

# MPS-AMS: MASKED PATCHES SELECTION AND ADAPTIVE MASKING STRATEGY BASED SELF-SUPERVISED MEDICAL IMAGE SEGMENTATION

Xiangtao Wang<sup>1</sup>, Ruizhi Wang<sup>1</sup>, Biao Tian<sup>1</sup>, Jiaojiao Zhang<sup>1</sup>, Shuo Zhang<sup>1</sup>,  
Junyang Chen<sup>2</sup>, Thomas Lukasiewicz<sup>3,4</sup>, Zhenghua Xu<sup>1,†</sup>

<sup>1</sup>State Key Laboratory of Reliability and Intelligence of Electrical Equipment, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin, China

<sup>2</sup>College of Computer Science and Software Engineering and Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen, China

<sup>3</sup>Institute of Logic and Computation, TU Wien, Vienna, Austria

<sup>4</sup>Department of Computer Science, University of Oxford, Oxford, United Kingdom

## ABSTRACT

Existing self-supervised learning methods based on contrastive learning and masked image modeling have demonstrated impressive performances. However, current masked image modeling methods are mainly utilized in natural images, and their applications in medical images are relatively lacking. Besides, their fixed high masking strategy limits the upper bound of conditional mutual information, and the gradient noise is considerable, making less the learned representation information. Motivated by these limitations, in this paper, we propose masked patches selection and adaptive masking strategy based self-supervised medical image segmentation method, named MPS-AMS. We leverage the masked patches selection strategy to choose masked patches with lesions to obtain more lesion representation information, and the adaptive masking strategy is utilized to help learn more mutual information and improve performance further. Extensive experiments on three public medical image segmentation datasets (BUSI, Hecker, and Brats2018) show that our proposed method greatly outperforms the state-of-the-art self-supervised baselines.

**Index Terms**— Self-supervised Learning, Conditional Entropy, Mutual Information, Medical Image Segmentation.

## 1. INTRODUCTION

Deep learning has demonstrated remarkable achievements in medical image analysis [1, 2]. In particular, self-supervised learning (SSL) has emerged as a crucial technique for medical image segmentation tasks [3, 4], which is mostly based on contrastive learning. Contrastive learning [5–7] enforces positive samples closer and negative samples further away in latent space to learn representation information. However, these methods only focus on the global semantics of the image and ignore the details of the image and non-subject areas [8]. To solve these problems, masked image modeling [9–12] for self-supervised pretraining has come into being and recently grown in popularity. Masked image modeling (MIM) aims to reconstruct corresponding discrete visual tokens from

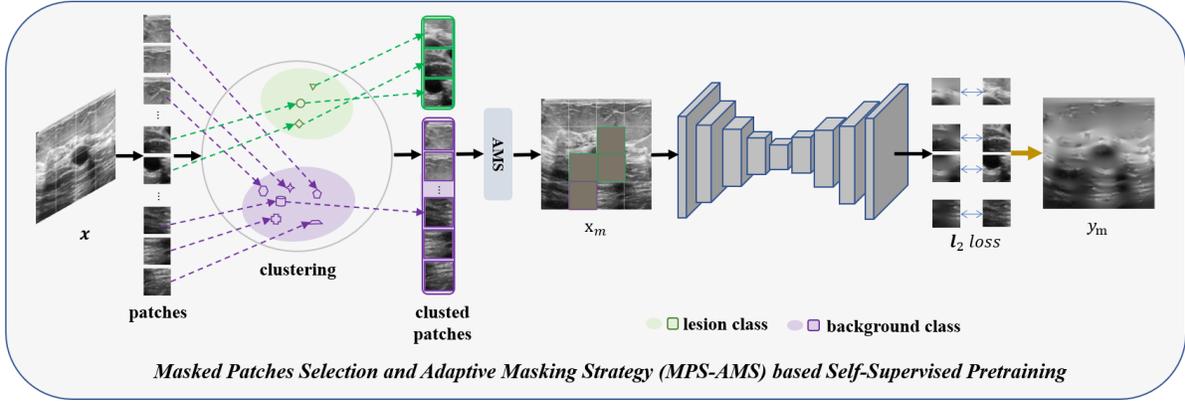
masked input, like MAE [9] and SimMIM [10]. MAE leverages an asymmetric encoder and decoder architecture to predict masked patches from unmasked ones directly. To further maintain image structure, SimMIM takes visible and masked patches as input, and it also lightweights decoder to accelerate pretraining process.

