherent relationships of different images, dubbed as *PolyCL*. Without requiring any pixel-level annotations or data augmentations, our PolyCL learns and transfers context-aware discriminant features useful for segmentation from an innovative surrogate, in a task-related manner. Experimental evaluations on the public LiTS dataset demonstrate significantly superior performance of PolyCL over multiple baselines in segmenting liver from abdominal computed tomography (CT) images, achieving a Dice improvement of up to 5.5%.

Index Terms— self-supervised learning, contrastive learning, medical imaging, segmentation, Computed tomography

1. INTRODUCTION

Medical image segmentation is essential in healthcare for clear interpretations of anatomical structures or lesions in images like computed tomography (CT) images, aiding accurate diagnoses and monitoring of health conditions. However, creating fully annotated datasets is expensive as it requires expert radiologists to label each pixel of a structure. This costliness limits the effectiveness of segmentation models, impeding their adoption in healthcare. Therefore, ongoing research is striving to optimize data usage, enabling models to effectively perform image segmentation even with reduced data [1, 2, 3]. Deviating from its supervised learning counterparts, self-supervised learning (SSL) has become an attractive direction to this end, e.g., [4]. In SSL, a model is first pre-trained on unlabeled samples to learn a pretext task and then fine-tuned on labeled samples to learn a downstream task for the actual evaluation.

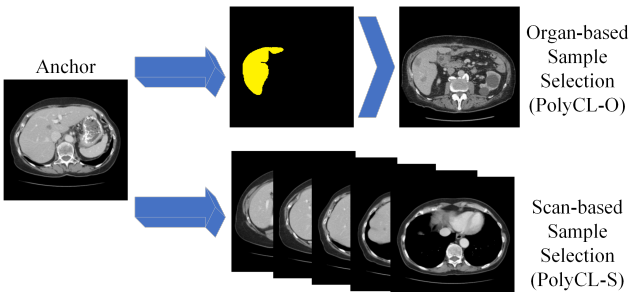

Fig. 1. Example selection strategies for the proposed PolyCL framework: PolyCL-O uses the information of whether each slice contains the organ-of-interest while PolyCL-S uses the information of the CT scan to which each slice belongs.

The representation learned by the model from the pre-training stage through a surrogate supervision (labels created from data itself), can be successfully transferred to various downstream tasks including segmentation [2].

A successful variant of self-supervised learning is contrastive learning which as a pre-training strategy helps cluster unlabeled examples in the latent space [5]. In contrastive learning-based pre-training, positive and negative image pairs are created for each slice in a dataset. A model is trained on aligning representations by increasing the similarity between positive pairs and increasing the difference between negative pairs in the embedding space. Most contrastive pre-training strategies use data augmentations to create positive examples, encouraging the model to represent transformed versions of the same image similarly [5, 6].

While it has shown benefits, this method fails to learn positive relationships between different images in the dataset. Previous work has investigated learning inter-scan relationships when looking at CT data, choosing positive example slices from similar locations within different scans [7]. This has been shown to be less data-hungry than standard supervised segmentation training but requires that the entire training dataset be obtained using the same modality as the images must be correctly aligned. Other papers have also looked at training the encoder to understand different views of the same subject or image [8] [9]. But these methods have not yet been adapted to medical images such as CT scans.

As mentioned, most previous work relies heavily on large pools of transformations to create positive examples for con-

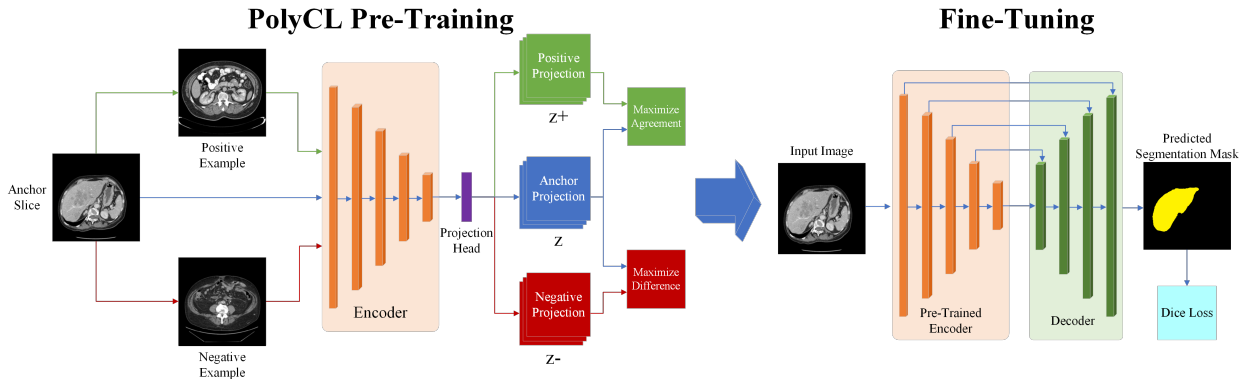

Fig. 2. Illustration of the proposed PolyCL-based medical image segmentation.

