Non-Salient Object Segmentation in Medical Images via Pre-trained Multi-Granularity Masked Autoencoders

Bin Zhang[1†], Dongsheng Ruan[2†], Ronghui Qi[1], Chenchu Xu[1], Yanping Zhang[1], Chengjin Yu[3⋆], Lei Xu[4], and Rui Wang[4]

[1] School of Computer Science and Technology, Anhui University, Hefei, China
[2] School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou, China
[3] School of Big Data and Statistics, Anhui University, Hefei, China
[4] Department of Radiology, Beijing AnZhen Hospital, Beijing, China

Abstract. The segmentation of non-salient objects in medical images plays a crucial role in the early detection and diagnosis of diseases. However, due to the low contrast and unbalanced distribution of the non-salient objects, their feature extraction still suffers from dimensional collapse. To address the inherent feature representation challenges of non-salient objects, we propose a pre-trained **M**ulti-**G**ranularity **M**asked **A**uto**E**ncoder (MG-MAE) framework with diversified feature learning capabilities. In the global level, masked image reconstruction captures holistic structural and contextual features. Subsequently, in the local level, patches are extracted from the global visible patches, and the Histogram of Oriented Gradient (HOG) features of these patches are then reconstructed to enhance the texture details. Based on local perception, the framework integrates Nuclear Norm Maximization (NNM) constraint to foster diversity of the local representations in the feature encoding process. In the HOG reconstruction process, the framework also adopts a Dynamic Weight Adjustment (DWA) strategy, assigning greater reconstruction weights to challenging image patches, thereby solving the problem of representation bias towards salient objects. We evaluate our method on a private dataset, CCTA139, and two public datasets, BTCV and LiTS, respectively. Our method achieves DSC of 80.71%, 82.60%, and 71.77%, respectively, surpassing the performance of current state-of-the-art methods. The code is available at https://github.com/zhangbbin/mgmae.

Keywords: Non-salient object segmentation · Masked Autoencoders.

1 Introduction

In medical image analysis, the accurate identification of non-salient objects carries substantial clinical significance [1,14]. This clinical relevance manifests across

⋆ Corresponding author: Chengjin Yu (23073@ahu.edu.cn). †Bin Zhang and Dongsheng Ruan—Contributed equally to this work.

multiple domains: coronary artery stenosis detection enables timely intervention in coronary artery disease progression [9, 10], precise localization of small abdominal organs (e.g., gallbladder and pancreas) supports surgical planning and diagnostic accuracy [21], and early detection of sub-centimeter tumors (e.g., stage I lung/hepatic malignancies) critically determines therapeutic outcomes [13]. Although these lesions may occupy minimal volume and exhibit barely perceptible contrast from surrounding tissues, their segmentation holds critical importance for clinical diagnosis and therapeutic decision-making.

However, the segmentation of non-salient objects still suffers from dimensional collapse stemming from two interrelated challenges. First, the non-salient objects have inherently low contrast compared to adjacent tissues, making them easily overlooked in image feature extraction. Second, this perceptual ambiguity is further complicated by severe distribution imbalances, which manifest through two mechanisms: inter-class imbalance reflects the statistical rarity of certain anatomical categories (e.g., accessory spleen occurring in $<2\%$ of abdominal scans); intra-class imbalance arises from significant distribution differences within individual categories, including size disparities (larger tumors 2-3 cm, smaller tumors 0.50-1 cm), shape polymorphism, and contrast heterogeneity. These challenges cause dimensional collapse: the latent space degenerates into a low-rank manifold biased toward dominant features (e.g., high-contrast structures), suppressing the representation of non-salient patterns.

