PiMPiC: An Overlap-Aware Contrastive Learning Framework for 3D Patch-Based Medical Image Segmentation

Reza Karimzadeh[1], Ehsan Yousefzadeh-Asl-Miandoab[2], Hossein Arabi[3], Pinar Tözün[2], and Bulat Ibragimov[1(✉)]

[1] Copenhagen University, Copenhagen, Denmark
[2]IT University of Copenhagen, Copenhagen, Denmark
[3] Geneva University Hospital, Geneva, Switzerland
bulat@di.ku.dk

Abstract. Deep learning models for 3D medical image segmentation typically require large annotated datasets to achieve high accuracy. However, collecting such datasets is time-consuming, costly, and constrained by privacy regulations. Contrastive learning, a self-supervised technique, enables models to learn meaningful data representations without any labeled data. However, applying traditional contrastive learning methods to medical images is challenging due to the structural similarity of human tissues, which often results in false negatives when similar tissues are treated as dissimilar. Additionally, slice-wise contrastive learning approaches rely on relative slice positions to form positive and negative pairs, limiting generalization to 3D patches and requiring image preregistration. To address these issues, we propose two novel modules for contrastive learning-based pretraining of 3D segmentation. The first, Patch Intersection Measurement (PiM), estimates the overlap between two patches in the embedding space. The second, Patch Intersection Contrast (PiC), encourages embeddings of overlapping regions to align closely while pushing non-overlapping regions apart. Experiments on two datasets for pancreas and kidney cancers segmentation demonstrated that our method outperforms both the state-of-the-art (SOTA) and the baseline segmentation models. Notably, for pancreas segmentation, even when trained with only 5% of the labeled data, our method achieves 12% and 4% improvement in Dice score compared to the baseline and SOTA, highlighting its effectiveness in low-data scenarios. The code is available at https://github.com/rezakarimzadeh/pimpic.

Keywords: Medical Image Segmentation · Self-supervised Learning · Contrastive Learning.

1 Introduction

Deep learning has demonstrated remarkable accuracy in medical image segmentation [10]. However, the performance of deep neural network models heavily depends on large volumes of annotated data, making them inherently data-hungry

[19]. Acquiring such labeled datasets is costly and time-consuming, requires extensive manual effort from clinical experts and often raises privacy concerns [28]. Self-supervised learning (SSL) has emerged as a favourable approach, capable of learning data representations without the need for labeled data [29]. SSL provides an approach of choice to the scarcity of annotated datasets, especially for tasks involving rare diseases, unique anatomies, or limited patient data [2].

SSL methods can be categorized into pretext task-based, generative model-based, and discriminative-based contrastive learning [20]. Pretext task-based methods design artificial tasks such as relative positioning [7], rotation prediction [11], and jigsaw puzzles [21] to learn meaningful features from unlabeled data. Generative model-based approaches, including autoencoders, context encoders [22], and bidirectional generative adversarial networks (BiGAN) [8], focus on reconstructing or generating data to capture its underlying structure. Discriminative-based contrastive learning methods, such as SimCLR [5], MoCo [14], and BYOL [13], learn by contrasting positive/similar and negative/dissimilar sample pairs to develop robust representations.

Most SSL methods, designed for natural images, struggle with 3D medical images due to challenges in meaningful augmentations, local feature learning, and high intra-class similarity [4]. Contrastive learning often generates false negatives from similar surrounding tissues, and existing SSL methods lack domain-specific adaptations for 3D segmentation [26,27]. To address these challenges, Zeng et al. [31] proposed a contrastive pretraining approach based on the relative positioning of slices in 3D images, while Chaitanya et al. [4] introduced a framework combining global volumetric and local per-pixel contrastive losses. Despite improving segmentation with limited labels, these methods remain constrained by their 2D nature, ignoring spatial continuity along the third axis [16,32].

Studies show that 3D segmentation models outperform 2D segmentation by leveraging spatial and contextual information more effectively while pre-training further enhances segmentation performance [25,3]. Goncharov et al. [12] introduced voxel-wise contrastive learning but focused only on local features, missing broader contextual relationships. Wu et al. [30] proposed VoCo, which predicts patch overlaps using a fixed 4×4 grid. While effective, its rigid structure may lead to overfitting and limited adaptability to anatomical variations. Additionally, increasing base patches beyond $4 \times 4 \times 1$ provides minimal gains, highlighting the need for more flexible selection strategies to better capture spatial variability.