trastive learning. Because of this, a major gap in the medical imaging field is contrastive pre-training strategies that leverage useful downstream task-related surrogates, enabling the model to learn useful information about the target task before seeing fully-labeled data. Our proposed framework, PolyCL, introduces such a strategy through our organ-based and scan-based example selection methods. Our specific contributions are summarized as follows:

- A novel self-supervised contrastive learning-based pre-training approach for medical image segmentation
- Innovative example-selection strategies leveraging different amounts of labeled data
- Thorough experimentation demonstrating our framework's effectiveness and generalizability across multiple datasets

2. METHODS

To formulate the problem, we assume an unknown data distributions $p(X, Y)$ over images X and segmentation labels Y. As shown in Fig. 2, our proposed training framework, PolyCL, we use a two-stage training approach. We first employ self-supervised constrastive learning through our *innovative* sample selection strategy (see Fig. 1) to pre-train the encoder of the model. We then add a decoder and fine-tune the entire model on labeled data for the downstream segmentation task. We employ a U-Net [10] like encoder-decoder architecture for the PolyCL-based segmentation.

2.1. PolyCL Pre-Training

The pre-training process begins by using one of our two *novel* example selection strategies. For each anchor slice, $s \in X$, a positive example, $s^+ \in X$, and a negative example, $s^- \in X$, are selected. Because of the limited medical data availability, we devise an organ-based and a scan-based strategy for selecting examples, each requiring a different level of information.

Organ-based example selection (PolyCL-O): PolyCL-O requires the knowledge of which slices contain the target organ in the dataset. If the anchor slice contains the target organ, its positive example will also contain the target organ, while its negative example will not. The opposite is true for anchor slices that do not contain the target organ. By choosing examples in this manner, the encoder learns how to represent the target structure before seeing fully annotated data, improving its performance in the actual downstream task. In addition, random selection over all CT scans in the dataset teaches the model inter-scan invariance.

Scan-based example selection (PolyCL-S): PolyCL-S, on the other hand, requires no additional information. For each slice in the dataset, a positive example is selected randomly from the same scan and a negative example is selected from any scan different from the anchor. This process teaches the encoder intra-scan relationships and enables to understand the images even without knowledge of the target structure.

After example selection is completed for each slice in the dataset, we use the encoder f with projection head g to obtain embedding vectors for the anchor slice, $z = g(f(s))$, and its positive and negative examples, $z^+ = g(f(s^+))$ and $z^- = g(f(s^-))$. The contrastive loss is calculated to enforce similar embeddings between the anchor and its positive example, and dissimilar embeddings between the anchor and its negative example.

$$L_C = -log \frac{exp(sim(z, z^+)/\tau)}{exp(sim(z, z^+)/\tau) + exp(sim(z, z^-)/\tau)}, \quad (1)$$

where, $sim(a, b)$ denotes cosine similarity between a and b ($\frac{a \cdot b}{|a| \cdot |b|}$), M is the minibatch size, and the temperature parameter τ is set to $1/M$.

2.2. Fine-Tuning

After pre-training, the projection head g is discarded from the contrastive learning and the pretrained encoder is used as the feature extractor. We add a decoder mirroring the encoder with U-Net style skip connections to facilitate pixel-level semantic segmentation. We then fine-tune the entire model on the target segmentation task using a Dice-based loss function.

Table 1. Quantitative evaluation demonstrates superior performance of PolyCL in segmenting liver from abdominal CT images. Mean±stdev scores are reported by calculating Dice coefficient and Hausdorff distance after 10 runs of each of the models.

Model	LiTS		TotalSegmentator	
	Dice	Hausdorff	Dice	Hausdorff
Baseline	0.860 ± 0.036	12.769 ± 2.444	0.569 ± 0.206	28.609 ± 13.178
SimCLR	0.883 ± 0.022	8.691 ± 1.363	$\mathbf{0.670 \pm 0.074}$	21.259 ± 3.748
PolyCL-S	0.881 ± 0.025	$\mathbf{8.628 \pm 1.397}$	0.655 ± 0.055	18.566 ± 1.942
PolyCL-O	$\mathbf{0.907 \pm 0.009}$	8.880 ± 1.303	0.644 ± 0.092	$\mathbf{17.924 \pm 1.931}$