In this paper, we propose a pre-trained multi-granularity framework for segmentation of non-salient objects in medical images, aiming to enhance the diversity of non-salient object features. To address the dimensional collapse issue caused by MAE [25], we introduce a local branch for fine-grained detail perception, extending MAE to a multi-granularity space. Based on local perception, we introduce a dual-path optimization scheme at the local level, incorporating Nuclear Norm Maximization (NNM) constraint and Dynamic Weight Adjustment (DWA) strategy for enhanced feature diversity and prioritization of challenging patches. Specifically, we establish a Multi-Granularity Masked Autoencoder (MG-MAE) framework, which hierarchically integrates global level to local level feature learning. In the global level, the input image is globally masked, and the masked patches are reconstructed to generate a coarse latent feature representation that captures the holistic understanding of the image [12]. In the local level, we further extract visible global patches and reconstruct the HOG features of the masked local patches. NNM aims to improve the effective rank [20] of the feature matrix to promote diverse feature representations. By encouraging diversity, NNM prevents the collapse of feature representations into narrow manifolds [15,25]. DWA dynamically adjusts the HOG reconstruction weights assigned to different image patches based on their learning difficulty. This ensures a more comprehensive optimization process and aids in learning the fine-grained features of non-salient objects.

Our contributions are summarized as follows: 1) We introduce a two-level MG-MAE framework that captures both global semantic structures and fine-grained local details, enhancing the model's ability to the segmentation of non-

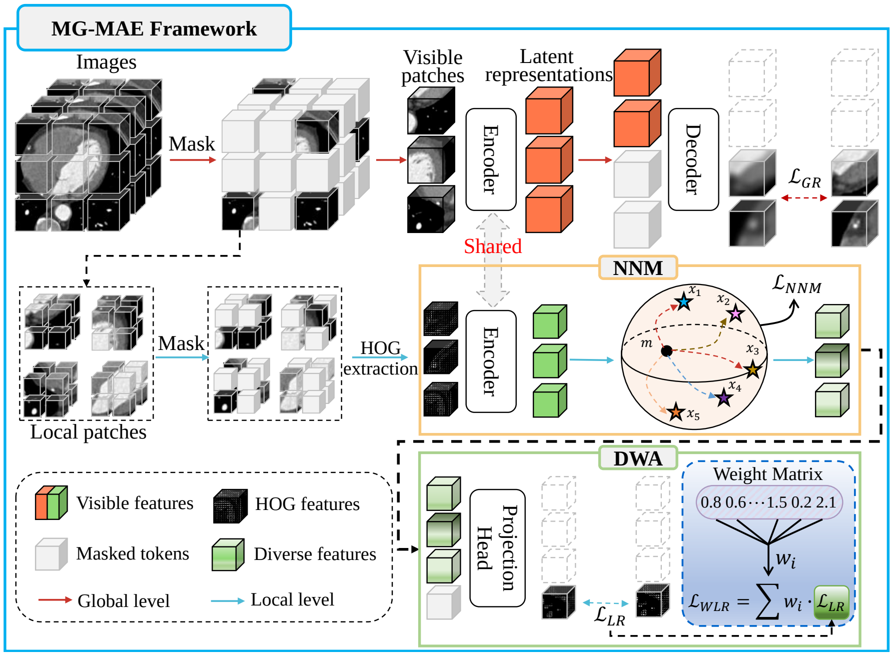

Fig. 1: The architecture of our proposed framework. MG-MAE combines global pixel reconstruction and local HOG feature refinement, optimized by NNM and DWA to prevent feature collapse and prioritize challenging patches.

salient objects. 2) We introduce a NNM constraint to enforce diverse feature learning, preventing the collapse of non-salient object features into narrow manifolds. 3) We introduce a DWA strategy to dynamically adjust the reconstruction weights based on the learning difficulty of each patch, placing higher emphasis on challenging non-salient objects.

2 Methodology

2.1 Preliminary: Masked Autoencoder

Masked Autoencoder (MAE) [12] is a scalable self-supervised vision learning method, where the core idea is to randomly mask a large portion of the input image (e.g., 75%) and reconstruct the missing pixels. The method employs an asymmetric encoder-decoder architecture: the encoder processes only the visible patches, while the lightweight decoder reconstructs the original image by combining the latent representations and the masked tokens. The high masking ratio reduces spatial redundancy, forcing the model to complete the task through a holistic understanding.

While MAE is powerful in representation learning, it lacks the ability to capture low-level image statistics [12], which leads to dimensional collapse when learning non-salient objects [25]. In this case, subtle and low-contrast objects tend to collapse into the background or salient objects. To enhance local perception of the non-salient objects, we further introduce a local branch specifically for the perception of fine-grained details in Section 2.2. Crucially, we employ a dual-path optimization. In the encoding phase, NNM is used to enhance the rank of the feature matrix and prevent collapse in Section 2.3, while the decoding phase incorporates DWA to prioritize challenging patches in Section 2.4.

