

Hierarchical Self-Supervised Learning for Medical Image Segmentation Based on Multi-Domain Data Aggregation

Hao Zheng, Jun Han, Hongxiao Wang, Lin Yang,
Zhuo Zhao, Chaoli Wang, and Danny Z. Chen

Department of Computer Science and Engineering, University of Notre Dame,
Notre Dame, IN 46556, USA
hzheng3@nd.edu

Abstract. A large labeled dataset is a key to the success of supervised deep learning, but for medical image segmentation, it is highly challenging to obtain sufficient annotated images for model training. In many scenarios, unannotated images are abundant and easy to acquire. Self-supervised learning (SSL) has shown great potentials in exploiting raw data information and representation learning. In this paper, we propose Hierarchical Self-Supervised Learning (HSSL), a new self-supervised framework that boosts medical image segmentation by making good use of unannotated data. Unlike the current literature on task-specific self-supervised pretraining followed by supervised fine-tuning, we utilize SSL to learn task-agnostic knowledge from heterogeneous data for various medical image segmentation tasks. Specifically, we first aggregate a dataset from several medical challenges, then pre-train the network in a self-supervised manner, and finally fine-tune on labeled data. We develop a new loss function by combining contrastive loss and classification loss, and pre-train an encoder-decoder architecture for segmentation tasks. Our extensive experiments show that multi-domain joint pre-training benefits downstream segmentation tasks and outperforms single-domain pre-training significantly. Compared to learning from scratch, our method yields better performance on various tasks (e.g., +0.69% to +18.60% in Dice with 5% of annotated data). With limited amounts of training data, our method can substantially bridge the performance gap with respect to denser annotations (e.g., 10% vs. 100% annotations).

Keywords: Self-supervised learning · Image segmentation · Multi-domain

1 Introduction

Although supervised deep learning has achieved great success on medical image segmentation [15, 17, 25, 34], it heavily relies on sufficient good-quality manual annotations which are usually hard to obtain due to expensive acquisition, data privacy, etc. Public medical image datasets are normally smaller than the generic image datasets (see Fig. 1(a)), and may hinder improving segmentation performance. Deficiency of annotated data has driven studies to explore alternative

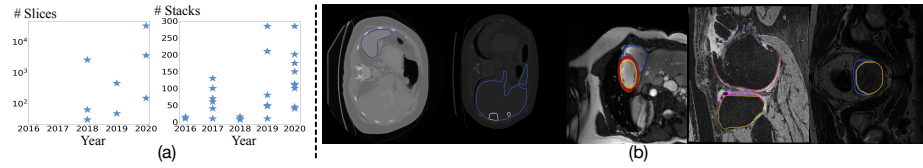


Fig. 1. (a) The number of images for each medical image segmentation challenge every year since 2016 at MICCAI (left: 2D images; right: 3D stacks). (b) Diverse medical image and mask examples: spleen, liver & tumours, cardiovascular structures, knee bones & cartilages, and prostate.

solutions. Transfer learning fine-tunes models pre-trained on ImageNet for target tasks [12, 36, 37], but it could be impractical and inefficient due to the pre-defined model architectures [18] and is not as good as transferred from medical images due to image characteristics differences [37]. Semi-supervised learning utilizes unlimited amounts of unlabeled data to boost performance, but it usually assumes that the labeled data sufficiently covers the data distribution, and needs to address consequent non-trivial challenges such as adversarial learning [20, 32] and noisy labels [31, 35]. Active learning selects the most representative samples for annotation [29, 33, 36] but focuses on saving manual effort and does not utilize unannotated data. Considering these limitations and the fact that considerable unlabeled medical images are easy to acquire and free to use, we seek to answer the question: *Can we improve segmentation performance with limited training data by directly exploiting raw data information and representation learning?*

Recently, self-supervised learning (SSL) approaches, which initialize models by constructing and training surrogate tasks with unlabeled data, attracted much attention due to soaring performance on representation learning [8–10, 14, 16, 21, 22, 24] and downstream tasks [4, 5, 23, 27, 37, 38]. It was shown that the learned representation by *contrastive learning*, a variant of SSL, gradually approaches the effectiveness of representations learned through strong supervision, even under circumstances when only limited data or a small-scale dataset is available [6, 11]. However, three key factors of contrastive learning have not been well explored for medical segmentation tasks: (1) A medical image dataset is often insufficiently large due to the intrusive nature of some imaging techniques or expensive annotations (e.g., 3D(+T) images), which suppresses self-supervised pre-training and hinders representation learning using a single dataset. (2) The contrastive strategy considers only congenetic image pairs generated by different transformations used in data augmentation, which suppresses the model from learning task-agnostic representations from heterogeneous data collected from different sources (see Fig. 1(b)). (3) Most studies focused on extracting high-level representations by pre-training the encoder while neglecting to learn low-level features explicitly and initialize the decoder, which hinders the performance of dense prediction tasks such as semantic segmentation.

