ASA: Learning Anatomical Consistency, Sub-volume Spatial Relationships and Fine-grained Appearance for CT Images

Jiaxuan Pang[1], DongAo Ma[1], Ziyu Zhou[2], Michael B. Gotway[3], and Jianming Liang[1]

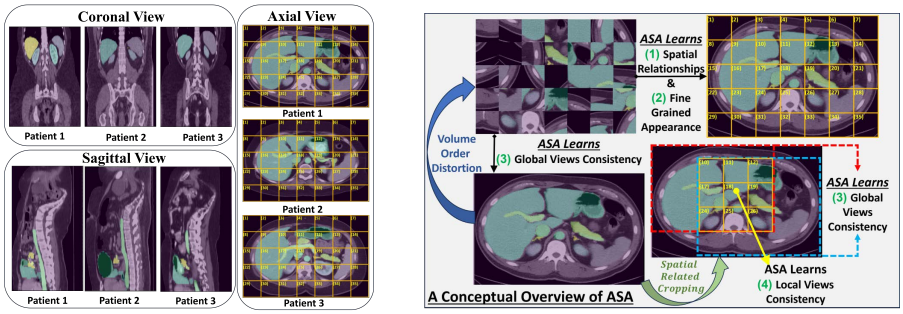
[1] Arizona State University, Tempe, AZ 85281, USA
{jpang12,dongaoma,jianming.liang}@asu.edu
[2] Shanghai Jiao Tong University, China
zhouziyu@sjtu.edu.cn
[3] Mayo Clinic, Scottsdale, AZ 85259, USA
Gotway.Michael@mayo.edu

Abstract. To achieve superior performance, deep learning relies on copiousness, high-quality, annotated data, but annotating medical images is tedious, laborious, and time-consuming, demanding specialized expertise, especially for segmentation tasks. Segmenting medical images requires not only macroscopic anatomical patterns but also microscopic textural details. Given the intriguing symmetry and recurrent patterns inherent in medical images, we envision a powerful deep model that exploits high-level context, spatial relationships in anatomy, and low-level, fine-grained, textural features in tissues in a self-supervised manner. To realize this vision, we have developed a novel self-supervised learning (SSL) approach called ASA to learn **a**natomical consistency, sub-volume **s**patial relationships, and fine-grained **a**ppearance for 3D computed tomography images. The novelty of ASA stems from its utilization of intrinsic properties of medical images, with a specific focus on computed tomography volumes. ASA enhances the model's capability to learn anatomical features from the image, encompassing global representation, local spatial relationships, and intricate appearance details. Extensive experimental results validate the robustness, effectiveness, and efficiency of the pretrained ASA model. With all code and pretrained models released at GitHub.com/JLiangLab/ASA, we hope ASA serves as an inspiration and a foundation for developing enhanced SSL models with a deep understanding of anatomical structures and their spatial relationships, thereby improving diagnostic accuracy and facilitating advanced medical imaging applications.

Keywords: Self-supervised Learning · Anatomical Structure Learning

1 Introduction

SSL has emerged as a transformative paradigm which enables the deep learning model to autonomously learn from data without expert labels, as annotating



(a) Three patients across 3 views (b) An overview of ASA

Fig. 1: (a) shows CT views from three different patients, highlighting significant anatomical similarities. To illustrate anatomical patterns, the axial view is divided into 35 non-overlapping grids. The inferior vena cava appears across grids #17 and #18, and the aorta is consistently located in grid #18 for all patients. The liver, the most prominent organ, occupies multiple grids shared by all patients. This uniformity underscores the importance of learning spatial relationships and high-level anatomical patterns. (b) provides an overview of ASA, which captures spatial relationships via order prediction, fine-grained features via appearance recovery, global features by maximizing agreement between two same-patient views, and local features by aligning the shared information within two views.

medical images is a laborious, time-consuming task that demands specialized expertise [10, 7]. Medical images often exhibit intriguing symmetry and recurrent patterns. As depicted in Fig. 1a, significant similarities are evident in axial, coronal, and sagittal views across diverse patients in CT images, and major organs appear in the same location. These similarities suggest that a robust model is expected to capture the overarching concept of shared appearances and features (i.e., anatomical structures) in CT volumes across all patients. By leveraging the symmetry and recurring nature of body structures, the model is expected to effectively identify high-level anatomical structures and intra-volume spatial relationships. However, despite the significant similarities observed across all CT volumes, there are still subtle differences present in each individual. Therefore, the model must also capture fine-grained features to discern and account for patient-specific distinctions.