2.2 Multi-Granularity Masked Autoencoders Framework

As shown in Fig. 1, MG-MAE is designed as a hierarchical framework, where each layer focuses on a distinct level of granularity. The global level of MG-MAE adopts the MAE [12] framework to reconstruct the masked patches of the input image. The global reconstruction loss can be defined as:

$$\mathcal{L}_{GR} = \frac{1}{|\mathcal{M}_{global}|} \sum_{i \in \mathcal{M}_{global}} (\tilde{X}_i - X_i)^2 \tag{1}$$

where $\mathcal{M}_{global}$ is the set of global masked indices, $\tilde{X}_i$ and X_i represent the reconstructed patches and the original patches.

In the local level, MG-MAE extracts visible patches from the global level and reconstructs HOG features [22] instead of raw pixels for the masked patches. HOG explicitly encodes gradient orientation distributions, which are critical for resolving subtle texture variations. This design aligns with the hierarchical MG-MAE: global pixel reconstruction preserves anatomical context, while local HOG refinement amplifies discriminative patterns essential for challenging patches. Similarly, the local patch p is masked at 50% to obtain p_v, which is then put into the encoder to obtain latent representation Z_{p_v}. The projection head $g_\phi(\cdot)$ [22] combine Z_{p_v} and masked tokens Z_{p_m} to reconstruct the missing HOG features:

$$\tilde{H}_{local} = g_\phi(Z_{p_v}, Z_{p_m}) \tag{2}$$

The reconstruction loss for the local HOG features can be defined as:

$$\mathcal{L}_{LR} = \frac{1}{|\mathcal{M}_{local}|} \sum_{i \in \mathcal{M}_{local}} (\tilde{H}_{local,i} - H_{local,i})^2 \tag{3}$$

where $\mathcal{M}_{local}$ is the set of local masked indices, $\tilde{H}_{local,i}$ and $H_{local,i}$ represent the reconstructed HOG features and the original features.

2.3 Nuclear Norm Maximization Constraint

To mitigate the risk of dimensional collapse in feature representations [25], where learned embeddings concentrate on into a low-rank manifold and fail to capture diverse patterns, we propose a nuclear norm maximization constraint that explicitly enforces feature diversity at the local encoder output. It has been theoretically verified that the nuclear norm is a convex relaxation of the matrix rank, which encourages the feature to maintain a high effective rank [5, 8, 20]. Specifically, this constraint operates on the latent feature matrix $Z_{p_v} \in \mathbb{R}^{m\times n}$, where m is the number of patches and n is the embedding dimension. By maximizing the nuclear norm $\|Z_{p_v}\|_*$, Z_{p_v} tends to maintain a high effective rank, ensuring that singular values are distributed broadly rather than collapsing to a few dominant dimensions. Then, the $\mathcal{L}_{NNM}$ can be defined as:

$$\mathcal{L}_{NNM} = -||Z_{p_v}||_* \tag{4}$$

2.4 Dynamic Weight Adjustment Strategy

To solve the problem of representation bias towards salient objects, we propose a theoretically grounded weight adjustment strategy that dynamically prioritizes patches based on their learning difficulty. The core idea is to quantify the gradient stability of each patch i. Then, we define its stability metric s_i as:

$$s_i = exp(-\lambda \cdot \frac{\left\|\nabla\mathcal{L}_i^t - \nabla\mathcal{L}_i^{t-1}\right\|_2}{\nabla\mathcal{L}_i^{t-1} + \varepsilon}) \tag{5}$$

where λ controls the sensitivity to gradient variations, and ε ensures numerical stability. A lower gradient variation (higher s_i) indicates persistent optimization difficulty, such as in low-contrast structures. The reconstruction weights w_i are then assigned through a entropy maximization weight allocation principle [16]:

$$w_i = \frac{s_i}{\sum_j s_j} \cdot log(1 + \frac{1}{s_i}) \tag{6}$$

This formula combines the information entropy principle with gradient stability, emphasizing difficult areas while preventing weight polarization through entropy constraints. Consequently, the dynamic weighted loss can be defined as:

$$\mathcal{L}_{WLR} = \sum_i w_i \cdot \mathcal{L}_{LR_i} \tag{7}$$

where $\mathcal{L}_{LR_i}$ represents the loss of each patch. Finally, the overall loss function in the pre-training phase is defined as:

$$\mathcal{L}_{pretrain} = \mathcal{L}_{GR} + \alpha\mathcal{L}_{WLR} + \beta\mathcal{L}_{NNM} \tag{8}$$

where α and β are used to balance the relative contributions of these loss terms.