In this study, we propose the Patch Intersection Measure and Contrast (PiMPiC) method for SSL, which analyzes image patch intersections to measure overlaps and apply contrastive learning to distinguish between overlapping and non-overlapping embeddings in feature maps. To the best of our knowledge, this is the first self-supervision strategy that capitalizes on both patch overlap prediction and contrastive learning performed on overlapping and non-overlapping regions. By enforcing this scheme, we encourage the model to capture the semantic representation of each patch, allowing it to learn positional contextual information without explicitly encoding spatial coordinates while having high randomness in the path selection, preventing shortcut learning and considering

contextual information instead of voxel-level information. We also visualize the learned representations, providing qualitative insights into the effectiveness of our self-supervised learning approach in structuring the embedding space.

2 Method

Our method consists of two key modules: PiM and PiC, both analyzing the intersection of random 3D patches (Fig. 1). PiM predicts the overlap between patches using projected embeddings, assigning zero for non-overlapping pairs and a percentage for overlapping ones. PiC applies a contrastive loss, enforcing similarity in the feature embeddings of intersecting regions while distinguishing non-intersecting ones. We employ a 3D U-Net encoder [6] as the backbone for self-supervised training. After pretraining, this encoder is fine-tuned with a randomly initialized decoder for downstream segmentation tasks. Further details on each module are provided in the following sections.

2.1 Patch intersection Measurement (PiM)

In the encoder, we employ a feature pyramid that starts with a base number of features (K) in the first layer, which has the same spatial dimensions as the patch size (W, H, D). After each downsampling by a factor of 2, the number of extracted features is doubled. Therefore, the input to the encoder has the dimensions $X \in \mathbb{R}^{N\times 1\times W\times H\times D}$, where N represents the batch size. The extracted feature maps at each level of the feature pyramid are represented as $F_l \in \mathbb{R}^{N\times(K\times 2^l)\times\frac{W}{2^l}\times\frac{H}{2^l}\times\frac{D}{2^l}}$, where l denotes the level of the feature pyramid.

This module utilizes the extracted features from the feature pyramid to measure the intersection between every two patches in the batch. To achieve this, we use the last three levels of the extracted features to predict the overlap. First, we perform channel-wise global average pooling on the features, resulting in feature maps of size $PF_l \in \mathbb{R}^{N\times\frac{W}{2^l}\times\frac{H}{2^l}\times\frac{D}{2^l}}$. Next, we concatenate the flattened versions of the features from the last three levels, specifically PF_{L-2}, PF_{L-1}, and PF_L. Finally, the concatenated features are passed through a projection head denoted as $f(\cdot)$ to obtain the projected features P of the batch (Fig. 1), defined as $P = f([PF_{L-2}, PF_{L-1}, PF_L])$, where $P \in \mathbb{R}^{N\times d}$ and d represents the dimension of the projected features. The optimization loss is the following:

$$\mathcal{L}_{PiM} = -\frac{2}{N(N-1)}\sum_{i=0}^{N-1}\sum_{j=0}^{i-1} log(1 - D_{i,j}) \tag{1}$$

$$D = |S - C|, \qquad C_{i,j} = Sim(P_i, P_j) = \frac{P_i.P_j}{||P_i||.||P_j||}, \quad i, j \in N,$$

where S represents the intersection matrix between all pairs of patches in the batch, which is known based on the positions of the patches in the original

