Anatomical Embedding-Based Training Method for Medical Image Segmentation Foundation Models

Mingrui Zhuang[1[0000-0002-2991-0919]], Rui Xu[1], Qinhe Zhang[2], Ailian Liu[2], Xin Fan[1] and Hongkai Wang*[1,3[0000-0002-1813-2162]]

[1] Dalian University of Technology, Dalian, China
wang.hongkai@dlut.edu.cn
[2] The First Affiliated Hospital of Dalian Medical University, Dalian, China
[3] Liaoning Key Laboratory of Integrated Circuit and Biomedical Electronic System

Abstract. Existing training methods for medical image foundation models primarily focus on tasks such as image restoration, overlooking the potential of harnessing the inherent anatomical knowledge of the human body. The discrepancy between the training tasks of foundation models and downstream tasks often necessitates model fine-tuning for each specific application. An insufficient scale of the downstream training set can lead to catastrophic forgetting of the foundational model. To address these issues, we propose a novel unsupervised training method for medical image foundation models. Our approach incorporates an anatomical embedding task, enabling the model to generate anatomically related embeddings for each voxel. To expedite the training and accommodate large-scale models, we employ the strategy of momentum contrast learning, which is further enhanced to adapt to the task of anatomical embedding. To improve the model's performance for specific targets, we introduce the region contrastive loss, utilizing a small set of segmentation labels (e.g., five samples) to identify the focused regions during training. In our experiments, we pre-train the foundation model using a dataset of 4000 unlabeled abdominal CT scans with the downstream task being the few-shot learning of 13 abdominal organ segmentation. The results showed significant improvements in the downstream segmentation task, particularly in the scenarios with limited segmentation annotations, compared to methods without pre-training and similar foundation models. The trained models and the downstream training code have been open sourced at https://github.com/DlutMedimgGroup/Anatomy-Embedding-Foundation-Model.

Keywords: Foundation Model, Pre-Trained Model, Anatomical Embedding Learning.

1 Introduction

Foundation models are commonly pretrained on extensive datasets to reduce the reliance on training data scale for specific downstream tasks. However, creating comprehensive medical image datasets for foundation training faces unique challenges, primarily due to patient privacy protection restrictions and the demanding workload of data annotation. Several methods have emerged to transfer models pretrained on large-scale natural image datasets to the domain of medical imaging [8, 15, 18]. Despite the potential performance enhancements achieved by such methods, they lack training on native medical data and fail to utilize the inherent anatomical knowledge present in the data.

Most existing foundation models employ the approach of fine-tuning the decoder when applied to downstream tasks [3, 7, 13]. In this process, fine-tuning the original parameters of the foundation model with a small training sample can potentially lead to forgetting the pretrained knowledge. Even with the commonly used encoder freezing strategy, fully retraining the decoder still faces overfitting and imperfect robustness issues when dealing with small training samples.

Knowledge of anatomical relationships is crucial for improving the performance of medical foundation models [16]. Most existing methods train foundational models using tasks that are unrelated to anatomical information, such as image restoration. Anatomical embedding involves generating a feature vector for each pixel in the image based solely on its anatomical position (e.g., lower liver, middle left of the kidney, etc.). The anatomical embedding task holds tremendous potential for training foundational models. The training of the anatomical embedding task employs unsupervised contrastive learning, taking advantage of its unsupervised nature to reduce the reliance on data annotations in medical image processing tasks[11, 17]. The encoded knowledge obtained from the trained foundation model can be directly used as inputs for downstream tasks, eliminating the need for fine-tuning the decoder for individual tasks.

This study introduces an unsupervised training method for medical segmentation foundation models. The method utilizes an unsupervised anatomical position encoding task, utilizing a substantial volume of unlabeled data for training. For specific downstream tasks, the features from the trained foundation model can be directly inherited by downstream models, eliminating the need for specific fine-tuning of the decoder. This strategy mitigates the performance degradation typically associated with fine-tuning on small datasets. Different foundation models can be specifically trained for mainstream imaging examinations (e.g., abdominal CT, brain MRI, etc.), thereby enhancing the performance of various downstream tasks tailored to the corresponding image types. Our method offers the following innovative contributions:

(a) By employing the anatomical embedding task, we train the medical image foundation model to effectively utilize the inherent anatomical knowledge of the human body. The anatomical embedding can be directly utilized for downstream tasks, thereby avoiding the performance loss associated with model fine-tuning.