3 Experiments

3.1 Experimental Setup

Datasets. Our method is validated on one private dataset and two public datasets. The CCTA139 dataset is a private dataset of coronary arteries containing 139 samples. BTCV is a public dataset for multi-organ segmentation.

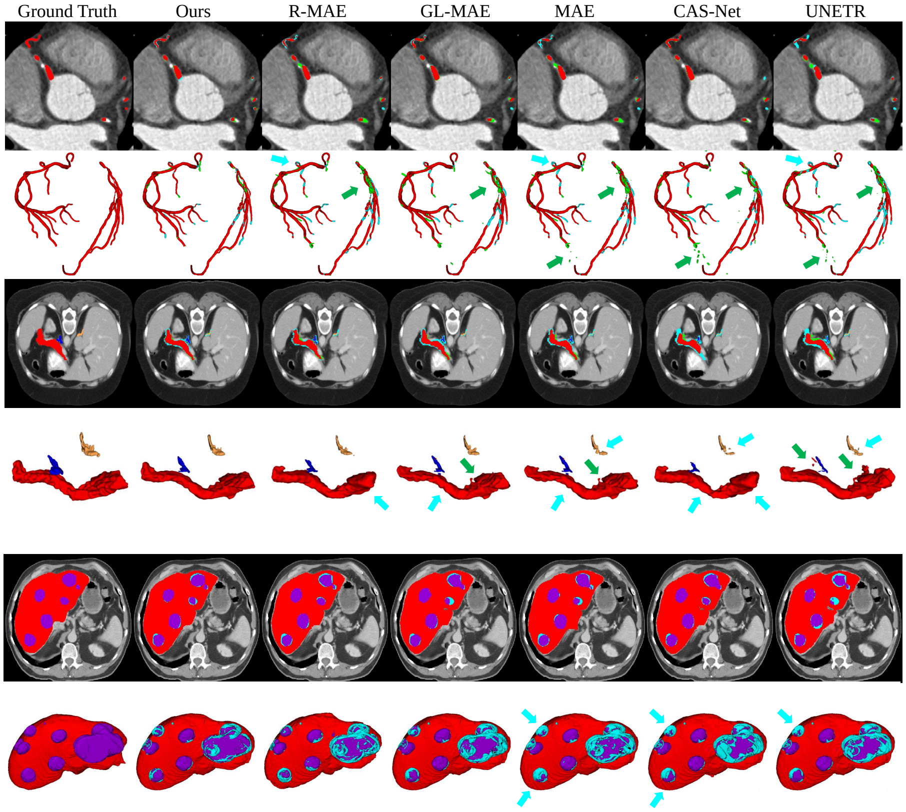

Fig. 2: Qualitative comparison on CCTA139, BTCV and LiTS datasets. The green area indicates False Positive, and the blue area indicates False Negative.

Organs with a lower voxel ratio are gallbladder, esophagus, veins, etc. LiTS is also a public dataset for liver tumor segmentation. The largest and smallest tumor volumes are 987.66 cm^3 and 0.04 cm^3. Compared with large tumors, extremely small non-salient objects are almost difficult to observe.

Evaluation Metrics. We employ the Dice Similarity Cofficient (DSC)[%], Average Surface Distance (ASD)[mm] and Hausdorff Distance (HD)[mm] as metrics.

Implementation Details. We adopt a self-supervised training approach, using the MG-MAE framework for pre-training and combining the MG-MAE encoder with the UNETR decoder for fine-tuning. We use Dice loss [17] as the loss function for segmentation of non-salient objects in the fine-tuning stage.