In organ segmentation, illustrated in Fig. 1a, larger organs such as liver and spleen exhibit larger regions of interest, while smaller organs like the esophagus and adrenal glands have smaller region of interest, necessitating more meticulous attention. An effective model must capture organ-specific relationships and appearance features, but also deliver precise pixel-level details of small regions. Therefore, a crucial question naturally arises: *How to develop an SSL framework that enables the model to acquire nuanced fine-grained features, com-*

prehend high-level global features, emphasize local-level embeddings, and capture contextual relationship features? To address this question, we have developed ASA to learn **a**natomical consistency, sub-volume **s**patial relationships, and fine-grained **a**ppearance. As illustrated in Fig. 1b, ASA incorporates four learning perspectives: (1) capturing sub-volume relationships through 3D sub-volume order prediction, (2) depicting fine-grained features within volumes through volume appearance recovery, (3) comprehending high-level global features by maximizing the agreement between two spatially related views using the student-teacher network, and (4) acquiring local features at the sub-volume level by aligning the shared local views within two spatially related views.

ASA is *different* from the distorted image recovery task [12, 19, 21] by focusing on reconstructing the correct volume from a set of displaced sub-volumes to capture fine-grained volume appearances and underlying structures. ASA is also distinguished from contrastive learning methods [1, 11] which aim to maximize agreement between two positive views, by further aligning the shared local views within these views. Moreover, ASA diverges from image context learning [17, 10, 2] by incorporating a student-teacher network to optimize global and local consistency between two spatially-related views, thereby facilitating the acquisition of generalized volume features. Inspired by [9], where the cyclic pretraining strategy and the student-teacher networks have been demonstrated effective in accumulating knowledge across various tasks, ASA further introduces an alternate learning strategy to enhance SSL from multiple perspectives. Through this work, we have made the following contributions:

1. A novel vision transformer-based SSL framework for 3D medical images that simultaneously captures high-level anatomical information, intra-volume relationships, and fine-grained appearance features.

2. Introduction of an alternate pretraining strategy involving a student-teacher network to facilitate learning from multiple perspectives.

3. Comprehensive experiments showcasing the transferability of ASA across diverse single-organ and multi-organ segmentation tasks, surpassing the performance of multiple supervised and SSL methods.

4. An efficient pretrained model that encapsulates rich semantic information, demonstrating superior label efficiency.

2 Method

To develop a comprehensive understanding of anatomical structures depicted in CT images, particularly focusing on spatial relationships and fine-grained features, as demonstrated in phase 1 of Fig. 2, ASA incorporates the following components: **(1)** a sub-volume order prediction module to capture intra-volume spatial relationships, and **(2)** a volume appearance recovery module to represent volume-wise fine-grained features. Additionally, to establish a high-level semantic context within a CT image, ASA employs the student-teacher learning paradigm, **(3)** aligning the global features extracted by the student's encoder with permuted volumes to those of the teacher's with original volumes. Fur-