To address these challenges, in this paper, we propose a new *hierarchical self-supervised learning* (HSSL) framework to pre-train on heterogeneous unan-

notated data and obtain an initialization beneficial for training multiple downstream medical image segmentation tasks with limited annotations. First, we investigate available public challenge datasets on medical image segmentation and propose to aggregate a multi-domain (modalities, organs, or facilities) dataset. In this way, our collected dataset is considerably larger than a task-specific dataset and the pretext model is forced to learn task-agnostic knowledge (e.g., texture, intensity distribution, etc). Second, we construct pretext tasks at multiple abstraction levels to learn hierarchical features and explicitly force the model to learn richer semantic features for segmentation tasks on medical images. Specifically, our HSSL utilizes contrasting and classification strategies to supervise image-, task-, and group-level pretext tasks. We also extract multi-level features from the network encoding path to bridge the gap between low-level texture and high-level semantic representations. Third, we attach a lightweight decoder to the encoder and pre-train the encoder-decoder architecture to obtain a suitable initialization for downstream segmentation tasks.

We experiment on our aggregated dataset composed of eight medical image segmentation tasks and show that our HSSL is effective in utilizing multi-domain data to initialize model parameters for target tasks and achieves considerably better segmentation, especially when only limited annotations are available.

2 Methodology

We discuss the necessity and feasibility of aggregating multi-domain image data and show how to construct such a dataset in Sect. 2.1, and then introduce our hierarchical self-supervised learning pretext tasks (shown in Fig. 2) in Sect. 2.2. After pre-training, we fine-tune the trained encoder-decoder network on downstream segmentation tasks with limited annotations.

2.1 Multi-Domain Data Aggregation

Necessity. As shown in Fig. 1(a), most publicly available medical image segmentation datasets are of relatively small sizes. Yet, recent progresses on contrastive learning empirically showed that training on a larger dataset often learns better representations and brings larger performance improvement in downstream tasks [6, 7, 11]. Similarly, a larger dataset is beneficial for supervised classification tasks and unsupervised image reconstruction tasks, because such a dataset tends to be more diverse and better cover the true image space distribution.

Feasibility. First, there are quite a few medical image dataset archives (e.g., TCIA) and public challenges (e.g., Grand Challenge). Typical imaging modalities (CT, MRI, X-ray, etc) of multiple regions-of-interest (ROIs, organs, structures, etc) are covered. Second, common/similar textures or intensity distributions are shared among different datasets (see Fig. 1(b)), and their raw images may cover the same physical regions (e.g., abdominal CT for the spleen dataset and liver dataset). Therefore, an aggregated multi-domain dataset can (1) enlarge the

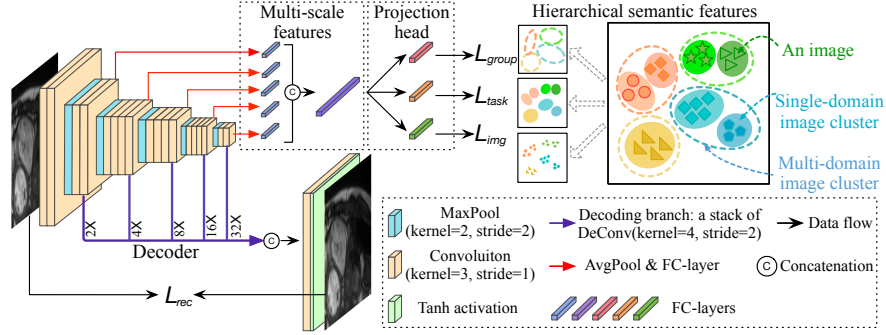


Fig. 2. An overview of our proposed hierarchical self-supervised learning (HSSL) framework (best viewed in color). The backbone encoder builds a pyramid of multi-scale features from the input image, forming a rich latent vector. Then it is stratified to represent hierarchical semantic features of the aggregated multi-domain data, supervised by different pretext tasks in the hierarchy. Besides, an auxiliary reconstruction pretext task helps initialize the decoder.

data size of a shared image space and (2) force the model to distinguish different contents from the raw images. In this way, task-agnostic knowledge is extracted.

Dataset Aggregation. To ensure the effectiveness of multi-domain data aggregation, three principles should be considered. (1) Representativeness: The datasets considered for aggregation should cover a moderate range of medical imaging techniques/modalities. (2) Relevance: The datasets considered should not drastically differ in content/appearance. Otherwise, it is easy for the model to distinguish them and a less common feature space is shared among them. (3) Diversity: The datasets considered should benefit a range of applications. In this work, we focus on CT and MRI of various ROIs (i.e., heart, liver, prostate, pancreas, knee, and spleen). The details of aggregated dataset are shown in Table 1.

2.2 Hierarchical Self-Supervised Learning (HSSL)

Having aggregated multiple datasets, $\mathcal{D} = \{D_1, D_2, \dots, D_N\}$, where D_i is a dataset for a certain segmentation task. A straightforward method to use \mathcal{D} is to directly extend some known pretext tasks (e.g., SimCLR [6]) and conduct joint pre-training. However, such pretext tasks only explicitly force the model to learn a global representation and are not tailored for the target segmentation tasks. Hence, taking imaging techniques and prior knowledge (e.g., appearance, ROIs) into account, we propose to extract richer semantic features from hierarchical abstract levels and devise the network for target segmentation tasks.