(b) In order to enable the utilization of larger model sizes and higher image resolutions, we employed the momentum contrast learning process strategy.

(c) To enhance the discriminative capability of the model's anatomical embedding features for important soft tissue organs, a region contrastive loss (RCL) designed for momentum-contrastive learning is proposed.

2 Methods

The proposed anatomical embedding-based framework of the foundation model training is shown in **Fig. 1**. Initially, contrastive learning with an anatomical encoding task is employed to train the foundation model M_F. $M_F{}'$ has the same structure and initialization as M_F, and its parameters are updated from M_F in the form of momentum [6]. During training, two patches with overlapping regions are randomly selected, as indicated by the yellow boxes in **Fig. 1**. The networks accept the patch as the input and outputs the anatomical embedding maps, where each pixel is a 64-dimensional feature vector. Finally, anatomical contrastive loss is used to supervise that the pixels with same anatomical positions in both patches (indicated by the red dots) have same feature embeddings, while different positions have distinct feature embeddings.

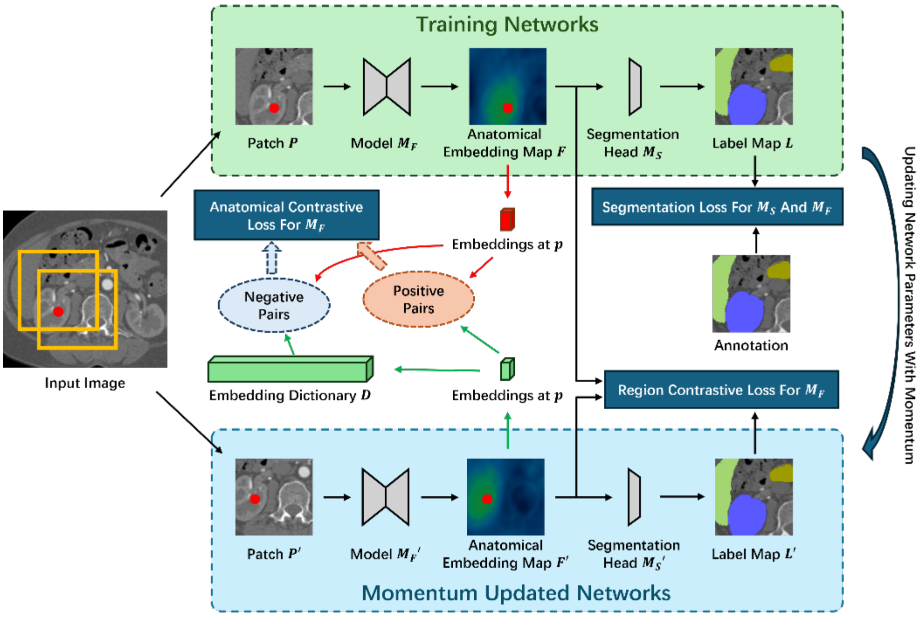

Training Networks
Patch P
Model M_F
Anatomical Embedding Map F
Segmentation Head M_S
Label Map L
Input Image
Anatomical Contrastive Loss For M_F
Embeddings at p
Segmentation Loss For M_S And M_F
Negative Pairs
Positive Pairs
Annotation
Embedding Dictionary D
Embeddings at p
Region Contrastive Loss For M_F
Patch P'
Model $M_F{}'$
Anatomical Embedding Map F'
Segmentation Head $M_S{}'$
Label Map L'
Momentum Updated Networks
Updating Network Parameters With Momentum

Fig. 1. Overview of the proposed framework.