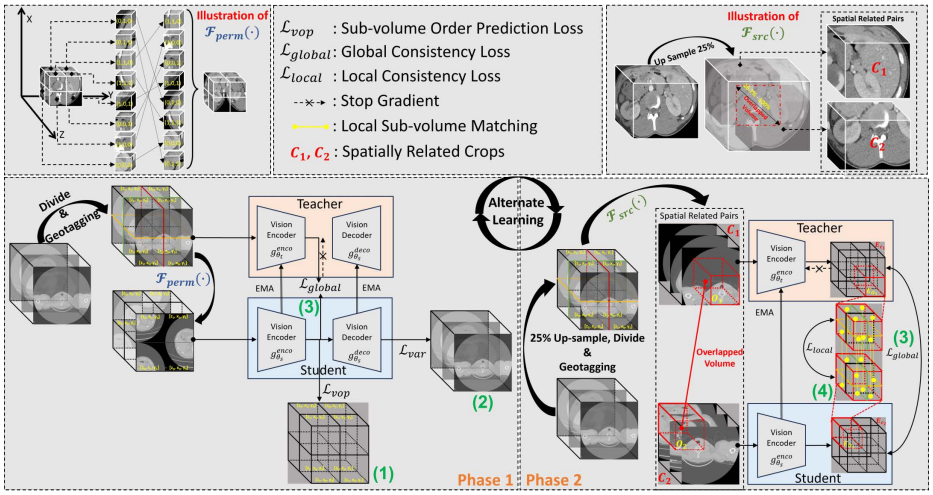


Fig. 2: ASA accquares anatomical knowledge through two learning phases. phase 1 focuses on capturing sub-volume relationships via sub-volume order prediction, depicting fine-grained features via volume appearance recovery, and constructing high-level semantic context of an image via global feature alignment between the permuted and original volumes. phase 2 enhances ASA's ability to summarize global features and delineate local features by optimizing the agreement between two spatially related views through the student-teacher learning paradigm. During phase 1, a CT image is divided into sub-volumes whose order is distorted using $\mathcal{F}_{perm}(\cdot)$ before being fed into the student model. The model is then trained to predict the original sub-volume order using the loss $\mathcal{L}_{vop}$, and to recover the volume's original appearance using the loss $\mathcal{L}_{ar}$. Additionally, the sub-volumes with the original order are fed into the teacher model to generate global features. These features are then utilized to compute the consistency loss $\mathcal{L}_{gCons}$ with the features obtained from the student model. In phase 2, the CT image is upsampled and cropped into two spatially related (overlapped) views using $\mathcal{F}_{src}(\cdot)$. Similar to **(3)**, these two views are fed into the student and teacher models to generate global features for maximizing agreement using $\mathcal{L}_{gCons}$. Concurrently, the local features from the overlapped region of the two views are utilized to further enhance agreement between the student and teacher models, employing the loss $\mathcal{L}_{lCons}$. In both phases, the *teacher* model is updated after each iteration using exponential moving average (EMA) based on the *student*'s weights. To stabilize and expedite training, the model alternates between these two learning phases. Once trained, the *teacher* model is transferred to downstream tasks.

thermore, to enhance ASA's ability to summarize global features and delineate local features, as depicted in phase 2 of Fig. 2, **(4)** spatially related cropping is employed, maximizing the agreement between teacher and student by align-

ing global and local features extracted from two spatially related crops. The following details ASA, we provide detailed training pseudo-code in Appendix D.

Learning intra-volume relationship and fine-grained appearance. In phase 1, depicted in Fig. 2, the original volume is given to the teacher network, generating embedding of the original volume appearance, while the order-distorted volume, obtained by $\mathcal{F}_{perm}(\cdot)$, is fed to the student network. The objective of sub-volume order prediction is to anticipate the accurate 3D coordinates of a sub-volume from its appearance (phase 1, **(1)**), while volume appearance recovery endeavors to rebuild the original volume from an order-distorted one (phase 1, **(2)**). Meanwhile, to stabilize the reconstruction process and ensure maximum preservation of global features, we instruct the student network to align the original appearance embedding generated by the teacher with the distorted appearance embedding produced by the student network (phase 1, **(3)**).

Acquiring global and local embedding consistency from two related views. As depicted in phase 2 of Fig. 2, two spatially related crops C_1 and C_2, obtained by $\mathcal{F}_{src}(\cdot)$, are input to the teacher and student networks, respectively. The objective of global embedding consistency is to enhance the general embedding level agreement between these two spatially related crops C_1 and C_2 (phase 2, **(3)**). To ensure alignment of the local embedding, we devise a sub-volume matching process that maximizes agreement between the local embeddings, generated from two overlapped sub-volumes, showing near phase 2, **(4)**.