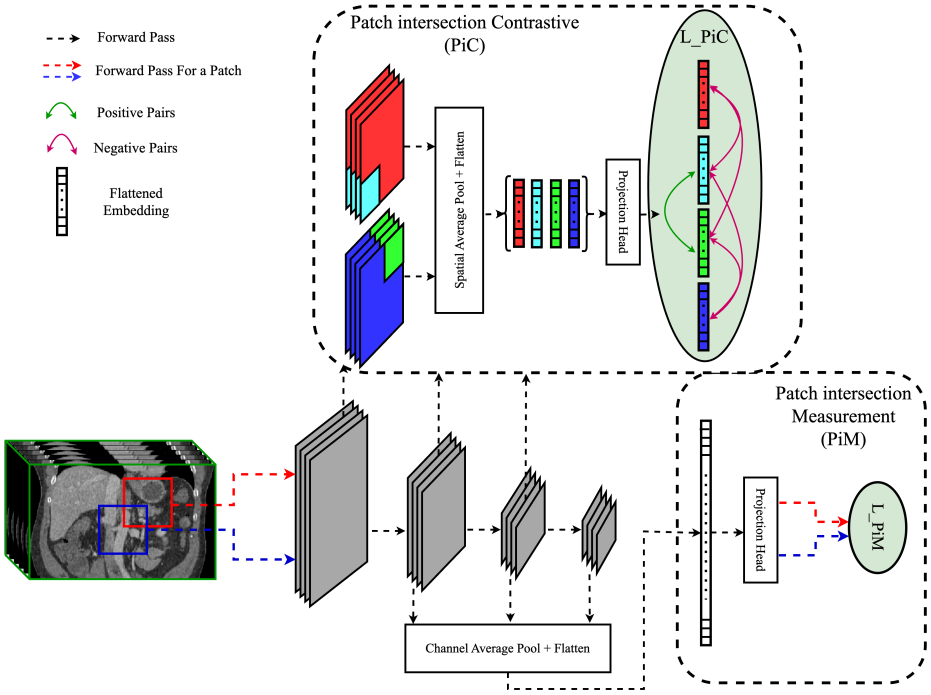

Fig. 1. Overall method description. The framework comprises two primary modules: Patch Intersection Measurement (PiM) and Patch Intersection Contrastive Module (PiC). PiM predicts the overlap between two random patches by analyzing their projected embeddings. PiC employs a contrastive loss to align embeddings of intersecting regions while distinguishing non-intersecting regions. A 3D U-Net encoder serves as the backbone for feature extraction.

3D image. $C_{i,j}$ denotes the cosine similarity between the projected features of patches i and j after passing through the projection head. This loss function minimizes the predicted intersection error between patch pairs.

2.2 Patch intersection Contrast (PiC)

Similar to PiM, PiC utilizes the extracted features from the feature pyramid, but exploits the first three feature maps. The objective is to use a contrastive setup to ensure that the embeddings of overlapping regions between two patches are close, while the embeddings of non-overlapping regions are distant.

As shown in Fig. 1, PiC considers the features for two overlapping patches represented as $F_l \in \mathbb{R}^{1\times(K\times 2^l)\times\frac{W}{2^l}\times\frac{H}{2^l}\times\frac{D}{2^l}}$. Based on the original image and the corresponding feature pyramid level, we can identify the overlapping coordinates. By applying spatial average pooling to both the overlapping and non-overlapping regions, we obtain four feature vectors corresponding to the two overlapping

patches' embeddings by the shape $\mathbb{R}^{4\times(K\times 2^l)}$. Finally, by passing the concatenated flattened features through a projection head denoted as $g(\cdot)$, we obtain a matrix $Q \in \mathbb{R}^{N'\times 4\times d}$, where N' represents the number of overlapping patch pairs in the batch, and d is the dimension of the projected features. The final loss function for this module is defined as follows:

$$\mathcal{L}_{PiC} = -\frac{1}{N'}\sum_{b=1}^{N'} log\frac{e^{Sim(Q_{b,0},Q_{b,1})/\tau}}{\sum_{i=0}^{3}\sum_{j=0}^{i-1} e^{Sim(Q_{b,i},Q_{b,j})/\tau}} \tag{2}$$

In which, Sim is the cosine similarity, and τ is the temperature factor that controls the concentration of the distribution in the contrastive loss. Terms $Q_{b,0}$ and $Q_{b,1}$ represent the positive (overlapping) pairs of feature vectors for the b-th overlapping patch pair in the batch. This loss function maximizes the similarity between overlapping regions while minimizing it for non-overlapping regions.

In PiM, we utilize channel-wise global average pooling, while in PiC, we apply spatial-wise average pooling. The rationale behind this design is that, for calculating the overlap in PiM, it is more meaningful to preserve spatial information, as it directly relates to the positional alignment of patches. Conversely, in PiC, when contrasting overlapping and non-overlapping regions, leveraging channel information is more beneficial, as it captures semantic feature differences that help distinguish between the two regions.