3. EXPERIMENTAL EVALUATION

3.1. Implementation Details

Data: We validated our proposed PolyCL by using 3 separate datasets at different data settings: Liver Tumor Segmentation (LiTS) [11], TotalSegmentator [12], and Medical Segmentation Decathlon (MSD) [13]. First, we used 17 abdominal CT scans from the LiTS dataset and split them into train/val/test: 10/2/5 ((1,664/298/787 slices). We randomly selected 34 scans from the TotalSegmentator dataset and split them into train/val/test: 10/4/20 (839/336/1324 slices). A smaller training set was used to simulate pre-training on a large dataset with no labels, then fine-tuning it on a smaller fully-labeled dataset. For generalizability evaluation, we used 4 scans (572 slices) randomly selected from the MSD dataset. **Inputs:** All image slices were 0–1 normalized and reshaped to $256 \times 256 \times 1$ before passing them to the models. We further preprocessed the images by window-leveling with a width of 400 HU and a center of 40 HU. Since the liver organ spans across a relatively smaller number of slices in a scan, we extracted only the middle 30% slices to avoid class imbalance. **Model Architecture:** Our PolyCL leverages the encoder-decoder architecture following [14]. We use Leaky ReLU (slope=0.2) and instance normalization between each convolution. We add a global average pool (GAP) at the end of f and a single fully-connected layer as g to generate 256-d embeddings. **Baselines:** For a baseline comparison, we used a fully-supervised version of U-Net with random initialization. We further compared our method by training and evaluating SimCLR [5]. All these networks are constructed with the same CNN backbone. **Training:** Each model, including each pre-trained encoder, was trained for 100 epochs with cosine-annealing learning rate scheduler and warm restarts after every 10 epochs. **Hyperparameters:** For pre-training, we used a learning rate of 0.0001 and a batch size of 20, while a learn rate of 0.001 and batch size of 10 was used for fine-tuning. Each model was trained 10 times to avoid any bias and enhance reliability in model predictions. **Evaluation:** For evaluation, we used Dice coefficient as a measure of similarity and Hausdorff distance as a distance metric. The average scores are reported across 10 iterations of each model.

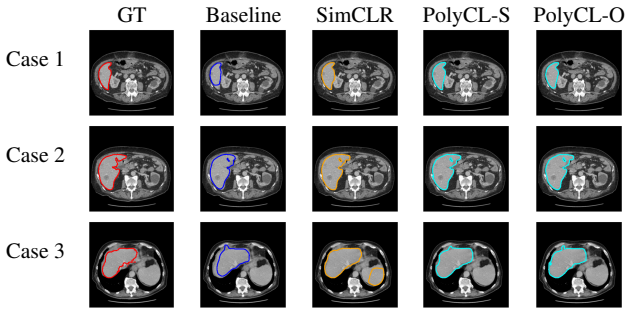

Fig. 3. Qualitative comparison of the segmented liver by our proposed PolyCL against baseline models and SimCLR on unseen CT slices. Please note that the rough boundary (Baseline–Case 2, SimCLR–Case 1), incomplete contour (Baseline–Cases 1 & 3), and incorrect delineation (SimCLR–Case 3), all are removed by our PolyCL-S and PolyCL-O.

3.2. Results and Discussion

Labeled data fine-tuning: Our primary findings comparing our proposed model to baseline U-Net and SimCLR when segmenting the liver from the LiTS dataset are reported in Table 1. We found that PolyCL-O performed significantly better than baseline U-Net and SimCLR ($p < 0.01$) in Dice coefficient. We also found that PolyCL-S showed improvements in Dice coefficient over the baseline model and a significant improvement in Hausdorff distance ($p < 0.01$) while being comparable to SimCLR. PolyCL-O consistently outperforms baseline models at scan level performance in 4/5 cases, with a marginal difference on one. This further validates our organ-based example selection as it indicates PolyCL-O would perform the best at the 3D scan level. The visualization of segmented liver boundaries in Fig. 3 further affirms the superiority of PolyCL-S and PolyCL-O over other methods.

Reduced-label fine-tuning: For reduced-label fine-tuning, we created fine-tuning sets using fewer CT scans from our LiTS train set, ranging from 1 to 9. As demonstrated in Fig. 4, using 5 train scans (45% slices), our PolyCL-O achieves comparable performance in Dice coefficient when compared to the baseline model trained on the entire dataset (0.8406 vs 0.8601, $p > 0.1$). Improved Hausdorff distance was obtained for 6 scans, or 56% of the train set. Similarly, we found that when using the PolyCL-S, improved Dice coefficient with 8 scans (81% slices) and improved Hausdorff distance was attained with 6 scans (56% slices).