To enhance the discriminative ability of the model towards important soft tissue organs with similar grayscale values, we propose a method using a limited number of labels (e.g., five samples) to guide the network's attention. Specifically, a lightweight fully connected segmentation head M_S is added after the foundation model. For the small amount of annotated data, a standard segmentation loss is employed to supervise the network. As the segmentation head is lightweight, this supervision primarily impacts the foundation model M_F. For the majority of unlabeled data, the result of momentum-updated segmentation head $M_S{}'$ is utilized for region contrastive loss. This loss function ensures that embeddings of the same region are close while the embeddings of different regions far apart.

2.1 Momentum Contrastive Learning for Large Training Dataset

When selecting patches with overlapping regions, we introduce randomness to enhance the diversity of the training data. The sizes of the two patches are determined randomly, with each side's length ranging from 0.9 to 1.1 times the preset value. The overlapping range in each direction is set as a fraction of the smaller side length between the two patches, varying from 0.5 to 1.0 times. Within the overlapping region, n_{pos} points are randomly selected as positive sample pairs for training, and this collection is denoted as $P = \{p_1, p_2, \dots, p_{n_{pos}}\}$. Finally, both patches are resampled to the predetermined size and individually input into M_F and $M_F{}'$.

The foundation model, M_F, is based on a 3D U-Net architecture. In this study, we adopt a five-layer network as an example, with each layer comprising two residual units. Unlike a typical U-Net, the decoder part of M_F differs in that the dimensionality reduction stops once the feature dimension reaches 64. Consequently, M_F produces anatomical embedding maps, F, containing 64 channels. $M_F{}'$ shares the same structure and training initialization as M_F. During training, M_F is guided by multiple loss functions, and its parameters are updated through backpropagation. Meanwhile, the parameters of $M_F{}'$ are updated using a momentum-based approach to gradually align with those of M_F. To ensure stability, a relatively larger momentum coefficient (i.e., 0.999) is commonly employed during training for $M_F{}'$.

We extract features from F and F' at the positions specified by p_i, denoting them as e_i and $e_i{}'$, respectively. Their collections are denoted as $E = \{e_1, e_2, \dots, e_{n_{pos}}\}$ and $E' = \{e_1{}', e_2{}', \dots, e_{n_{pos}}{}'\}$. e_i and $e_i{}'$ correspond to the same anatomical position p_i. The training goal is to minimize the differences between them by treating them as positive sample pairs in contrastive learning. However, selecting a large number of features as negative sample pairs in real-time can significantly decrease the training speed. To overcome this challenge, we adopt an approach inspired by MoCo [6]. Specifically, we incorporate E' into a feature dictionary, D, which serves as a source of negative sample points for subsequent training. The positions of previously generated features in D are considered different from the current point. E and D are treated as negative sample pairs. As a result, the anatomical contrastive loss can be represented as

$$L_{AC} = -\log \frac{\exp\left(\sum_{i=1}^{n_{pos}} e_i \cdot e_i' \Big/ \tau\right)}{\exp\left(\sum_{i=1}^{n_{pos}} \sum_{j=1}^{n_{neg}} e_i \cdot d_j \Big/ \tau\right) + \exp\left(\sum_{i=1}^{n_{pos}} e_i \cdot e_i' \Big/ \tau\right)} \tag{1}$$

where n_{neg} denotes the size of D and τ is a temperature hyper-parameter. Due to the normalization operation applied to the output of M_F, $\|e_i\| = 1$. As a result, $e_i \cdot e_i{}'$ can be directly interpreted as the cosine similarity between them.

2.2 Guiding Network Attention Using A Few Annotations

To enhance the discriminative ability of the model towards important soft tissue organs with similar grayscale values, we propose a region supervision loss to fine-tune the network for downstream tasks. In our approach, we introduce a lightweight segmentation head, denoted as M_S, which consists of two fully connected layers. M_S is appended to the end of M_F, facilitating the conversion of anatomical embedding maps F into segmentation results L. For the small amount of annotated data (i.e., five samples), a common segmentation loss function, denoted as L_S, is employed. Notably, due to the lightweight nature of M_S, L_S significantly contributes to supervising M_F.