It is worth noting that convolutional neural networks (CNNs) have a degree of equivariant to translation [18], meaning that in PiC, overlapping regions can naturally produce similar feature representations. However, our loss function enhances discriminability by pushing negative pairs away from the embeddings of overlapping sections. Introducing intensity-based data augmentations additionally increases the robustness of representations to intensity changes.

3 Experimental Setup

In this work, we use two datasets: the pancreas segmentation dataset from the Medical Segmentation Decathlon (MSD) [1] and the Kidney Tumor Segmentation (KiTS) dataset [15], containing 281 and 210 annotated cases, respectively. Both datasets are split into 70% training, 10% validation, and 20% test sets. Preprocessing involves: (1) clipping CT values (Hounsfield Units) to $[-200, 400]$ HU and mapping them to $[0, 1]$, and (2) interpolating and normalizing voxel spacings to $[1.5, 1.5, 1.5]$ mm^3.

In the pretraining stage, we trained the 3D U-Net encoder using different methods to evaluate the impact of each proposed module: PiM, PiC, and their combination (PiMPiC). Pretraining was conducted on raw training images, excluding segmentation masks. Models were trained for 20K steps with a batch size of 32, using the Adam optimizer [17] with a learning rate of 10^{-5} and a patch size of $[64, 64, 64]$. The encoder comprises four downsampling levels, where each layer includes three 3D convolutional layers followed by instance normalization

and ReLU activation, with skip connections to aid gradient flow. The base feature size (K) is set to 32, and the projection head output dimension for PiM and PiC is 128. Additionally, τ is set to 1 in Eq. 2.

After pretraining the encoder, we appended a decoder and trained the full model using segmentation masks. All encoder-decoder models share the same architecture, differing only in the pre-trained encoder weights obtained from various self-supervised methods. To assess performance under different data availability conditions, we trained segmentation models using 5%, 10%, 20%, 50%, and 100% of the training data. Training was conducted with the Dice loss function, and early stopping with a 15-epoch patience.

4 Results

The results in terms of Dice score are presented in Tables 1 and 2 for pancreatic and kidney cancer segmentation, respectively. As a baseline, we trained a model with randomly initialized weights and applied data augmentation techniques, including rotation, shift, scaling, and random noise. Additionally, we compared our method to the state-of-the-art pertaining algorithms proposed for medical image segmentation: VoCo [30] and Vox2Vec [12].

Table 1. Transfer learning comparison of segmentation performance across varying proportions of labeled training data. This table presents the mean Dice scores (± SD) for pancreas and pancreas tumor segmentation tasks, evaluated under different percentages of labeled training data . The performance of our proposed method, PiMPiC, is compared against a baseline model, VoCo, and Vox2Vec. An ablation study is also included to assess the individual contributions of the PiM and PiC modules.

Method/Training %	**5% (10)**	**10% (20)**	**20% (40)**	**50% (98)**	**100% (196)**
Pancreas					
Random + Aug	36.0 ± 16.9	41.7 ± 15.7	55.3 ± 15.4	63.5 ± 12.4	66.9 ± 11.0
VoCo [30]	37.2 ± 16.9	44.4 ± 18.8	55.7 ± 15.2	61.9 ± 14.0	65.9 ± 11.5
Vox2Vec [12]	44.7 ± 14.6	49.9 ± 13.8	58.0 ± 13.8	63.5 ± 11.7	65.1 ± 11.3
PiM	47.3 ± 14.7	51.6 ± 15.0	**60.5 ± 13.4**	61.3 ± 14.3	67.6 ± 11.2
PiC	38.9 ± 15.8	47.6 ± 17.7	56.3 ± 17.0	61.3 ± 15.7	65.0 ± 13.0
PiMPiC	**48.3 ± 15.3**	**52.2 ± 15.7**	59.2 ± 13.5	**65.4 ± 12.2**	**68.2 ± 10.5**
Tumor					
Random + Aug	2.2 ± 8.4	7.6 ± 14.5	14.1 ± 18.6	21.6 ± 21.5	25.9 ± 23.0
VoCo [30]	10.3 ± 14.4	12.3 ± 15.4	12.7 ± 16.0	22.1 ± 21.1	27.4 ± 23.2
Vox2Vec [12]	12.6 ± 15.6	16.5 ± 18.4	**18.4 ± 20.5**	22.7 ± 21.8	27.8 ± 24.1
PiM	9.6 ± 14.3	14.0 ± 17.0	15.0 ± 17.2	21.9 ± 21.5	23.3 ± 22.7
PiC	8.6 ± 13.3	15.3 ± 16.8	13.7 ± 16.4	**23.2 ± 21.3**	26.8 ± 22.5
PiMPiC	**13.9 ± 15.0**	**17.2 ± 19.0**	13.9 ± 17.6	23.1 ± 21.0	**27.8 ± 23.5**