3.2 Comparsion with Sate-of-the-Art Methods

We conduct experiments on the CCTA139, BTCV, LiTS datasets to evaluate our method, comparing it with several state-of-the-art methods, i.e., U-Net [19],

Table 1: Comparison with other methods on the CCTA139, BTCV, LiTS datasets. The best-performing results are highlighted in bold. SL and SSL represent supervised learning and self-supervised learning with fine-tuning.

	Methods	CCTA139			BTCV			LiTS		
		DSC↑	ASD↓	HD↓	DSC↑	ASD↓	HD↓	DSC↑	ASD↓	HD↓
SL	UNet(MICCAI'15)	62.75	9.42	43.22	75.65	10.28	32.40	63.49	25.61	34.61
	ResUNet(ISPRS'20)	69.06	7.57	24.88	77.35	8.84	23.81	66.81	21.56	27.33
	TransBTS(MICCAI'21)	75.48	6.13	17.22	80.87	4.45	18.41	66.48	17.58	18.72
	TransUNet(arXiv'21)	75.81	5.80	15.92	80.44	4.19	18.50	67.39	16.34	15.11
	UNETR(WACV'22)	76.68	2.33	14.21	79.45	5.53	18.44	66.99	16.81	15.32
	CAS-Net(MEDIA'23)	76.93	2.52	8.01	81.00	3.19	16.78	70.29	6.86	12.04
SSL	LoMaR(arXiv'22)	77.56	3.06	7.88	81.43	4.45	12.70	68.56	10.44	13.15
	MAE(CVPR'22)	78.62	2.52	7.52	81.49	3.07	12.76	68.83	7.11	12.94
	GL-MAE(arXiv'23)	78.44	**1.56**	5.18	82.05	3.34	9.03	71.10	4.79	10.87
	R-MAE(ICLR'24)	79.63	2.15	5.04	81.72	3.37	11.55	70.38	6.01	12.73
	FocusMAE(CVPR'24)	78.95	1.87	6.10	81.01	4.14	15.55	69.24	9.97	11.50
	SMA(ICLR'24)	79.97	2.47	7.50	81.56	3.53	10.27	69.42	6.25	11.36
	MG-MAE(Ours)	**80.71**	1.63	**4.02**	**82.60**	**2.02**	**8.64**	**71.77**	**3.86**	**10.53**

ResUNet [6], TransBTS [23], TransUNet [3], UNETR [11], CAS-Net [7], LoMaR [4], MAE [12], GL-MAE [26], R-MAE [18], FocusMAE [2], and SMA [24]. As shown in Table 1, our method achieves 80.71%, 82.60%, 71.77% on DSC and 4.02mm, 8.64mm, 10.53mm on HD, respectively. Compared with the supervised method UNETR, the self-supervised method has been significantly improved; in the task of non-salient object segmentation, our method is more suitable for this task. We further conduct a qualitative comparison among these methods.

Table 2: Effectiveness of different components. We employ UNETR as the baseline and successively add loss functions for effectiveness analysis.

Model	Loss component				CCTA139		BTCV		LiTS	
	$+\mathcal{L}_{GR}$	$+\mathcal{L}_{LR}$	$+\mathcal{L}_{WLR}$	$+\mathcal{L}_{NNM}$	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓
Baseline	-	-	-	-	76.68	14.21	79.45	18.44	66.99	15.32
Model 1	√	-	-	-	78.62	7.52	81.49	12.76	68.83	12.94
Model 2	√	√	-	-	78.99	7.30	82.11	8.53	69.58	12.55
Model 3	√	√	√	-	79.28	6.49	82.24	7.65	70.13	11.20
Model 4	-	-	√	-	77.21	8.02	81.16	14.57	69.24	12.64
Ours	√	√	√	√	**80.71**	**4.02**	**82.60**	**8.64**	**71.77**	**10.53**

We select methods with better performance for comparison, such as R-MAE, GL-MAE, MAE, CAS-Net, and UNETR. As shown in Fig. 2, we show the 2D and 3D visualizations of each dataset, from top to bottom: CCTA139 dataset, BTCV dataset, LiTS dataset. It is evident that our method exhibits better performance with fewer false positives and false negatives in segmentation of