Although MAE and its variants [10–12] have shown promising results, their strategies for selecting masked patches and masking ratio are still unsatisfactory. Specifically, they have not been extensively applied in medical images, where the lesion area is usually small and may be overlooked, resulting in less lesion representation information and limiting the performance of downstream tasks. Additionally, a fixed high masking rate leads to a small learnable conditional mutual information and large gradient noise, which lowers the upper bound of representation information learned and makes optimization challenging [13, 14]. Therefore, the need for masked patches selection and adaptive masking strategy in medical images is compelling.

In this paper, we innovatively propose **Masked Patches Selection and Adaptive Masking Strategy** based self-supervised medical image segmentation (**MPS-AMS**). First, we leverage the masked patches selection strategy to focus on lesions to learn more lesion representation information, which is achieved by choosing the masked patches with a high probability of containing lesions through covariance matrix and k-means clustering. Then, we propose an adaptive masking ratio strategy to improve the upper bound of conditional mutual information to learn more representation information.

The contributions of this paper are briefly summarized as follows: (i) We propose a novel masked patches selection strategy specifically for medical images and an adaptive masking strategy to overcome the shortcomings of existing masked image modeling methods. (ii) To enhance the lesion representation information, we use the masked patches selection strategy to select patches with a higher probability of containing lesions and the adaptive masking ratio strategy to reduce gradient noise and improve the upper bound of conditional mutual information. (iii) Extensive experiments on three public medical image datasets demonstrate that MPS-AMS outperforms state-of-the-art self-supervised methods,

<sup>†</sup>Corresponding author: zhenghua.xu@hebut.edu.cn (Zhenghua Xu).



**Fig. 1.** The illustration of our MPS-AMS architecture. Green and purple areas represent lesion and background, respectively.

and the proposed strategies are effective and essential for improving model performance.

## 2. METHODOLOGY

Figure 1 illustrates the overall structure of our proposed MPS-AMS, which comprises two main processing steps. Firstly, MPS-AMS conducts masked image modeling pretraining using a large set of unlabeled medical images. The resulting modules are then utilized in fully supervised downstream segmentation tasks with a small amount of labeled images.

### 2.1. Masked Patches Selection Strategy

Since current masked image modeling works are mostly focused on natural images, we first propose masked patches selection strategy special for medical images. We sort all the patches in the order of lesion first, then background, and then mask patches in this order. The input image  $x$  contains two parts,  $x = \{x_i, x_{-i}\}$ , where  $x_i$  indicates visible patches and  $x_{-i}$  indicates masked patches.

To get the selected  $x_i$  and  $x_{-i}$ , we define two initialized cluster centers, which are predicted to represent lesion and background class respectively. After dividing  $x$  into patches, we construct a covariance matrix after a softmax layer according to the degree of similarity between different patches. Then, we take k-means to divide all patches into two categories. Considering that most lesions in medical images only occupy a small overall area, we can suppose the category with a small number of clusters as lesions. Besides, we choose k-means because it can achieve good performance with lower complexity and faster efficiency compared with other clustering methods like hierarchical, t-SNE, and so on, and the reason why we choose k-means is well discussed in the section of results.