Overall training scheme. As depicted in Fig. 2, we conduct pretraining of the student network by alternately propagating the loss $\mathcal{L}_{vopar} = \lambda_{vop} * \mathcal{L}_{vop} + \lambda_{var} * \mathcal{L}_{var} + \lambda_{global} * \mathcal{L}^{global}_{\theta_s,\theta_t}$ in phase 1 and $\mathcal{L}_{consistency} = \lambda_{global} * \mathcal{L}^{global}_{\theta_s,\theta_t} + \lambda_{local} * \mathcal{L}^{local}_{\theta_s,\theta_t}$ in phase 2, where λ_{vop}, λ_{var}, λ_{global} and λ_{local} are regularization factor contributing the importance of the learning task. θ_s and θ_t are student and teacher networks, respectively. We optimize $\mathcal{L}^{global}_{\theta_s,\theta_t}$ by minimizing l_2 distance between the predicted volume and the original volume, both $\mathcal{L}^{global}_{\theta_s,\theta_t}$ and $\mathcal{L}^{local}_{\theta_s,\theta_t}$ are optimized by minimizing the l_2 distance between two normalized volume and sub-volume embedding, respectively. Finally, we define $\mathcal{L}_{vop}$ as a regression task by minimizing l_2 distance between the predicted sub-volume coordinates and the randomly shuffled coordinates generated by $\mathcal{F}_{perm}(\cdot)$. Only the student's encoder and decoder are updated by $\mathcal{L}_{vopar}$, while $\mathcal{L}_{consistency}$ updates only the student's encoder. The weights of all learnable networks are shared between the two phases. Additionally, to summarize and consolidate the knowledge acquired from the two phases, we introduce a teacher model with the same architecture as the student. The teacher network is updated using EMA [16] based on the learning experience of the student. Consequently, the learned sub-volume-wise relationships, volume-wise fine-grained features, and overall context are refined within the teacher model for future application-specific downstream tasks.

Table 1: ASA excels in both supervised and SSL techniques, achieving the highest average *dice* score in segmenting all organs on the BTCV dataset. With three comprehensive learning objectives, ASA outperforms supervised competitors in segmenting 9/12 organs and SSL competitors in another 9/12 organs.

Methods/ Organs‡	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Vins	Pan	AG	Avg.
RandPatch† [14]	95.82%	88.52%	90.14%	68.31%	75.01%	96.48%	82.93%	88.96%	82.49%	73.54%	75.48%	66.09%	81.98%
TransBTS† [18]	94.59%	89.23%	90.47%	68.50%	75.59%	96.14%	83.72%	88.85%	82.28%	74.25%	75.12%	66.74%	82.12%
nnFormer† [20]	94.51%	88.49%	93.39%	65.51%	74.49%	96.10%	83.83%	88.91%	80.58%	75.94%	77.71%	68.19%	82.30%
UNETR† [4]	94.91%	92.10%	93.12%	**76.98%**	74.01%	96.17%	79.98%	89.74%	81.20%	75.05%	80.12%	62.60%	83.00%
nnU-Net† [5]	95.92%	88.28%	92.62%	66.58%	75.71%	96.49%	**86.05%**	88.33%	82.72%	**78.31%**	79.17%	67.99%	83.18%
SimMIM [19]	92.03%	93.66%	92.13%	69.16%	75.06%	96.21%	76.36%	89.80%	83.91%	72.46%	73.61%	68.24%	81.89%
Swin UNETR [15]	95.30%	**94.29%**	**94.22%**	74.01%	76.35%	96.71%	80.56%	90.42%	84.70%	75.12%	80.61%	67.25%	84.13%
ASA	**96.89%**	94.28%	94.10%	75.53%	**76.66%**	**96.79%**	82.42%	**92.03%**	**86.02%**	74.77%	**80.98%**	**70.73%**	**85.10%**

† Values are obtained from [8]The baseline performances were established by [8].
‡ Spl: spleen, RKid: right kidney, LKid: left kidney, Gall: gallbladder, Eso: esophagus, Liv: liver, Sto: stomach, Aor: aorta, IVC: inferior vena cava, Veins: portal and splenic veins, Pan: pancreas, AG: left and right adrenal glands.