We formulate three hierarchical levels (see Fig. 3(a)). (1) *Image-level*: Each image I is a learning subject; we want to extract distinguishable features of I w.r.t. another image, regardless of which dataset it originally comes from or what ROIs it contains. Specifically, we follow the state-of-the-art SimCLR [6] and build positive and negative pairs with various data augmentations. (2) *Task-level*: Each

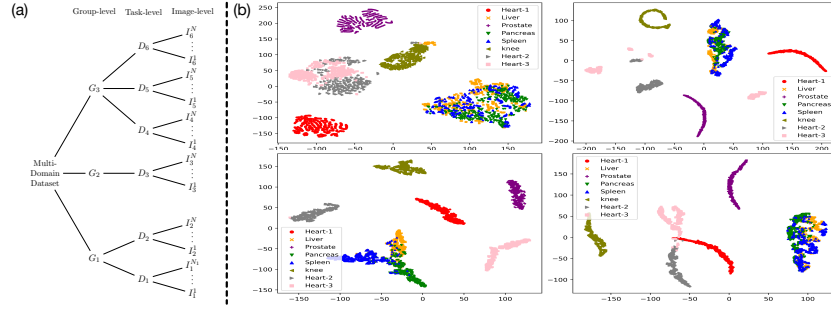


Fig. 3. (a) An example of the hierarchical structure of a multi-domain dataset. Each chosen dataset/task D_i forms a domain consisting of a set of images $\{I_i^k\}_{k=1}^{N_i}$, where N_i is the total number of images in D_i . Multiple tasks form a multi-domain cluster called a *group* (G_j). (b) t-SNE projection [19] of extracted features (best viewed in color). Top-left: F_{VGG-19} ; top-right: F_{image} ; bottom-left: F_{task} (forming single-domain task-level clusters as in Table 1); bottom-right: F_{group} (forming multi-domain group-level clusters as in Table 1).

D_i is originally imaged for a specific purpose (e.g., CT for spleen). Generally, images belonging to a same dataset are similar inherently. As shown in Fig. 3(b), images of different modalities and ROIs are easier to distinguish. For abdominal CTs of spleen and liver, although the images are similar, their contents are different. Thus, each task’s dataset forms a single domain of certain ROI and image types. (3) *Group-level*: Despite the differences among different segmentation tasks, the contents of images may show a different degree of similarity. For example, in the physical space, liver CT scans have overlapping with spleen CT scans; cardiac MRIs scanned for different purposes (e.g., diverse cardiovascular structures) contain the same ROI (i.e., the heart) regardless of the image size and contrast. In this way, we categorize multiple domains of images into a group, which forms a multi-domain cluster in the feature space. Assigned with both task-level and group-level labels, each image constitutes a tuple (I, y^t, y^g) , where t and g are task-class and group-class, respectively (see Table 1).

Further, to better aggregate low- and high-level features from the encoder, we compress multi-scale feature vectors from the feature pyramid and concatenate them together, and then attach three different projection heads to automatically extract hierarchical representations (see Fig. 2).

Image-Level Loss. Given an input image I , the contrastive loss is formulated as: $l(\tilde{I}, \hat{I}) = -\log \frac{\exp\{\text{sim}(\tilde{z}, \hat{z})/\tau\}}{\exp\{\text{sim}(\tilde{z}, \hat{z})/\tau\} + \sum_{I \in \Lambda^-} \exp\{\text{sim}(\tilde{z}, \bar{z})/\tau\}}$, where $\tilde{z} = P_l(E(\tilde{I}))$, $\hat{z} = P_l(E(\hat{I}))$, $\bar{z} = P_l(E(\bar{I}))$, $P_l(\cdot)$ is the image-level projection head, $E(\cdot)$ is the encoder, \tilde{I} and \hat{I} are two different augmentations of image I (i.e., $\tilde{I} = \tilde{t}(I)$ and $\hat{I} = \hat{t}(I)$), $\bar{I} \in \Lambda^-$ consisting of all negative samples of I , and $\tilde{t}, \hat{t} \in \mathcal{T}$ are two augmentations. The augmentations \mathcal{T} include random cropping, resizing, blurring, and adding noise. $\text{sim}(\cdot, \cdot)$ is cosine similarity, and τ is a temperature scaling parameter. Given our multi-domain dataset \mathcal{D} , the image-level loss is

Table 1. Details of our data obtained from public sources. The left two columns: their task-classes and group-classes based on our multi-domain data aggregation principles.