We evaluate the effectiveness of the proposed masked patches selection strategy by estimating the conditional entropy to represent the uncertainty of the sampling strategy. After classifying patches, the uncertainty of  $x_i$  is reduced, which indicates an improvement of the lower bound. Concretely, We leverage  $H_j$  to indicate the uncertainty of the sampling strategy.  $H_1$  is the lowest bound of uncertainty,

$$H_1 = \mathbb{E}_p(x_i, x_{-i}) \log P(x_i, x_{-i}). \quad (1)$$

$H_2$  is used to indicate the uncertainty in the learning process of neural networks,

$$H_2 = \mathbb{E}_p(x_i, x_{-i}) \log Q(x_i, x_{-i}). \quad (2)$$

Suppose  $H_3$  is the optimal upper bound with the Monte Carlo sampling strategy,

$$H_3 = \mathbb{E}_p(\hat{x}_i, \hat{x}_{-i}) \log P(x_i, x_{-i}). \quad (3)$$

where  $\hat{x}_i$  and  $\hat{x}_{-i}$  represent the best sampling results.

The Monte Carlo sampling strategy is shown below, in the interval  $[a, b]$ ,  $f(x)$  represents the size of the value,  $p(x)$  represents the probability of occurrence, the meaning of the integral result is the output total value.

$$\int_a^b h(x) dx = \int_a^b f(x) p(x) dx = E_{p(x)}[f(x)] \quad (4)$$

Because  $KL(P||Q) > 0$ , we can get  $H_2 \leq H_1$ . Combining the definitions of conditional entropy described in detail [15], we can get  $H_3 \leq H_2 \leq H_1$ , which means that our proposed masked patches selection strategy reduces uncertainty and can help to learn more lesion representation information.

### 2.2. Adaptive Masking Strategy

In contrast to the fixed high masking ratio used in MAE and SimMIM, we propose a novel adaptive masking ratio strategy that combines following insights. Fine-tuning performance is limited under high and fixed masking ratios using in MAE. Furthermore, we note that model's ability to learn representation information and conditional mutual information is higher under larger masking ratios.

Concretely, the initial adaptive ratio  $\sigma_0$  is 25%, which is set according to MAE and it increases with the training process.

$$\sigma = \sigma_0 + \ln(x_e)/\tau, \quad (5)$$

where  $x_e$  donates the training epoch and  $\tau$  is a constant.

Combining with masked patches selection strategy, we can get the numbers of masked patches  $n$  and  $x_m$ , where  $n = \lfloor N \times \sigma \rfloor$  and  $x_m$  indicates the image with masked patches. As illustrated in Fig 1, it is achieved by the  $l_2$  loss.

$$L = \sum_{i=1}^n (y_{m_i} - x_{m_i})^2, \quad (6)$$

where  $x_{m_i}$  indicates the  $i$ -th masked patch,  $y_{m_i}$  indicates the  $i$ -th reconstructed patch.

## 3. EXPERIMENTS AND RESULTS

### 3.1. Experimental settings

To evaluate the effectiveness of our proposed MPS-AMS, we perform extensive experiments on three publicly medical im-