Similar to $M_F{}'$, $M_S{}'$ is a momentum-updated network based on M_S. Due to the larger momentum coefficient, the output of $M_S{}'$, represented by L', exhibits greater stability during the training process compared to L. L' is employed to supervise the network M_F in distinguishing the feature outputs between regions of interest. Inspired by the local contrastive loss [2], we propose a region contrastive loss designed for momentum-contrastive learning to supervise the similarity of features within the same region in F, while encouraging larger dissimilarity between features from different regions. This loss function, denoted as L_{RC}, can be represented as

$$L_{RC} = -\log \frac{1+\exp\left(\sum_{m=1}^{C} \overline{e_m} \cdot \overline{e_m{}'}\right)}{1+\exp\left(\sum_{m=1}^{C} \overline{e_m} \cdot \overline{e_m{}'}\right)+\exp\left(\sum_{m=1}^{C} \sum_{n=1,n\neq m}^{C} \overline{e_m} \cdot \overline{e_n{}'}\right)} \quad (2)$$

where C represents the number of segmentation categories. $\overline{e_m}$ and $\overline{e_m{}'}$ denote the averages of embeddings belonging to segmentation label m in the map E and E', respectively.

3 Experiments and Results

Datasets. To ensure the training effectiveness of the foundation model, we utilized the FLARE23 dataset [9], which comprises more than 4250 abdominal CT images, of which 4000 randomly selected data samples were used for unsupervised training of the foundation mode. To evaluate the performance of the foundation model, we employed the rest 250 samples with 13 abdominal organ labels to train a downstream segmentation network. Among these, 200 samples were allocated for training, while the remaining 50 samples were set aside for testing. Prior to training, all data underwent standardized preprocessing steps, including gray-scale clipping within the range of [-500, 750], gray-scale normalization, and resampling to an image resolution of 1×1×1mm^3.

Implementation and Evaluation Criteria. The proposed framework is implemented on MONAI[1] and PyTorch. To accommodate the large model scale, a deep learning server equipped with 8 NVIDIA A800 GPUs was employed for training the foundation model. We empirically set $n_{pos} = 1000$ and $n_{neg} = 100000$ to strike a balance between training speed and resource utilization. During training, patch pairs with overlapping regions are alternately fed into M_F and $M_F{}'$, and the mean of the two losses is used as the final loss. In the first 500 epochs of training, only L_C is utilized, and in the subsequent 200 epochs, the segmentation head and the region loss function are introduced.

Foundation Model Test. To validate if the foundation model has learned the anatomical embedding from the training images, a test of inter-subject anatomical correspondence was conducted, as shown in **Fig. 2**. The template image is fed into the network to obtain anatomical embeddings for specific points. The predicted position of the prediction image is determined by selecting the anatomical embedding point with the highest cosine similarity to the input point. The examples shown in **Fig. 2** represent scenarios where the target points are located both on the boundaries and within the organs. Notably, despite the anatomical structural differences between the template and predicted images, the network of this method succeeds in accurately identifying the correct anatomical correspondences. The improved correspondence accuracy can also be attributed to the introduction of momentum contrastive learning, enabling training of the foundation model at a spatial resolution of 1×1×1mm^3.

It is worth mentioning that the proposed training method may possess potential advantages in the context of swarm learning[12, 14]. According to formula (1), the computation of L_C depends solely on e_i, e_i', and D, with D derived from e_i'. By sharing e_i and e_i' among the swarm learning clients, it becomes feasible to perform synchronous training of the model across different centers under the same initialization conditions. By eliminating the need to transmit model parameters, the risk of model leakage is reduced. To guarantee data security, the transmitted anatomical embeddings are randomly sampled from discrete positions in the images and then shuffled, making it impossible to reconstruct the original data. This approach enables each client to leverage the entire dataset while maintaining data security. This eliminates the need for model aggregation algorithms (e.g., FedAvg[10]) and the associated performance losses.