Task ID	Group ID	ROI-Type	Segmentation class	# of slices	Source
1	1	Heart-MRI	1: left atrium	1262	LASC [28]
2	2	Liver-CT	1: liver, 2: tumor	4342	LiTS [3]
3	3	Prostate-MRI	1: central gland, 2: peripheral zone	483	MSD [26]
4	2	Pancreas-CT	1: Pancreas, 2: tumor	8607	MSD [26]
5	2	Spleen-CT	1: spleen	1466	MSD [26]
6	4	Knee-MRI	1: femur bone, 2: tibia bone, 3: femur cartilage, 4: tibia cartilage	8187	Knee [30]
7	1	Heart-MRI	1: left ventricle, 2: right ventricle,	1891	ACDC [2]
8	1	Heart-MRI	3: myocardium	3120	M&Ms [1]

defined as: $\mathcal{L}_{img} = \frac{1}{|\Lambda^+|} \sum_{(\tilde{I}, \hat{I}) \in \Lambda^+} [l(\tilde{I}, \hat{I}) + l(\hat{I}, \tilde{I})]$, where Λ^+ is a set of all similar pairs sampled from \mathcal{D} .

Task-Level Loss & Group-Level Loss. Given task-class and group-class, we formulate task- and group-level pretext tasks as classification tasks. The training objectives are: $\mathcal{L}_{task} = -\sum_{c=1}^T y_c^t \log(p_c^t)$; $\mathcal{L}_{group} = -\sum_{c=1}^G y_c^g \log(p_c^g)$, where $p_c^t = P_t(E(I))$, $p_c^g = P_g(E(I))$, $P_t(\cdot)$ (or $P_g(\cdot)$) is the task-level (or group-level) projection head, $E(\cdot)$ is the encoder, y_c^t (or y_c^g) is the task-class (or group-class) of input image I , and T (or G) is the number of classes of tasks (or groups).

We visualize some sample learned features in Fig. 3(b), in which the hierarchical layout is as expected, implying that our model is capable of extracting richer semantic features at different abstract levels of the input images.

Decoder Initialization. A decoder is also indispensable for semantic segmentation tasks. We devise a multi-scale decoder and formulate a reconstruction pretext task. The loss is defined as: $\mathcal{L}_{rec} = \frac{1}{|\mathcal{D}|} \sum_{I \in \mathcal{D}} \|S(E(I)) - I\|_2$, where $E(\cdot)$ is the encoder, $S(\cdot)$ is the decoder, and $\|\cdot\|_2$ is the L_2 norm.

In summary, we combine the hierarchical self-supervised losses at all the levels and the auxiliary reconstruction loss to jointly optimize the model: $\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{img} + \lambda_2 \mathcal{L}_{task} + \lambda_3 \mathcal{L}_{group} + \lambda_4 \mathcal{L}_{rec}$, where $\lambda_i (i = 1, 2, 3, 4)$ are the weights to balance loss terms. For simplicity, we let $\lambda_1 = \lambda_2 = \lambda_3 = 1/3, \lambda_4 = 50$.

Segmentation. Once trained, the encoder-decoder can be fine-tuned for downstream multi-domain segmentation tasks. For a give task D_i , we acquire some annotations (e.g., 10%) and optimize the network with cross-entropy loss.

3 Experiments and Results

Datasets and Experimental Setup. We employ multiple MRI and CT image sets from 8 different data sources with distribution shift, and sample 2D slices from each stack (see Table 1 for a summary of their sample numbers and downstream tasks). Each dataset is split into X_{tr} , X_{val} , and X_{te} in the ratios of 7 : 1 : 2. We use all images for the pre-training stage and then fine-tune the pre-trained network with labeled images from X_{tr} . We experiment with different amounts of training data X_{tr}^s , where $s \in \{5\%, 10\%, 100\%\}$ denotes the ratio of $\frac{X_{tr}^s}{X_{tr}}$. The segmentation accuracy is measured by the Dice-Sørensen Coefficient.

Table 2. Quantitative results on Task-1 (heart), Task-3 (prostate), and Task-5 (spleen). Dice scores for each class are listed and the average scores are in parentheses. TFS: training from scratch. Same network architecture is used for fair comparison in all the experiments. Our HSSL achieves the best performance in most settings (in bold).

Task-#	Anno.	TFS	Rotation [9]	In-painting [24]	MoCo [11]	SimCLR [6]	HSSL (Ours)
1	5%	71.56	72.83	65.40	75.97	73.45	81.46
	10%	79.64	82.31	81.99	79.07	81.19	81.79
	100%	85.81	87.43	86.56	87.19	87.06	87.65
3	5%	20.65; 47.56 (34.10)	28.74; 67.11 (47.93)	20.13; 52.16 (36.14)	29.55; 64.95 (47.25)	39.67; 68.35 (54.01)	35.30; 70.08 (52.69)
	10%	40.10; 66.95 (53.53)	44.15; 70.63 (57.39)	33.81; 67.14 (50.48)	40.16; 67.98 (54.07)	46.04; 70.39 (58.22)	46.97; 72.21 (59.59)
	100%	50.19; 76.74 (63.47)	55.21; 78.21 (66.71)	53.19; 77.97 (65.59)	56.31; 77.59 (66.95)	56.53; 77.86 (67.20)	58.80; 78.35 (68.58)
5	5%	48.75	56.74	47.86	54.91	63.40	67.35
	10%	67.44	74.68	71.30	68.22	78.25	80.95
	100%	85.88	86.96	85.96	85.75	87.76	88.45