3 Experiments and Results

ASA: We independently pretrained ASA on the AMOS2022 [6] dataset for the major evaluations (Table 1, Fig. 3, 4) and on the LUNA16 [13] dataset for the ablation study (Table 2, 3). Both models follow the same pretraining protocol. In phase 1, volumes are resized to $128 \times 128 \times 128$, with the sub-volume size $16 \times 16 \times 16$, leading to 512 unique sub-volumes and coordinates. The volume in phase 2 is up-sampled to $160 \times 160 \times 160$ before two spatially-related crops sized $128 \times 128 \times 128$ are obtained. The Swin UNETR [15, 3] architecture is employed as both the student and teacher networks. Our model undergoes a thorough comparison with both supervised and SSL baselines, revealing its superior performance across various metrics related to multi-organ segmentation task (Table 1), full finetuning, and linear probing evaluation on single organ segmentation tasks (Fig. 3), as well as the label efficiency examination task (Fig. 4). We conduct ablation studies to demonstrate effectiveness on 1D and 3D order encoding predictions (Table 2) and to examine performance differences across various ASA learning tasks (Table 3).

SimMIM and Swin UNETR: We pretrain the SimMIM [19] baseline on AMOS2022 [6] dataset, adhering to the official implementation and implementing the method in 3D on the Swin UNETR architecture with a 50% masking ratio. Swin UNETR [15, 3] undergoes a pretraining phase involving three common SSL tasks on five publicly accessible CT datasets. We obtained the pretrained model from its official GitHub release. Pretraining and evaluation protocols are detailed in Appendix A and B.

1) ASA outperforms state-of-the-art supervised and SSL methods on multi-organ segmentation challenge on average.

Experimental Setup: To showcase the performance enhancements achieved through ASA pretraining, we compare the ASA model with state-of-the-art supervised and SSL models. For a fair comparison, we obtained the supervised baseline performances and followed the train/valid/test split proposed by [8]. All models are finetuned on the same setup and follow the same evaluation protocol.

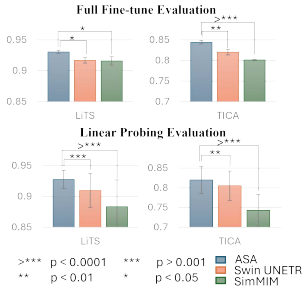


Fig. 3: Full finetuning and linear probing evaluation on single-organ segmentation tasks.

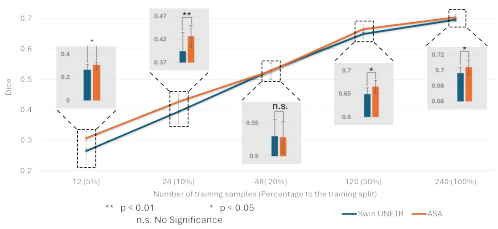


Fig. 4: ASA surpasses the SoTA Swin UNETR in label-efficient transfer learning on AMOS2022, underscoring its robust features learned from three advantageous tasks.

Result and Analysis: Showing in Table 1, ASA surpasses all reported methods on 13 organ segmentation tasks (with left and right adrenal glands combined) on the BTCV validation set. Its superior performance over all five supervised learning methods in segmenting 9/12 organs highlights ASA's effectiveness in acquiring generic appearance features for a variety of abdominal organs, despite being pretrained on only one abdominal dataset. Furthermore, ASA outperforms state-of-the-art SSL pretraining methods, SimMIM and Swin UNETR, recognized in the 2D natural/medical imaging and 3D medical imaging domains, respectively [19, 15]. The substantial margin achieved by ASA over SimMIM underscores the efficacy of learning anatomical relationships. Additionally, ASA outperforms Swin UNETR, which is pretrained via three proxy tasks to learn volume-level discriminative and rotation-invariant features for the thoracic and abdominal regions using five datasets. This suggests that more robust features can be learned by capturing anatomical structure through sub-volume order prediction and by depicting fine-grained appearance features through volume appearance recovery.

2) ASA offers generalized representations for single-organ segmentation tasks under both fully finetuned and linear probing scenarios.

