Supplementary Material
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

A Pretraining Details

ASA: We independently pretrained ASA on the AMOS2022 dataset for the major evaluations and on the LUNA16 dataset for the ablation study. Both models follow the same pretraining protocol. In phase 1, the volumes are resized to $1\times128\times128\times128$, with the sub-volume size $16 \times 16 \times 16$, leading to 512 unique shuffleable sub-volumes and coordinates. The volume in phase 2 is up-sampled to $160 \times 160 \times 160$ before two spatially-related crops sized $1\times128\times128\times128$ are obtained. The Swin UNETR architecture is employed as both the student and teacher networks. We followed the default network configuration with patch size = 2, window size =7, feature size = 48, layer depth = (2, 2, 2, 2), and attention head number = (3, 6, 12, 24). We optimize the student model using the SGD optimizer with a momentum of 0.9 and a cosine learning rate scheduler, starting with a base learning rate of 0.001 and a learning rate warmup period of 5 epochs. The training data, with a batch size of 12, is distributed across 4 Nvidia A100 GPUs, each with 80 GB of memory. The model undergoes pretraining for 400 epochs. A stop-gradient operator is applied to the teacher, which is updated using an iteration-wise EMA of the student parameters, starting with an initial momentum of 0.9.

SimMIM We pretrain the SimMIM baseline on AMOS2022 dataset, adhering to the official implementation and implementing the method in 3D on the Swin UNETR using its default configuration. Similar to ASA, the volumes are first resized to $1 \times 128 \times 128 \times 128$, with each sub-volume size of $16 \times 16 \times 16$ masked at a 50% ratio. The optimization and learning rate scheduling strategy are the same as those used in ASA pertaining.

Swin UNETR undergoes a pretraining phase involving three common self-supervised learning tasks on five publicly accessible CT datasets. We obtained the pretrained model from its official GitHub release. Given that only the encoder weights were available, we randomly initialized the decoder for all subsequent evaluations in downstream tasks.

Table S1 summarizes and compares the pretraining tasks, datasets, and the number of training samples between ASA and baseline pretrained models.

Table S1: A concise comparison of self-supervised pretraining tasks, datasets, and the number of training samples between ASA and baseline pretrained models.

Method Name	Pretrianing Tasks	Dataset (# Samples for pretraining)	Organs
ASA	Order prediction, appearance recovery,	AMOS2022 (#240)	Abdomen
	global&local consistency	LUNA16 (#843)	Lung
Swin UNETR	Rotation, masked sub-volume recovery, contrastive learning	LUNA16, LiDC, TCIA Covid19 HNSCC, TCIA Colon (#5,050)	Chest CT, head & neck cancer, abdomen and pelvis
SimMIM	Masked sub-volume recovery	AMOS2022 (Totally #240)	Abdomen

B Downstream Tasks

We fine-tune the ASA model and baseline models on diverse abdomen organ segmentation tasks. In the preprocessing step for all tasks, we first re-sample all scans to a uniform voxel space, (1.5 (2.0 for BTCV), 1.5, 1.5) for z, x, and y dimensions. Subsequently, we clip the intensity values within the range of -175 to 250 and normalize them to a scale between 0 and 1. Moreover, during training, we randomly sample $128 \times 128 \times 128$ voxels, incorporating spatial padding if any dimension is smaller than the specified input size. Data augmentation techniques, including random flips, rotations, and intensity shifts, are employed during training with probabilities of 0.1, 0.1, and 0.5, respectively. We fine-tune all tasks utilizing the AdamW optimizer with a learning rate of $1e^{-4}$. The training is performed on the Dice similarity coefficient loss for 30,000 iterations, employing a batch size of 1. The implementation of all downstream tasks is carried out using PyTorch[1] and MONAI[2] and is run on a single NVIDIA A100 GPU. In each experiment, we perform five independent runs and present the average Dice score as the metric for evaluating the experiment results. For linear probing, we initialize the pretrained model's weights and then freeze the backbone while allowing the decoder to undergo fine-tuning.