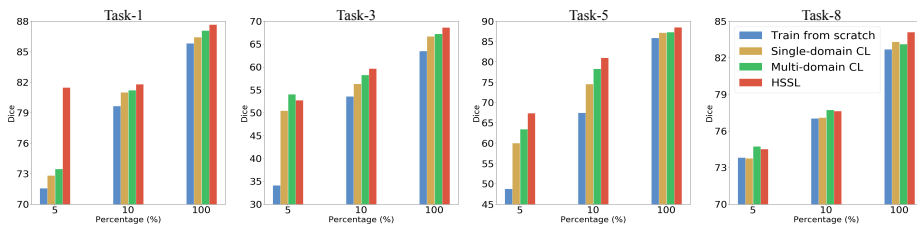


Fig. 4. Quantitative results of TFS *vs.* single-domain CL *vs.* multi-domain CL *vs.* HSSL for Task-1/-3/-5/-8 with different ratios (5%, 10%, 100%) of labeled data, respectively.

Implementation Details. For self-supervised pre-training, we use ResNet-34 [13] as the base encoder network, FC layers to obtain latent vectors, and sequential DeConv layers to reconstruct images. Detailed structures can be found in Supplementary Material. The model is optimized using Adam with a linear learning rate scaling for 1k epochs (initial learning rate = $3e^{-4}$). For segmentation tasks, we optimize the network using Adam with the “poly” learning rate policy, $L_r \times (1 - \frac{epoch}{\#epoch})^{0.9}$, where the initial learning rate $L_r = 5e^{-4}$ and $\#epoch = 10k$. Random cropping and rotation are applied for augmentation. In all the experiments, the mini-batch size is 30 and input image size is 192×192 .

Main Results. Our approach contributes to the “pre-training + fine-tuning” scheme in two aspects: hierarchical self-supervised learning (HSSL) and multi-domain data aggregation. **(1) Effectiveness of HSSL.** We compare with state-of-the-art pretext task training methods [6, 9, 11, 24] on seven downstream segmentation tasks, and quantitative results of three representative tasks are summarized in Table 2. First, our method surpasses training from scratch (TFS) substantially, showing the effectiveness of better model initialization. More can be found in Supplementary Material. Second, our approach outperforms known SSL-based methods in almost all the settings, indicating a better capability to extract features for segmentation tasks. Third, our HSSL can more effectively boost performance, especially when extremely limited annotations are available (e.g., +18.60% with 5% annotated data on Task-3), implying potential applicability when abundant images are acquired but few are labeled.

Table 3. Quantitative results of different models on Task-1/-3/-5 with 5%, 10%, and 50% annotated data, respectively. Our HSSL achieves the best performance in most the settings (highest scores in bold).

Method	Param. (M)	Task-1 (heart)			Task-3 (prostate)			Task-5 (spleen)		
		5%	10%	50%	5%	10%	50%	5%	10%	50%
UNet [25]	39.40	75.43	77.72	86.75	38.19	49.44	62.61	54.71	62.81	81.48
UNet3+ [15]	26.97	78.48	78.81	87.52	42.06	50.94	63.50	60.05	64.83	82.74
HSSL (Ours)	22.07	81.46	81.79	87.02	52.69	59.59	66.64	67.35	80.95	85.86

Table 4. Ablation study of loss functions on Task-1 & Task-5 w/ 5% anno. data.

L_{rec}	L_{img}	L_{task}	L_{group}	Task-1	Task-5
✓				65.71	46.13
	✓			73.45	63.40
✓	✓			77.26	65.01
✓		✓		79.32	66.67
✓	✓	✓	✓	81.46	67.35

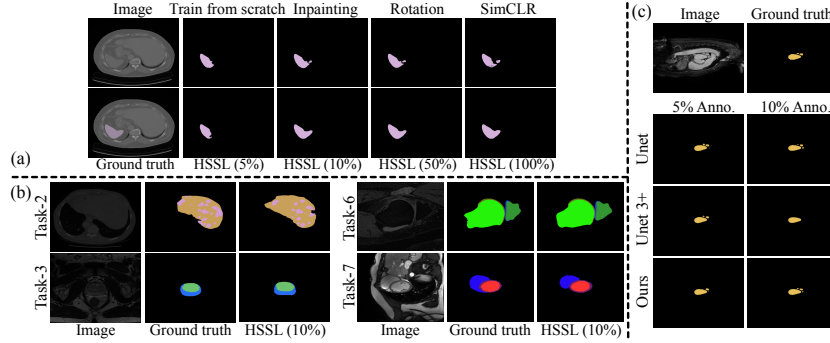


Fig. 5. Qualitative comparison (best viewed in color). (a) Top: results of different methods on Task-5 (10% annotated data); Bottom: results of our HSSL with different ratios of annotated data. (b) Results of Task-2/-3/-6/-7 (10% annotated data). (c) Results of different models on Task-1 trained with 5% and 10% annotated data, respectively.