Experimental Setup: To assess ASA's generalizability, we transfer the ASA model, along with two SSL models, to pancreas and liver segmentation tasks. Following the data split specified by [8], we evaluate the models on the Pancreas-CT dataset (80 scans) and the LiTS dataset (130 scans) under both fine-tuning and linear probing setups. For linear probing, we initialize the networks with the pretrained model's weights, freeze the encoder, and train only the decoder.

Result and Analysis: As depicted in Fig. 3, ASA surpasses both state-of-the-art SSL methods under both fine-tuning and linear probing setups. As the liver is a sizable organ in the abdominal region, all models demonstrate high performances as measured by the dice score. ASA attains the highest score, underscoring its superiority in delineating intricate edge details. In pancreas segmentation, ASA outperforms SimMIM by a significant margin and surpasses Swin UNETR by a more modest margin, highlighting the enhanced adaptability of features acquired

Table 2: Target task performance on 1D and 3D volume order prediction.

Predicted Order	BTCV	TCIA Pan	LiTS
1D sequence	82.48	82.02	73.29
3D coordinates	**83.16**	**82.79**	**73.52**

Table 3: Study on different ASA tasks and strategies, mean dice score for segmenting all organs is reported.

	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$	$\mathcal{S}_5$	$\mathcal{S}_6$	$\mathcal{S}_7$	$\mathcal{S}_8$	$\mathcal{S}_9$	$\mathcal{S}_{10}$	$\mathcal{S}_{11}$
@P1 $\mathcal{T}_{sop}$	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓
@P1 $\mathcal{T}_{gc}$	×	×	×	×	✓	×	✓	✓	×	×	✓
@P1 $\mathcal{T}_{var}$	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓
@P2 $\mathcal{T}_{lc}$	✓	×	×	×	×	×	×	×	×	✓	✓
@P2 $\mathcal{T}_{gc}$	✓	✓	×	×	×	✓	✓	✓	✓	✓	✓
@P3 $\mathcal{T}_{lc}$	×	×	×	×	×	×	×	✓	✓	×	×
BTCV	82.12	82.38	82.52	83.16	82.79	84.01	83.97	84.05	84.03	83.86	**84.42**

$\mathcal{T}_{sop}$ represents sub-volume order prediction task only, $\mathcal{T}_{var}$ represents volume appearance recovery task only, $\mathcal{T}_{gc}$ represents global consistency task only, and $\mathcal{T}_{lc}$ represents local consistency task only. @P1-3 represents the task performed at which learning stage. All models are pretrained on LUNA16 only.

through our method, which effectively captures spatial relationships and fine-grained features.

3) ASA exhibits better datalabel efficiency for multi-organ segmentation.

Experimental Setup: We finetune both ASA and Swin UNETR pretrained models on subsets comprising 12 (5%), 24 (10%), 48 (20%), 120 (50%), and 240 (100%) randomly selected samples from the official training split of AMOS2022 [6]. To ensure fairness across diverse random samples, we conduct five independent experiments and report their average performances. The mean Dice score for segmenting 15 organs is reported.

Result and Analysis: ASA demonstrates its superiority by surpassing Swin UNETR, a state-of-the-art SSL method that employs three learning objectives on 3D medical segmentation task benchmarks [15]. This achievement highlights the effectiveness of the ASA model in providing richer and more utilizable information. As depicted in Fig. 4, ASA significantly outperforms Swin UNETR in lower data regimes, with 12 (5%) and 24 (10%) training samples. While the performance of both models is comparable with 48 (20%) training samples, ASA continues to outperform Swin UNETR as the number of training samples increases (50% and 100%). This underscores the effectiveness of ASA in extracting fine-grained features and organ appearance information, even when pretrained on fewer datasets.