5 Discussion

We proposed PiMPiC, a straightforward and powerful SSL method for pretraining 3D segmentation networks. It studies the overlap between patches (PiM)

while leveraging contrastive learning (PiC) to pull the embeddings of intersecting regions closer and push the embeddings of non-intersecting regions further apart within the encoder feature pyramid. Unlike existing 3D SSL methods, PiMPiC does not rely on positional information but learns patch intersections and their spatial relationships directly. This eliminates the need to define fixed bases for calculating patch intersections, allowing the model greater flexibility to learn inter-patient relationships between patches without predefined constraints.

In the low-data regime, PiMPiC achieved notable improvements over the baseline and SOTA methods. As shown in Table 1, for pancreas tissue segmentation with only 5% labeled data, the baseline, VoCo, and Vox2Vec obtained Dice scores of 36.0%, 37.2%, and 44.7%, respectively, while PiMPiC outperformed them with a 48.3% Dice score. A similar trend is observed in pancreas tumor segmentation, as well as kidney and tumor segmentation (Tables 1 and 2), demonstrating the effectiveness of our method. Fig. 2 provides an example of pancreas and tumor segmentation using 10% of the training segmentation data.

To conduct an ablation study, we separately evaluated the contributions of the PiM and PiC modules. With several exceptions, the combined use of PiM and PiC results in up to a 30% improvement in segmentation accuracy (Table 1). Analysis of the validation loss curves confirms this observation: PiMPiC converges to a lower loss value than PiM, PiC, and alternative methods, indicating that PiMPiC enables the network to learn representative embeddings for image structures using PiM, while PiC discriminates between similar and dissimilar regions, maximizing the separation of embeddings. This synergy enhances the network's performance in downstream tasks (Figure 3 (E)).

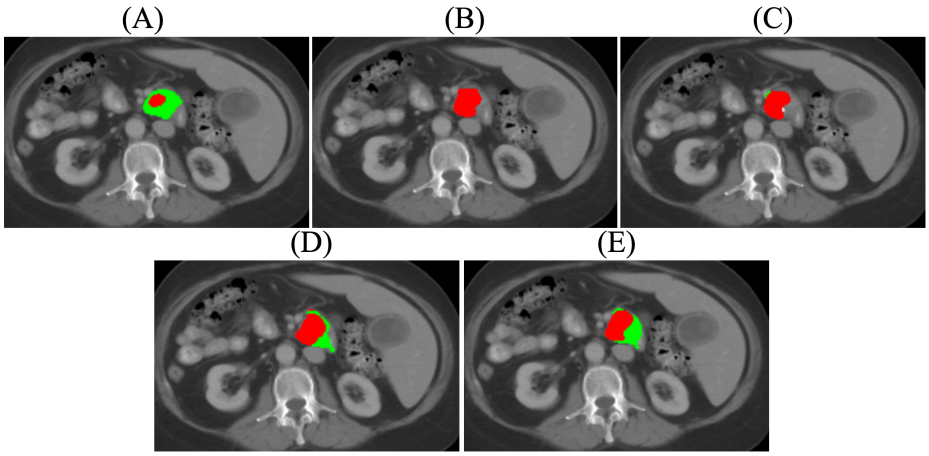

Fig. 2. Pancreas and tumor segmentation with output with 10% Training Data. Axial views from a representative patient show: (A) ground truth segmentation; (B) result from a model with random initialization; (C) VoCo; (D) Vox2Vec; and (E) our proposed PiMPiC method. PiMPiC effectively captures both pancreas and tumor tissues, leading to improved segmentation performance.