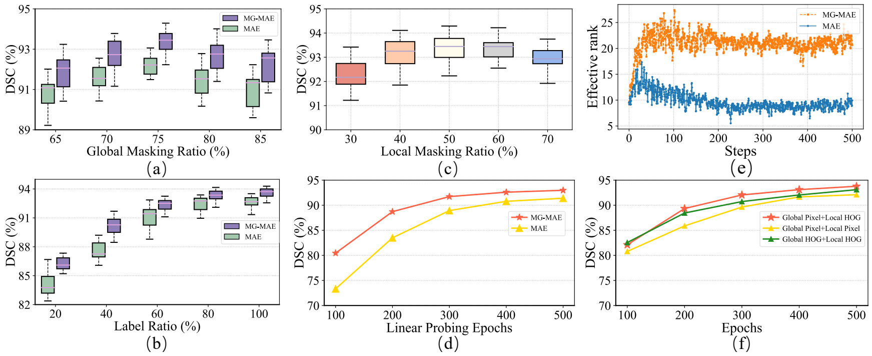

Fig. 3: Analysis of MG-MAE on CCTA139 dataset. We add the aorta of the coronary artery to show the changes of DSC more clearly. (a) and (b) show the analysis of global masking ratio and label ratio of different methods. (c) shows the analysis of local masking ratio of MG-MAE. (d) shows the linear probing experiment. (e) shows the change of effective rank during MAE and MG-MAE training. (f) shows the experiments with different reconstruction targets.

non-salient objects. For non-salient objects in the BTCV dataset, such as the pancreas, gallbladder, and veins, our method achieves DSC scores of 74.56%, 72.75%, and 71.96%, surpassing the current state-of-the-art methods.

3.3 Ablation Study

Effectiveness of different components. We conduct a series of experiments to demonstrate the effectiveness of each component. Our method includes the $\mathcal{L}_{GR}$, $\mathcal{L}_{LR}$, $\mathcal{L}_{WLR}$, $\mathcal{L}_{NNM}$. As shown in Table 2, after integrating these loss functions, our method achieves progressively better performance. Compared with Model 1, Model 2 adds a local branch, achieving DSC improvements of 0.37%, 0.62%, 0.75% across three datasets. Subsequently, Model 3 verifies the effectiveness of the DWA strategy with DSC improvements of 0.29%, 0.13%, 0.55%. Model 4 proves that a single local branch is not as good as the global branch in Model 1. Compared with Model 3, our method adds a NNM constraint, and the effectiveness of this constraint is verified through experiments with DSC improvements of 1.43%, 0.36%, 1.64%.

Masking Ratio and Label Ratio. We systematically analyze the performance of the model during pre-training and fine-tuning. As shown in Fig. 3(a)(c), experimental results on the CCTA139 dataset show that a global masking ratio of 75% and a local masking ratio of 50% yields the best performance, with a DSC of 93.78%. For label usage, 60% label utilization achieves near-optimal accuracy, and full labels are recommended to enhance model robustness in Fig. 3(b).

Linear Probing. We compare the convergence of MG-MAE and MAE in Fig. 3(d), MG-MAE achieves faster convergence and higher DSC in linear probing.

Effective Rank. We compare the effective rank [25] of MG-MAE and MAE in

Fig. 3(e). As the effective rank of MAE decreases, the dimensional features gradually collapse during the training process. According to the orange line, NNM effectively solves this issue as the effective rank increases in MG-MAE.
HOG vs Pixel. We compare reconstruction targets in Fig. 3(f), MG-MAE achieves optimal performance when integrating global Pixel with local HOG.

4 Conclusion

In this paper, we propose the MG-MAE framework for non-salient object segmentation, incorporating both global and local level feature representations. Crucially, the framework utilizes NNM for enhancing feature diversity and DWA for optimizing reconstruction by prioritizing challenging patches. Experimental results demonstrate significant performance improvements over state-of-the-art methods, highlighting the effectiveness of our method.

Acknowledgments. This study was funded by The National Natural Science Foundation of China (No.62101491, No.62401508, No.62106001), The University Synergy Innovation Program of Anhui Province under Grant (GXXT-2021-007), The Anhui Provincial Natural Science Foundation (2208085Y19).