4) Ablation Study: Comparison among different learning tasks. Extensive ablation studies are conducted to compare different combinations of training tasks and strategies, demonstrating the superiority of current ASA setup in the BTCV task. Table 3 shows that models trained solely with crop consistency tasks ($\mathcal{S}_1$, $\mathcal{S}_2$) or focused only on recovering volume appearance ($\mathcal{S}_3$) exhibit the lowest performance. Adding sub-volume order prediction alongside appearance recovery ($\mathcal{S}_4$) slightly enhances performance, underscoring the importance of learning sub-volume relationships. Notably, adding $\mathcal{T}_{gc}$ ($\mathcal{S}_5$) causes the teacher to collapse, resulting in poor performance. Integrating alternative training and infusing global consistency in both stages ($\mathcal{S}_6$, $\mathcal{S}_7$) boosts performance. However, including local consistency and adding a third learning stage ($\mathcal{S}_8$, $\mathcal{S}_9$) do not yield any discernible benefits. The current ASA setup ($\mathcal{S}_{11}$), which incorporates $\mathcal{T}_{sop}$, $\mathcal{T}_{var}$, and $\mathcal{T}_{gc}$ in the first learning stage, and $\mathcal{T}_{gc}$ and $\mathcal{T}_{lc}$ in the second

learning cycle, showcases the most prominent performance. There is a significant performance drop when $\mathcal{T}gc$ is removed in the first stage ($\mathcal{S}10$), highlighting the importance of consistency between the original view embedding from the teacher and the expected view embedding from the student network.

5) Ablation Study: 1D sub-volume sequences and 3D sub-volume coordinates prediction. We evaluate the efficacy of using 1D sub-volume order presentation (e.g., 1, 2, 3, ..., k) versus 3D sub-volume order presentation (e.g., (0,0,0), ..., (3,3,5), ..., (z,x,y)). All models presented in Table 2 are pretrained on LUNA16 [13] using a combination of volume appearance recovery and sub-volume order prediction tasks. The results show that predictions based on 3D sub-volume order consistently outperform those based on 1D sub-volume order in all downstream tasks, highlighting the importance of 3D sub-volume order presentation for improved model performance.

4 Conclusion and Feature Work

We have developed a novel SSL method, ASA, which leverages the symmetry and recurrent attributes inherent in medical images to acquire robust global representation, intra-volume relationships, and detailed appearance features. ASA employs a student-teacher network to alternately learn from diverse learning perspectives. Extensive experiments have demonstrated the effectiveness and efficiency of ASA. Despite its excellent performance in segmenting CT images, the training process for ASA is relatively intricate and relies on consistent and recurrent anatomical structures due to the proxy tasks involving positional encoding predictions. Our future work will focus on simplifying the pretraining process, removing the need for consistent anatomical patterns, and expanding ASA to cover a broader range of tasks, including detection and registration. We hope ASA inspires the development of enhanced SSL models with a deep understanding of anatomical structures and their spatial relationships, thereby improving diagnostic accuracy and facilitating advanced medical imaging applications.

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

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: Simclr: A simple framework for contrastive learning of visual representations. In: International Conference on Learning Representations. vol. 2 (2020)
2. 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)
3. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2021)

4. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
5. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
6. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)
7. Jing, L., Tian, Y.: Self-supervised visual feature learning with deep neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(11), 4037–4058 (2020)
8. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21152–21164 (2023)
9. Ma, D., Pang, J., Gotway, M.B., Liang, J.: Foundation ark: Accruing and reusing knowledge for superior and robust performance. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 651–662. Springer (2023)
10. Pang, J., Haghighi, F., Ma, D., Islam, N.U., Hosseinzadeh Taher, M.R., Gotway, M.B., Liang, J.: Popar: Patch order prediction and appearance recovery for self-supervised medical image analysis. In: MICCAI Workshop on Domain Adaptation and Representation Transfer. pp. 77–87. Springer (2022)
11. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 319–345. Springer (2020)
12. 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)
13. Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., Van Den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis **42**, 1–13 (2017)
14. Tang, Y., Gao, R., Lee, H.H., Han, S., Chen, Y., Gao, D., Nath, V., Bermudez, C., Savona, M.R., Abramson, R.G., et al.: High-resolution 3d abdominal segmentation with random patch network fusion. Medical image analysis **69**, 101894 (2021)
15. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740 (2022)
16. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
17. Wang, H., Fan, J., Wang, Y., Song, K., Wang, T., Zhang, Z.: Droppos: Pre-training vision transformers by reconstructing dropped positions. arXiv preprint arXiv:2309.03576 (2023)

18. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 109–119. Springer (2021)
19. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: International Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
20. Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y.: nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201 (2021)
21. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical image analysis **67**, 101840 (2021)