Table 2. Mean Dice scores (± SD) for kidney and tumor segmentation. This table compares our PiMPiC method against the baseline, VoCo, and Vox2Vec models across various labeled data percentages.

Method/Training %	5% (8)	10% (15)	20% (30)	50% (74)	100% (147)
Kidney					
Random + Aug	49.1 ± 13.8	82.5 ± 15.3	83.7 ± 14.7	87.8 ± 14.2	89.1 ± 11.3
VoCo [30]	68.2 ± 15.3	84.0 ± 15.3	82.3 ± 18.4	87.6 ± 14.6	**89.1 ± 14.8**
Vox2Vec [12]	65.5 ± 16.8	83.4 ± 12.1	81.4 ± 18.0	87.3 ± 14.7	87.6 ± 14.3
PimPiC	**77.6 ± 18.0**	**84.2 ± 16.9**	**85.3 ± 14.2**	**88.6 ± 13.4**	89.1 ± 14.3
Tumor					
Random + Aug	5.5 ± 11.7	11.3 ± 17.3	21.0 ± 21.5	34.7 ± 27.1	41.2 ± 29.6
VoCo [30]	7.6 ± 13.4	12.1 ± 16.1	16.5 ± 21.0	29.9 ± 25.4	**46.4 ± 28.5**
Vox2Vec [12]	9.3 ± 11.5	16.2 ± 17.5	20.0 ± 18.6	30.3 ± 23.4	38.4 ± 25.6
PimPiC	**10.4 ± 14.0**	**14.6 ± 18.4**	**23.9 ± 24.1**	**35.3 ± 27.5**	42.1 ± 29.0

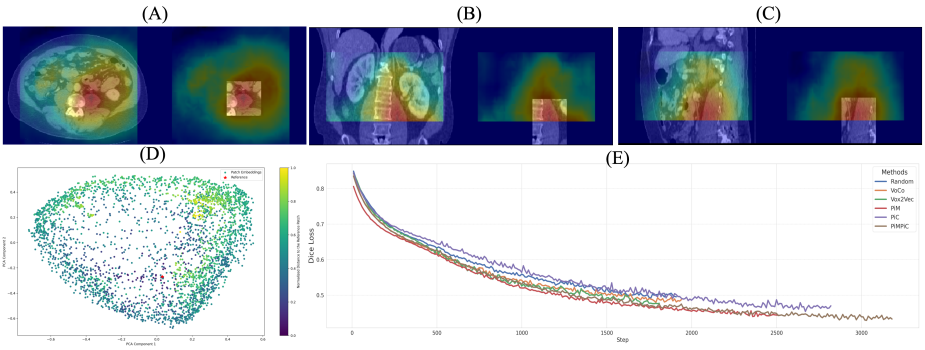

Fig. 3. (A-C) PiMPiC embedding similarities for a CT image visualized as heatmaps indicating cosine similarity between a reference patch (right sub-figure) and other patches. Higher similarity near the reference patch suggests effective capture of local contextual information. (D) A 2D PCA projection shows embeddings colored by normalized center-to-center distances from the reference patch, confirming that closer patches have more similar embeddings. (E) Validation loss curves during training.

The purpose of the PiM module is to enforce spatial coherence in the embedding space, ensuring that similar anatomical regions exhibit higher feature similarity while maintaining distinctiveness across different structures. This should improve segmentation by preserving local contextual information. To validate this, we selected a reference patch from a random CT image, extracted its embedding using PiM's projection head, and computed cosine similarity with all other patches in the image. As expected, the resulting similarity measurements formed a distance map-like image (Fig. 3 A-C), highlighting strong correlations near the reference patch. A 2D PCA projection (Fig. 3 D) shows that closer

patches have more similar embeddings, while distant patches are mapped further apart, reinforcing the model's ability to distinguish local structures.

In our experiments, we pre-trained the models using SSL techniques with 70% of the available data for each dataset. We then fine-tuned the models for the segmentation task using varying percentages of labeled data. This setup allowed us to evaluate SSL methods under limited pretraining data conditions. Unlike existing SSL approaches, our framework demonstrated strong performance even with a reduced amount of pretraining data, highlighting its efficiency in leveraging limited annotations while maintaining high segmentation accuracy.