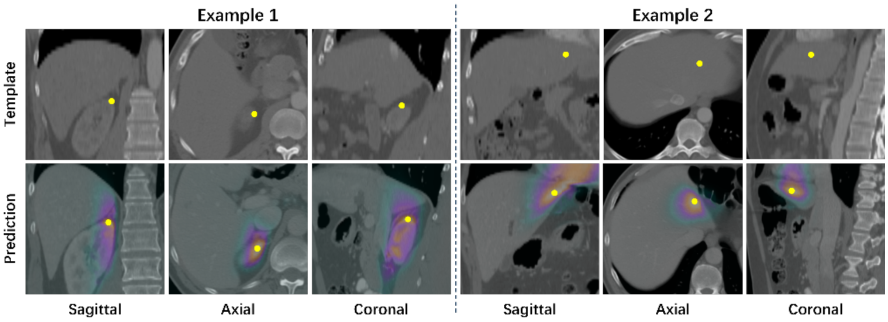
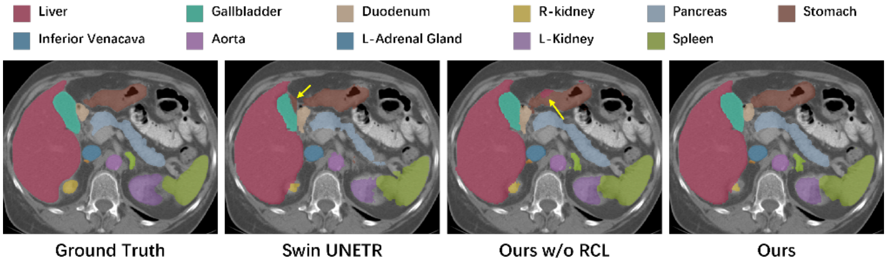


Fig. 2. Point-wise anatomical correspondence test results of the foundation model. The yellow dots indicate the input and output positions. The heatmap in the prediction image represents the similarity of features to the input points.

Downstream Segmentation Task. To evaluate the effectiveness of the trained foundation model in downstream tasks, we conducted validation using a segmentation task focused on 13 major abdominal organs. The downstream segmentation network is based on a conventional U-Net architecture, where the anatomical embeddings generated by the trained foundation model are combined with the original image as input. As control methods, we also employed the U-Net and the UNETR [5] without anatomical embedding input. To further validate the performance of the foundation model, we employed the Swin UNETR method [4], initialized with a self-supervised pre-trained Swin Transformer backbone [13], as an additional control. Furthermore, we conducted an ablation experiment where we excluded the proposed region contrast loss (RCL).

Fig. 3 showcases a qualitative comparison of the segmentation results obtained from three pre-trained methods when trained with 20 data examples. The results demonstrate that our method exhibits the least number of flaws in the segmentation results. **Fig. 4** presents a comparison of the results obtained from each method when trained with 20 data examples. An additional set of 50 data samples was used as the test dataset. The networks utilizing anatomical embedding input outperformed non-pretraining methods across all segmentation targets. Notably, our method outperforms the pre-trained Swin UNETR in terms of segmentation results for the majority of targets. Furthermore, the incorporation of GLC (Global Local Contrast) further enhanced the segmentation accuracy for the majority of targets, confirming its effectiveness.



Ground Truth Swin UNETR Ours w/o RCL Ours

Fig. 3. Qualitative demonstration of segmentation results using different methods. The yellow arrows indicate noticeable defects.