Fourth, with more annotations, our method can further improve accuracy and achieve state-of-the-art performance (e.g., +1.84% to +2.57% with 100% annotated data over TFS). Qualitative results are given in Fig. 5 and Supplementary Material. **(2) Effectiveness of Multi-Domain Data Aggregation.** We conduct pre-training on single-domain and aggregated multi-domain data, and compare the segmentation performances. “Single-domain CL” and “Multi-domain CL” are all based on the state-of-the-art SimCLR [6]. As sketched in Fig. 4, one can see that multi-domain data aggregation consistently outperforms (sometimes significantly) single-domain pre-training (e.g., with 10% annotated data on Task-5, multi-domain CL and HSSL outperform single-domain CL by 3.74% and 6.41%, respectively). This suggests that more data varieties can provide complementary information and help improve the overall performance.

Discussions. **(1) Comparison with State-of-the-Art Models.** As shown in Table 3, our method outperforms the state-of-the-art UNet3+ [15] significantly in almost all the settings. Further, with limited annotated data (e.g., 5%), our method bridges the performance gap significantly w.r.t. the results obtained by training with more annotated data. Also, our model is most lightweight, and thus efficient as well. Qualitative results are given in Fig. 5(c). **(2) Ablation Study.** As shown in Table 4, each hierarchical loss contributes to representation learning and leads to segmentation improvement.

4 Conclusions

In this paper, we proposed *hierarchical self-supervised learning*, a novel self-supervised framework that learns hierarchical (image-, task-, and group-levels) and multi-scale semantic features from aggregated multi-domain medical image data. A decoder is also initialized for downstream segmentation tasks. Extensive experiments demonstrated that joint training on multi-domain data by our method outperforms training from scratch and conventional pre-training strategies, especially in limited annotation scenarios.

Acknowledgement. This research was supported in part by the U.S. National Science Foundation through grants IIS-1455886, CCF-1617735, CNS-1629914, and IIS-1955395.

References

1. Multi-centre, multi-vendor & multi-disease cardiac image segmentation challenge (M&Ms). <https://www.ub.edu/mmms/>, accessed: 2021-07-01
2. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging* **37**(11), 2514–2525 (2018)
3. Bilic, P., Christ, P.F., Vorontsov, E., Chlebus, G., Chen, H., Dou, Q., Fu, C.W., Han, X., Heng, P.A., Hesser, J., et al.: The liver tumor segmentation benchmark (LiTS). *arXiv preprint arXiv:1901.04056* (2019)
4. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. In: *NeurIPS*. pp. 12546–12558 (2020)
5. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D.: Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis* **58**, 101539 (2019)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *ICML*. pp. 1597–1607 (2020)
7. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: *NeurIPS*. pp. 22243–22255 (2020)
8. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: *ICCV*. pp. 1422–1430 (2015)
9. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: *ICLR* (2018)
10. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. In: *NeurIPS*. pp. 21271–21284 (2020)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: *CVPR*. pp. 9729–9738 (2020)
12. He, K., Girshick, R., Dollár, P.: Rethinking ImageNet pre-training. In: *ICCV*. pp. 4918–4927 (2019)

13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
14. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. In: ICLR (2019)
15. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: UNet 3+: A full-scale connected UNet for medical image segmentation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1055–1059 (2020)
16. Larsson, G., Maire, M., Shakhnarovich, G.: Learning representations for automatic colorization. In: ECCV. pp. 577–593 (2016)
17. Liang, P., Chen, J., Zheng, H., Yang, L., Zhang, Y., Chen, D.Z.: Cascade decoder: A universal decoding method for biomedical image segmentation. In: Proceedings of the IEEE International Symposium on Biomedical Imaging. pp. 339–342 (2019)
18. Liu, S., Xu, D., Zhou, S.K., Grbic, S., Cai, W., Comaniciu, D.: Anisotropic hybrid network for cross-dimension transferable feature learning in 3D medical images. In: Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics, pp. 199–216 (2019)
19. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**(11), 2579–2605 (2008)
20. Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T.: Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation. In: ISBI. pp. 1038–1042 (2018)
21. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV. pp. 69–84 (2016)
22. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018)
23. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In: ECCV. pp. 762–780 (2020)
24. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR. pp. 2536–2544 (2016)
25. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
26. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* (2019)
27. Tao, X., Li, Y., Zhou, W., Ma, K., Zheng, Y.: Revisiting rubik’s cube: self-supervised learning with volume-wise transformation for 3D medical image segmentation. In: MICCAI. pp. 238–248 (2020)
28. Tobon-Gomez, C., Geers, A.J., Peters, J., Weese, J., Pinto, K., Karim, R., Amar, M., Daoudi, A., Margeta, J., Sandoval, Z., et al.: Benchmark for algorithms segmenting the left atrium from 3D CT and MRI datasets. *IEEE Transactions on Medical Imaging* **34**(7), 1460–1473 (2015)
29. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: MICCAI. pp. 399–407 (2017)
30. Yin, Y., Zhang, X., Williams, R., Wu, X., Anderson, D.D., Sonka, M.: LOGISMOS—layered optimal graph image segmentation of multiple objects and surfaces:

- Cartilage segmentation in the knee joint. *IEEE Transactions on Medical Imaging* **29**(12), 2023–2037 (2010)
31. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation. In: *MICCAI*. pp. 605–613 (2019)
 32. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: *MICCAI*. pp. 408–416 (2017)
 33. Zheng, H., Yang, L., Chen, J., Han, J., Zhang, Y., Liang, P., Zhao, Z., Wang, C., Chen, D.Z.: Biomedical image segmentation via representative annotation. In: *AAAI*. pp. 5901–5908 (2019)
 34. Zheng, H., Yang, L., Han, J., Zhang, Y., Liang, P., Zhao, Z., Wang, C., Chen, D.Z.: HFA-Net: 3D cardiovascular image segmentation with asymmetrical pooling and content-aware fusion. In: *MICCAI*. pp. 759–767 (2019)
 35. Zheng, H., Zhang, Y., Yang, L., Wang, C., Chen, D.Z.: An annotation sparsification strategy for 3D medical image segmentation via representative selection and self-training. In: *AAAI*. pp. 6925–6932 (2020)
 36. Zhou, Z., Shin, J., Zhang, L., Gurudu, S., Gotway, M., Liang, J.: Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally. In: *CVPR*. pp. 7340–7351 (2017)
 37. Zhou, Z., Sodha, V., Siddiquee, M.M.R., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J.: Models genesis: Generic autodidactic models for 3D medical image analysis. In: *MICCAI*. pp. 384–393 (2019)
 38. Zhuang, X., Li, Y., Hu, Y., Ma, K., Yang, Y., Zheng, Y.: Self-supervised feature learning for 3D medical images by playing a Rubik’s cube. In: *MICCAI*. pp. 420–428 (2019)

Supplementary Materials

Table 1. The configurations of Encoder, Decoder, Feature Fusion, & Projection Head.

Encoder			
Layer name	Operators	Output size	
Input	—	$192 \times 192 \times 3$	
conv1-1	$3 \times 3, 64$	$192 \times 192 \times 64$	
conv1-2	$3 \times 3, 64, /2$	$96 \times 96 \times 64$	
layer-1	$\left\{ \begin{matrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{matrix} \right\} \times 3$	$48 \times 48 \times 64$	
layer-2	$\left\{ \begin{matrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{matrix} \right\} \times 4$	$24 \times 24 \times 128$	
layer-3	$\left\{ \begin{matrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{matrix} \right\} \times 6$	$12 \times 12 \times 256$	
layer-4	$\left\{ \begin{matrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{matrix} \right\} \times 3$	$6 \times 6 \times 512$	

Decoder			
Previous layer	Layer name	Operators	Output size
layer-4	up-4	$4 \times 4, 64, 2 \times$	$12 \times 12 \times 64$
		$4 \times 4, 32, 2 \times$	$24 \times 24 \times 32$
		$4 \times 4, 16, 2 \times$	$48 \times 48 \times 16$
		$4 \times 4, 8, 2 \times$	$96 \times 96 \times 8$
layer-3	up-3	$4 \times 4, 4, 2 \times$	$192 \times 192 \times 4$
		$4 \times 4, 32, 2 \times$	$24 \times 24 \times 32$
		$4 \times 4, 16, 2 \times$	$48 \times 48 \times 16$
		$4 \times 4, 8, 2 \times$	$96 \times 96 \times 8$
layer-2	up-2	$4 \times 4, 4, 2 \times$	$192 \times 192 \times 4$
		$4 \times 4, 16, 2 \times$	$48 \times 48 \times 16$
		$4 \times 4, 8, 2 \times$	$96 \times 96 \times 8$
		$4 \times 4, 4, 2 \times$	$192 \times 192 \times 4$
layer-1	up-1	$4 \times 4, 8, 2 \times$	$96 \times 96 \times 8$
		$4 \times 4, 4, 2 \times$	$192 \times 192 \times 4$
up-1~up-4	Fout-0	concatenation	$192 \times 192 \times 16$
Fout-0	Fout-1	$4 \times 4, 16$	$192 \times 192 \times 16$
Fout-1	Fout-2	$4 \times 4, N$	$192 \times 192 \times N$

Feature Fusion			
Previous layer	Layer name	Operators	Output size
layer-4	V4	Avg-Pool(6), 2 <i>fc</i> -layers	512
layer-3	V3	Avg-Pool(12), 2 <i>fc</i> -layers	256
layer-2	V2	Avg-Pool(24), 2 <i>fc</i> -layers	128
V2~V4	V _{all}	concatenation	896

Projection Head			
Previous layer	Layer name	Operators	Output size
V _{all}	H1	2 <i>fc</i> -layers	512
V _{all}	H2	2 <i>fc</i> -layers	512
V _{all}	H3 (i.e., F_I)	2 <i>fc</i> -layers	512
H1	F_G	2 <i>fc</i> -layers	4
H2	F_T	2 <i>fc</i> -layers	8