We also evaluated the energy consumption of our proposed method and observed that during pre-training, it consumed 24.50 MJ (6.80 kWh). As shown in Table 1, by fine-tuning the model with only 5% of annotated data, our approach required approximately 20 fewer annotated samples to achieve the same level of accuracy as a model with random initialization. Considering that expert annotation typically requires 30–60 minutes per case, our SSL method is, on average, 30 times more cost-efficient in the case of limited data availability [23,9,24]. Thus, beyond leveraging unlabeled data, our approach also reduces the financial cost of model training.

Acknowledgments. This work received support from Novo Nordisk Foundation under grant NFF20OC0062056.

References

1. 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)
2. Arabi, H., Zaidi, H.: Mri-guided attenuation correction in torso pet/mri: Assessment of segmentation-, atlas-, and deep learning-based approaches in the presence of outliers. Magnetic resonance in medicine **87**(2), 686–701 (2022)
3. Avesta, A., Hossain, S., Lin, M., Aboian, M., Krumholz, H.M., Aneja, S.: Comparing 3d, 2.5 d, and 2d approaches to brain image auto-segmentation. Bioengineering **10**(2), 181 (2023)
4. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. Advances in neural information processing systems **33**, 12546–12558 (2020)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
7. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE international conference on computer vision. pp. 1422–1430 (2015)

8. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. arXiv preprint arXiv:1605.09782 (2016)
9. ERIERI: Radiologist diagnostic salary in denmark (2025), https://www.erieri.com/salary/job/radiologist-diagnostic/denmark, accessed: 2025-02-21
10. Fernando, K.R.M., Tsokos, C.P.: Deep and statistical learning in biomedical imaging: State of the art in 3d mri brain tumor segmentation. Information Fusion **92**, 450–465 (2023)
11. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (2018)
12. Goncharov, M., Soboleva, V., Kurmukov, A., Pisov, M., Belyaev, M.: vox2vec: A framework for self-supervised contrastive learning of voxel-level representations in medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 605–614. Springer (2023)
13. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
15. Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 (2019)
16. Hu, X., Zeng, D., Xu, X., Shi, Y.: Semi-supervised contrastive learning for label-efficient medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 481–490 (2021)
17. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Kondor, R., Trivedi, S.: On the generalization of equivariance and convolution in neural networks to the action of compact groups. In: International conference on machine learning. pp. 2747–2755. PMLR (2018)
19. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
20. Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., Tang, J.: Self-supervised learning: Generative or contrastive. IEEE transactions on knowledge and data engineering **35**(1), 857–876 (2021)
21. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. pp. 69–84. Springer (2016)
22. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016)
23. Qu, C., Zhang, T., Qiao, H., Tang, Y., Yuille, A.L., Zhou, Z., et al.: Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. Advances in Neural Information Processing Systems **36**, 36620–36636 (2023)
24. Services, A.W.: Amazon ec2 on-demand pricing (2025), https://aws.amazon.com/ec2/pricing/on-demand, accessed: 2025-02-21

25. Shivdeo, A., Lokwani, R., Kulkarni, V., Kharat, A., Pant, A.: Evaluation of 3d and 2d deep learning techniques for semantic segmentation in ct scans. In: 2021 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD). pp. 1–8. IEEE (2021)
26. Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C.: 3d self-supervised methods for medical imaging. Advances in neural information processing systems **33**, 18158–18172 (2020)
27. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 776–794. Springer (2020)
28. Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., Yang, R., et al.: Annotation-efficient deep learning for automatic medical image segmentation. Nature communications **12**(1), 5915 (2021)
29. Wang, W.C., Ahn, E., Feng, D., Kim, J.: A review of predictive and contrastive self-supervised learning for medical images. Machine Intelligence Research **20**(4), 483–513 (2023)
30. Wu, L., Zhuang, J., Chen, H.: Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22873–22882 (2024)
31. Zeng, D., Wu, Y., Hu, X., Xu, X., Yuan, H., Huang, M., Zhuang, J., Hu, J., Shi, Y.: Positional contrastive learning for volumetric medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 221–230 (2021)
32. Zhai, P., Cong, H., Zhu, E., Zhao, G., Yu, Y., Li, J.: Mvcnet: multiview contrastive network for unsupervised representation learning for 3-d ct lesions. IEEE Transactions on Neural Networks and Learning Systems (2022)