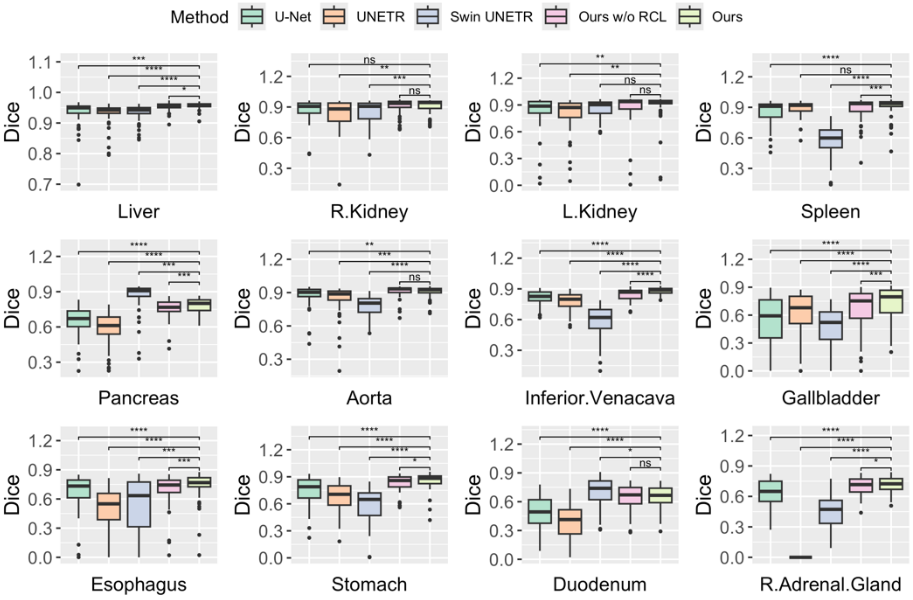


Fig. 4. Quantitative comparisons of segmentation performance using a training dataset consisting of 20 volumes. * indicates p < .05; ** indicates p < .01; *** indicates p < .001; **** indicates p < .001; **** indicates p < .0001.

Comparison with existing foundation models. To compare our method with existing medical foundation models, we conducted a comprehensive comparison with MedSAM[8]. MedSAM is officially pre-trained on the Flare22 dataset (similar to the Flare23 dataset we used) using much more computing resources (20 A100 GPUs), requiring an additional text prompt indicating the presence of each organ in each 2D slice. Our model yielded higher Dice score (0.84±0.09) than MedSAM (0.74±0.17).

The robustness of downstream tasks. The foundation model also exhibits superior robustness than the compared method against small training sets and variations in input data grayscale. **Fig. 5** (a) presents the performance of each method under different training set sizes. The results reveal that methods pre-trained with the foundation model exhibit a significant performance advantage when dealing with small training sample sizes. The reason for the comparatively poor performance of the two Transformer-based methods may be attributed to the larger model size, which makes them more susceptible to the limited training dataset. **Fig. 5** (b) compares the robustness of each method to grayscale transformations. By default, the range for grayscale clipping and grayscale normalization is [-500, 750]. During grayscale transformation, this range is proportionally scaled down. The degree of range transformation is depicted on the horizontal axis of **Fig. 5** (b). The results indicate that pre-trained models exhibit robustness for downstream segmentation networks, and our proposed method demonstrates minimal degradation in performance when the transformation degree is kept below 48%.

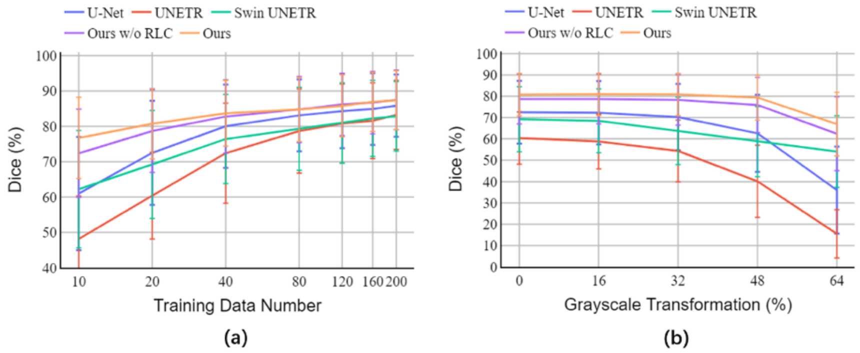


Fig. 5. (a) Impact of training set size on segmentation results. (b) Robustness comparison. The horizontal axis represents the degree of grayscale normalization range transformation of the input data.