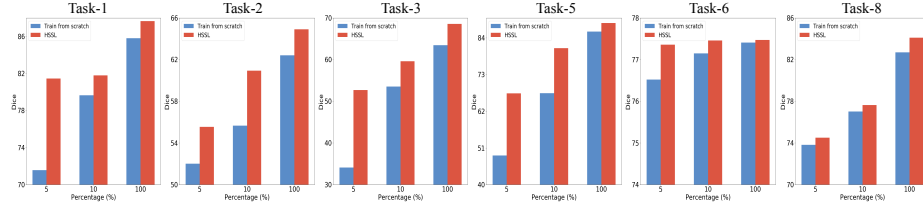


Fig. 1. Quantitative results of Task-1 (heart), Task-2 (liver), Task-3 (prostate), Task-5 (spleen), Task-6 (knee), and Task-8 (heart) with 5%, 10%, and 100% annotated data, respectively.

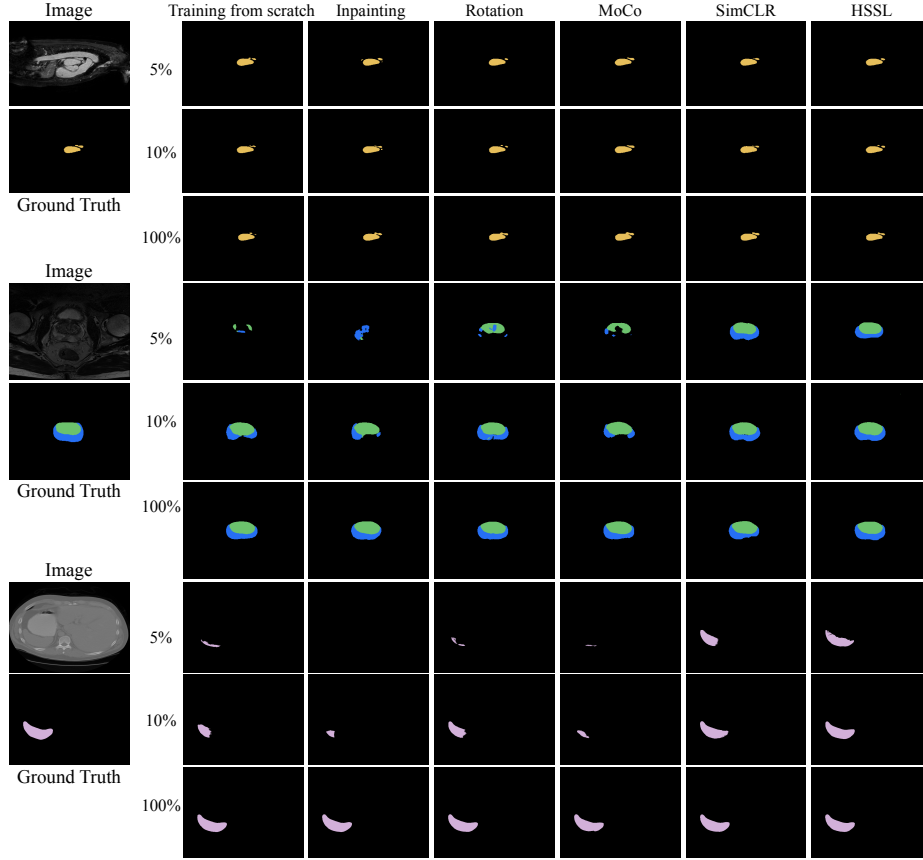


Fig. 2. Qualitative results of several known methods and our HSSL on Task-1 (heart), Task-3 (prostate), and Task-5 (spleen) trained with different ratios (5%, 10%, and 100%) of annotated data, respectively. One can observe that our HSSL consistently yields good results, especially in extremely sparse annotation scenarios.

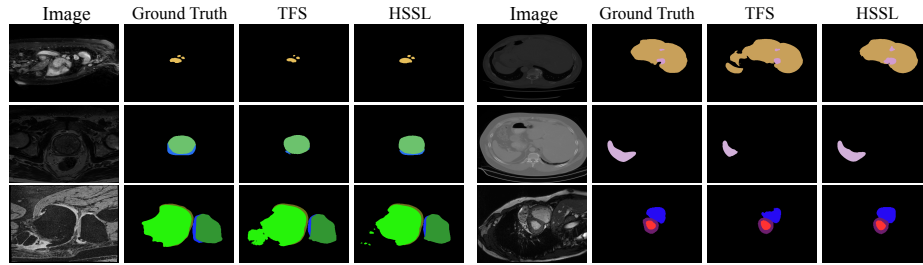


Fig. 3. Comparison of qualitative results between TFS (training from scratch) and our proposed HSSL using 10% annotated data on six segmentation tasks: Task-1 (heart), Task-2 (liver), Task-3 (prostate), Task-5 (spleen), Task-6 (knee), and Task-7 (heart) (from left to right, top to bottom).