Model generalization: To investigate the potential generalizability of the models, we fine-tuned and evaluated all of the LiTS pre-trained models on the TotalSegmentator dataset, reported in 1. We found that both PolyCL-O and PolyCL-S performed about equivalently, outperforming the baseline model by large margins. Additionally, PolyCL performs comparably to SimCLR in terms of Dice coefficient ($p = 0.5$); but significantly improved over SimCLR in terms of Hausdorff distance ($p < 0.05$). For additional generalization testing, we

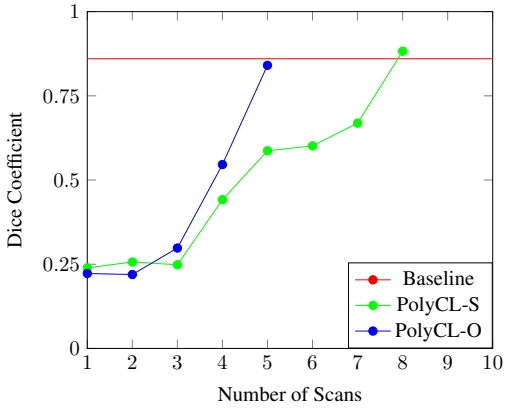

Fig. 4. Effectiveness of our PolyCL in limited labeled data fine-tuning. A set of 5 models were fine-tuned with each listed number of scans using the PolyCL-O and PolyCL-S strategies with encoder initializations randomly selected from their pre-trained encoders. To establish statistical significance, another 5 models were fine-tuned using 5 and 6 scans with PolyCL-O and 6 and 8 scans using PolyCL-S.

used the MSD dataset to evaluate the models already trained on the LiTS dataset without additional fine-tuning. We found that all contrastive pre-training approaches resulted in similar performance improvements over the baseline models, with PolyCL-O performing the best, achieving a Dice improvement of 1.4% over the next-best performing model.

4. CONCLUSIONS

We have presented a novel contrastive learning-based self-supervised learning method (PolyCL) employing innovative organ-based and scan-based example selection strategies for medical image segmentation. We have found that our pre-training strategy generates significant benefits and can use existing data more efficiently than fully supervised and other contrastive learning-based methods. We have also discovered that PolyCL is less data-hungry than fully supervised training, as comparable performance was achieved with a 55% reduction during fine-tuning. Additionally, PolyCL demonstrated improved generalization when fine-tuned and tested on out-of-distribution data. Our future work will focus on further exploring the architectural adjustment with different backbones as well as on larger-scale datasets across various segmentation tasks.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access. Ethical approval was not required as confirmed by the license attached with the open-access data.

6. REFERENCES

[1] Jialin Peng and Ye Wang, "Medical image segmentation with limited supervision: a review of deep network models," *IEEE Access*, vol. 9, pp. 36827–36851, 2021.

[2] Krishna Chaitanya et al., "Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation," *Medical Image Analysis*, vol. 87, pp. 102792, 2023.

[3] Abdullah-Al-Zubaer Imran, *From fully-supervised, single-task to scarcely-supervised, multi-task deep learning for medical image analysis*, Ph.D. thesis, UCLA, 2020.

[4] Saeed Shurrab and Rehab Duwairi, "Self-supervised learning methods and applications in medical imaging analysis: A survey," *PeerJ Computer Science*, vol. 8, pp. e1045, 2022.

[5] Ting Chen et al., "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.

[6] Ke Yan et al., "Sam: Self-supervised learning of pixel-wise anatomical embeddings in radiological images," *IEEE TMI*, vol. 41, no. 10, pp. 2658–2669, 2022.

[7] Jinxi Xiang et al., "Self-ensembling contrastive learning for semi-supervised medical image segmentation," *arXiv preprint arXiv:2105.12924*, 2021.

[8] Yonglong Tian, Dilip Krishnan, and Phillip Isola, "Contrastive multiview coding," in *ECCV*, 2020, pp. 776–794.

[9] Shekoofeh Azizi et al., "Big self-supervised models advance medical image classification," in *ICCV*, 2021, pp. 3478–3488.

[10] Olaf Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[11] Patrick Bilic et al., "The liver tumor segmentation benchmark (LiTS)," *Medical Image Analysis*, vol. 84, pp. 102680, 2023.

[12] Jakob Wasserthal et al., "Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images," *Radiology: AI*, vol. 5, no. 5, 2023.

[13] Michela Antonelli et al., "The medical segmentation decathlon," *Nature Communications*, vol. 13, no. 1, 2022.

[14] Ayaan Haque et al., "Multimix: sparingly-supervised, extreme multitask learning from medical images," in *ISBI*, 2021, pp. 693–696.