BTCV: The Beyond the Cranial Vault (BTCV) dataset comprises CT scans from 30 patients, with each scan accompanied by 14 manual segmentation annotations. These annotations consist of 1 background and 13 different organs. Following the approach in [6,2], we establish a split of 24 samples for training and

[1] https://pytorch.org/

[2] https://monai.io/

6 samples for testing. We formulated a 14-class segmentation task, encompassing segmenting background, spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein, splenic vein, pancreas, right adrenal gland, and left adrenal gland.

Pancreas-CT: The Pancreas-CT dataset [5] comprises 80 abdominal contrast-enhanced 3D CT scans from 53 male and 27 female subjects, each scan paired with 2 manual segmentation annotations for background and pancreas. Following the protocol outlined in [4], a split of 64 samples for training and 16 samples for testing is established. We have formulated a 2-class segmentation task, including the background and pancreas classes.

LiTS: The Liver Tumor Segmentation Benchmark (LiTS) [1] comprises 130 CT scans, each paired with 3 manual segmentation annotations. These annotations include 1 for the background, 1 for the liver organ, and 1 for the liver tumor. Following the methodology outlined in[4], we establish a split of 94 samples for training and 36 samples for testing. The segmentation task is structured as a three-class problem, encompassing the background, liver, and liver tumor.

AMOS2022: The Multi-Modality Abdominal Multi-Organ Segmentation Challenge (AMOS2022) [3] consists of 360 CT scans, each scan paired with voxel-level annotations for 15 abdominal organs and 1 background class. Adhering to the official training split of 240 samples and 120 samples for testing, we formulated a 16-class segmentation task. This task includes segmenting the background, spleen, right kidney, left kidney, gall bladder, esophagus, liver, stomach, aorta, postcava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, and prostate/uterus.

Below is a brief summary of the aforementioned downstream tasks, including the number of training/testing samples and the target organs.

Table S2: A brief information of downstream tasks, number of training/ testing samples, and target organs.

Tasks	Train/ Test	Organs (Excluding Background)
Beyond the Cranial Vault (BTCV)	24/ 6	Spleen, Right kidney, Left kidney, Gallbladder, Esophagus, Liver, Stomach, Aorta, Inferior vena cava, Portal and splenic veins Pancreas, Left and right adrenal glands.
Pancreas-CT (TCIA)	64/ 16	Pancreas
Liver Tumor Segmentation Benchmark (LiTS)	96/ 36	Liver, liver Tumor
Multi-Modality Abdominal Multi-Organ Segmentation Challenge (AMOS2022)	240/ 120	spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate/uterus

C Additional Experiment Results and Ablation Study

Table S3: Supplementary to Table 1 with standard deviation measurements, ASA achieves the highest average dice score in segmenting all organs compared to SoTA self-supervised methods. Specifically, ASA outperforms competitors in segmenting 9 out of 12 organs.

Methods/ Organs[‡]	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Vins	Pan	AG	Avg.
SimMIM	92.03±0.93	93.66±0.67	92.13±1.32	69.16±2.22	75.06±2.07	96.21±0.21	76.36±1.23	89.80±0.39	83.91±0.87	72.46±1.08	73.61±1.45	68.24±0.71	81.89±1.10
Swin UNETR	95.30±0.35	**94.29±0.26**	**94.22±0.30**	74.01±3.68	76.35±1.23	96.71±0.08	80.56±1.44	90.42±0.27	84.70±0.54	**75.12±0.33**	80.61±0.99	67.25±1.00	84.13±0.87
ASA	**96.89±0.85**	94.28±0.15	94.10±0.06	**75.53±1.35**	**76.66±0.84**	**96.79±0.13**	**82.42±1.97**	**92.03±2.22**	**86.02±0.70**	74.77±0.59	**80.98±1.39**	**70.73±0.34**	**85.10±0.88**

[‡] 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.