4 Conclusion

In this paper, we propose a novel training framework for medical image foundation models. Our approach leverages the anatomical embedding task to train the foundational model, eliminating the need for downstream fine-tuning and mitigating the associated performance degradation. By incorporating the proposed momentum contrastive learning method and region contrastive loss, we have successfully improved training efficiency and enhanced the model's ability to discern soft tissues. Through experiments, we demonstrate the effectiveness of our approach through the foundation model and downstream segmentation model results.

Acknowledgments. This work was supported in part by the National Key Research and Development Program No. 2020YFB1711500, 2020YFB1711501 and 2020YFB1711503, the general program of National Natural Science Fund of China (No. 81971693, 61971445), the funding of Dalian Key Laboratory of Digital Medicine for Critical Diseases，the Fundamental Research Funds for the Central Universities (No. DUT22YG229 and DUT22YG205), the funding of Liaoning Key Lab of IC & BME System and Dalian Engineering Research Center for Artificial Intelligence in Medical Imaging.

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

References

1. Cardoso, M.J. et al.: MONAI: An open-source framework for deep learning in healthcare, http://arxiv.org/abs/2211.02701, (2022). https://doi.org/10.48550/arXiv.2211.02701.
2. Chaitanya, K. et al.: Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Medical Image Analysis. 87, 102792 (2023). https://doi.org/10.1016/j.media.2023.102792.
3. Dosovitskiy, A. et al.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, http://arxiv.org/abs/2010.11929, (2021). https://doi.org/10.48550/arXiv.2010.11929.
4. Hatamizadeh, A. et al.: Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In: Crimi, A. and Bakas, S. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 272–284 Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-08999-2_22.
5. Hatamizadeh, A. et al.: UNETR: Transformers for 3D Medical Image Segmentation, http://arxiv.org/abs/2103.10504, (2021). https://doi.org/10.48550/arXiv.2103.10504.
6. He, K. et al.: Momentum Contrast for Unsupervised Visual Representation Learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9726–9735 IEEE, Seattle, WA, USA (2020). https://doi.org/10.1109/CVPR42600.2020.00975.
7. Liu, Z. et al.: Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision (2021).
8. Ma, J. et al.: Segment anything in medical images. Nat Commun. 15, 1, 654 (2024). https://doi.org/10.1038/s41467-024-44824-z.
9. Ma, J. et al.: Unleashing the Strengths of Unlabeled Data in Pan-cancer Abdominal Organ Quantification: the FLARE22 Challenge, http://arxiv.org/abs/2308.05862, (2023). https://doi.org/10.48550/arXiv.2308.05862.
10. McMahan, B. et al.: Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. pp. 1273–1282 PMLR (2017).
11. Park, T. et al.: Contrastive Learning for Unpaired Image-to-Image Translation, http://arxiv.org/abs/2007.15651, (2020). https://doi.org/10.48550/arXiv.2007.15651.
12. Saldanha, O.L. et al.: Swarm learning for decentralized artificial intelligence in cancer histopathology. Nat Med. 28, 6, 1232–1239 (2022). https://doi.org/10.1038/s41591-022-01768-5.

13. Tang, Y. et al.: Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022).
14. Warnat-Herresthal, S. et al.: Swarm Learning for decentralized and confidential clinical machine learning. Nature. 594, 7862, 265–270 (2021). https://doi.org/10.1038/s41586-021-03583-3.
15. Wu, J. et al.: Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation, http://arxiv.org/abs/2304.12620, (2023).
16. Yan, K. et al.: SAM: Self-Supervised Learning of Pixel-Wise Anatomical Embeddings in Radiological Images. IEEE Transactions on Medical Imaging. 41, 10, 2658–2669 (2022). https://doi.org/10.1109/TMI.2022.3169003.
17. Yu, Z. et al.: Cross-grained Contrastive Representation for Unsupervised Lesion Segmentation in Medical Images. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 2339–2346 (2023). https://doi.org/10.1109/ICCVW60793.2023.00248.
18. Zhang, Y. et al.: Input Augmentation with SAM: Boosting Medical Image Segmentation with Segmentation Foundation Model. In: Celebi, M.E. et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops. pp. 129–139 Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-47401-9_13.