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

References

1. Alexander, R.G., Yazdanie, F., Waite, S., Chaudhry, Z.A., Kolla, S., Macknik, S.L., Martinez-Conde, S.: Visual illusions in radiology: untrue perceptions in medical images and their implications for diagnostic accuracy. Frontiers in Neuroscience **15**, 629469 (2021)
2. Basu, S., Gupta, M., Madan, C., Gupta, P., Arora, C.: Focusmae: Gallbladder cancer detection from ultrasound videos with focused masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11715–11725 (2024)
3. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
4. Chen, J., Hu, M., Li, B., Elhoseiny, M.: Efficient self-supervised vision pretraining with local masked reconstruction. arXiv preprint arXiv:2206.00790 (2022)
5. Cui, S., Wang, S., Zhuo, J., Li, L., Huang, Q., Tian, Q.: Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3941–3950 (2020)
6. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing **162**, 94–114 (2020)
7. Dong, C., Xu, S., Dai, D., Zhang, Y., Zhang, C., Li, Z.: A novel multi-attention, multi-scale 3d deep network for coronary artery segmentation. Medical Image Analysis **85**, 102745 (2023)

8. Fazel, M.: Matrix rank minimization with applications. Ph.D. thesis, PhD thesis, Stanford University (2002)
9. Gać, P., Jaworski, A., Parfianowicz, A., Karwacki, J., Wysocki, A., Poręba, R.: Discrepancies between coronary artery calcium score and coronary artery disease severity in computed tomography angiography studies. Diagnostics **14**(17), 1928 (2024)
10. Gharleghi, R., Chen, N., Sowmya, A., Beier, S.: Towards automated coronary artery segmentation: A systematic review. Computer Methods and Programs in Biomedicine **225**, 107015 (2022)
11. 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)
12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
13. Huang, W., Liu, W., Zhang, X., Yin, X., Han, X., Li, C., Gao, Y., Shi, Y., Lu, L., Zhang, L., et al.: Lidia: Precise liver tumor diagnosis on multi-phase contrast-enhanced ct via iterative fusion and asymmetric contrastive learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 394–404. Springer (2024)
14. Ibrahim, A., Primakov, S., Beuque, M., Woodruff, H., Halilaj, I., Wu, G., Refaee, T., Granzier, R., Widaatalla, Y., Hustinx, R., et al.: Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. Methods **188**, 20–29 (2021)
15. Jing, L., Vincent, P., LeCun, Y., Tian, Y.: Understanding dimensional collapse in contrastive self-supervised learning. arXiv preprint arXiv:2110.09348 (2021)
16. Ma, W., Qu, H., Zhao, J., Chen, B., Principe, J.C.: Sparsity aware minimum error entropy algorithms. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2179–2183. IEEE (2015)
17. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
18. Nguyen, D.K., Li, Y., Aggarwal, V., Oswald, M.R., Kirillov, A., Snoek, C.G., Chen, X.: R-mae: Regions meet masked autoencoders. In: The Twelfth International Conference on Learning Representations (2024)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
20. Roy, O., Vetterli, M.: The effective rank: A measure of effective dimensionality. In: 2007 15th European signal processing conference. pp. 606–610. IEEE (2007)
21. Shen, N., Wang, Z., Li, J., Gao, H., Lu, W., Hu, P., Feng, L.: Multi-organ segmentation network for abdominal ct images based on spatial attention and deformable convolution. Expert Systems with Applications **211**, 118625 (2023)
22. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14668–14678 (2022)
23. Wenxuan, W., Chen, C., Meng, D., Hong, Y., Sen, Z., Jiangyun, L.: Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference

on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 109–119 (2021)

24. Xie, J.W., Lee, Y., Chen, A.S., Finn, C.: Self-guided masked autoencoders for domain-agnostic self-supervised learning. In: The Twelfth International Conference on Learning Representations (2024)
25. Zhang, Q., Wang, Y., Wang, Y.: How mask matters: Towards theoretical understandings of masked autoencoders. Advances in Neural Information Processing Systems **35**, 27127–27139 (2022)
26. Zhuang, J.X., Luo, L., Chen, H.: Advancing volumetric medical image segmentation via global-local masked autoencoder. arXiv preprint arXiv:2306.08913 (2023)