D Pseudocode for ASA's Alternate Pretraining

```
Algorithm 1 One round of ASA alternate pretraining
Data: Input patient volumes: $\mathcal{S} = \{S_1, S_2, ..., S_M\}$, $S_i \in \mathbb{R}^{C \times D \times H \times W}$
Functions: Data augmentation: $\mathcal{F}_{perm}(\cdot)$, $\mathcal{F}_{src}(\cdot)$, Local sub-volume matching: $\mathcal{F}_M(\cdot)$;
           Order prediction loss: $\mathcal{L}_{vop}$; Appearance recovery loss: $\mathcal{L}_{var}$; Global con-
           sistency loss: $\mathcal{L}^{global}_{\theta_s,\theta_t}$; Local consistency loss: $\mathcal{L}^{local}_{\theta_s,\theta_t}$ ; Loss update by SGD
           optimizer: $Update_{sgd}(\cdot,\cdot)$
Trainable Parameters (Randomly Initialized): Student's encoder and decoder:
           $g^{enco}_{\theta_s}(\cdot)$, $g^{deco}_{\theta_s}(\cdot)$
Stop Gradient: Teacher's encoder and decoder: $g^{enco}_{\theta_t}(\cdot)$, $g^{deco}_{\theta_t}(\cdot)$
Hyperparameters: EMA Momentum: $\kappa$; Loss regularization parameter: $\lambda_{vop}$, $\lambda_{var}$,
           $\lambda_{global}$, $\lambda_{local}$
$\{g^{enco}_{\theta_t}, g^{deco}_{\theta_t}\} \leftarrow \{g^{enco}_{\theta_s}, g^{deco}_{\theta_s}\}$ // initialize teacher with student's parameters
/* train student for one epoch on the sub-volume order prediction and
   volume appearance recovery task                                      */
for $S_i$ in $S_1, S_2, ..., S_M$ do
    $S^{perm}_i, \mathcal{C}^{perm}_i = \mathcal{F}_{perm}(S_i)$;

    $\mathcal{P}^{vo}_i, \overline{y_s} = g^{enco}_{\theta_s}(S^{perm}_i)$,;
    $\_, \overline{y_t} = g^{enco}_{\theta_t}(S_i)$;
    $\mathcal{P}^{va}_i = g^{deco}_{\theta_s}(g^{enco}_{\theta_s}(S^{perm}_i))$

    $Loss = \lambda_{vop} * \mathcal{L}_{vop}(\mathcal{P}^{vo}_i, \mathcal{C}^{perm}_i) + \lambda_{var} * \mathcal{L}_{var}(\mathcal{P}^{va}_i, S_i) + \lambda_{global} * \mathcal{L}_{global}(\overline{y_t}, \overline{y_s})$ ;

    $Update(\{g^{enco}_{\theta_s}, g^{deco}_{\theta_s}\}, Loss)$;

    $\{g^{enco}_{\theta_t}, g^{deco}_{\theta_t}\} \leftarrow \kappa\{g^{enco}_{\theta_t}, g^{deco}_{\theta_t}\} + (1-\kappa)\{g^{enco}_{\theta_s}, g^{deco}_{\theta_s}\}$;
end
/* train student for one epoch on learning global consistency          */
for $S_i$ in $S_1, S_2, ..., S_M$ do
    $Crop_1, Crop_2 = \mathcal{F}_{src}(S_i)$;

    $\overline{y_s}, \overline{y_t}, y_s, y_t = g^{enco}_{\theta_s}(Crop_1), g^{enco}_{\theta_t}(Crop_2)$;

    $Loss = \mathcal{L}^{global}_{\theta_s,\theta_t}(\overline{y_s}, \overline{y_t}) + \mathcal{L}^{local}_{\theta_s,\theta_t}(\mathcal{F}_M(y_s, y_t))$ ;

    $Update(g^{enco}_{\theta_s}, Loss)$;

    $\{g^{enco}_{\theta_t}\} \leftarrow \kappa\{g^{enco}_{\theta_t}\} + (1-\kappa)\{g^{enco}_{\theta_s}\}$;
end
```

E Acknowledgements

This research has been supported in part by ASU and Mayo Clinic through a Seed Grant and an Innovation Grant, and in part by the NIH under Award Number R01HL128785. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work has utilized the GPUs provided in part by the ASU Research Computing and in part by the Bridges-2 at Pittsburgh Supercomputing Center through allocation BCS190015 and the Anvil at Purdue University through allocation MED220025 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The content of this paper is covered by patents pending.

References

1. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
2. 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)
3. 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)
4. 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)
5. Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B., Summers, R.M.: Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18. pp. 556–564. Springer (2015)
6. 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)