**Table 1.** Results of the proposed MPS-AMS and baselines on BUSI, Hecktor, and Brats2018 datasets.

Methods		BUSI			Hecktor			Brats2018		
		DSC	PPV	Sen	DSC	PPV	Sen	DSC	PPV	Sen
5%	U-Net	0.3863	0.5234	0.4531	0.1762	0.2803	0.1755	0.2059	0.2253	0.2606
	SimCLR	0.4172	0.4129	0.3554	0.2201	0.2385	0.3113	0.2908	0.3009	0.4376
	BYOL	0.4291	0.6991	0.4311	0.1967	0.2179	0.2555	0.2811	0.2867	0.4545
	SwAV	0.4017	0.6128	0.4470	0.2186	0.1909	<b>0.3793</b>	0.2277	0.1884	0.4466
	MAE	0.4793	0.6568	0.5463	0.2560	<b>0.2975</b>	0.2966	0.2898	0.3012	0.4596
	SimMIM	0.4644	0.6847	0.4951	0.2413	0.2745	0.2972	0.2801	<b>0.3095</b>	0.4297
	<b>MPS-AMS</b>	<b>0.5002</b>	<b>0.7034</b>	<b>0.5661</b>	<b>0.2711</b>	<b>0.2975</b>	0.3347	<b>0.2973</b>	0.3035	<b>0.4708</b>
10%	U-Net	0.4876	0.6360	0.5262	0.2541	0.3002	0.2875	0.2529	0.2677	0.3366
	SimCLR	0.5396	0.6439	0.5759	0.2947	0.3325	0.3900	0.3551	<b>0.3459</b>	0.4868
	BYOL	0.5491	0.7044	0.5761	0.3013	0.3106	0.3930	0.3458	0.3058	0.3535
	SwAV	0.5163	0.6325	0.5372	0.2550	0.2669	0.3340	0.2914	0.2562	0.4694
	MAE	0.5639	0.6603	0.6104	0.3195	0.3443	0.3794	0.3578	0.3220	0.4878
	SimMIM	0.5537	0.6918	<b>0.6262</b>	0.2920	0.3325	0.3511	0.3246	0.3149	0.4725
	<b>MPS-AMS</b>	<b>0.5914</b>	<b>0.7305</b>	0.6211	<b>0.3554</b>	<b>0.3681</b>	<b>0.4125</b>	<b>0.3633</b>	0.3163	<b>0.5019</b>
50%	U-Net	0.5714	0.6339	0.6058	0.3090	0.3801	0.3160	0.3535	0.3530	0.4139
100%	U-Net	0.6821	0.8005	0.6542	0.3927	0.4523	0.4736	0.4294	0.4497	0.5224

**Table 2.** Results of our ablation studies on three datasets.

Methods		BUSI			Hecktor			Brats2018		
		DSC	PPV	Sen	DSC	PPV	Sen	DSC	PPV	Sen
5%	<i>base</i>	0.4584	0.6773	0.4841	0.2370	0.2636	0.3051	0.2757	0.2302	0.4227
	<i>base+AMS</i>	0.4629	0.6690	0.4498	0.2479	0.2778	0.3087	0.2790	0.2806	0.3653
	<i>base+MPS</i>	0.4732	<b>0.7459</b>	0.4859	0.2521	0.2865	0.2940	0.2801	<b>0.3095</b>	0.4297
	<b><i>base+AMS+MPS</i></b>	<b>0.5002</b>	0.7034	<b>0.5661</b>	<b>0.2711</b>	<b>0.2975</b>	<b>0.3347</b>	<b>0.2973</b>	0.3035	<b>0.4708</b>

age datasets with supervised learning and state-of-the-art SSL approaches. The results are shown in Table 1.

### 3.1.1. Datasets

(i) The BUSI dataset [16] contains ultrasound scans of breast cancer and consists of 780 images categorized into normal, benign, and malignant. The average image size is  $500 \times 500$  pixels. (ii) The Hecktor dataset [17, 18] contains 25923 slides with CT and PET modalities for head and neck tumor segmentation. (iii) The BraTS2018 dataset [19–21] was released for segmenting brain tumors and includes 22963 scans with four MRI modalities: T1, T1CE, T2, and FLAIR volumes. Furthermore, we selected CT and T1 modalities for experiments on the Hecktor and Brats2018 datasets as they are challenging to segment and can demonstrate the effectiveness of the proposed method on complex datasets.

### 3.1.2. Implementation details

Our MPS-AMS is implemented based on Torch 1.7.0 and CUDA-10.1. For pretraining, we employ ADAM [22] optimizer with a learning rate of 0.0002, and the batch size is 32 for BUSI, 48 for Hecktor, and 36 for BraTS2018. We set  $\tau$  to 12 to ensure a final masking ratio of nearly 80%. For transfer learning, we use the U-Net [23] for downstream segmentation task with ADAM optimizer, an initial learning rate of 0.0002, weight decay of 0.0001, and the learning rate strategy is warmup-cosine-lr. The batch size is set to 32, 31, and 56 for datasets with 5% labeled data, and 32, 90, and 70 for datasets with 10% labeled data, respectively. All datasets are randomly divided by 8:1:1. The training epochs are set to 200 for contrastive learning methods, 800 for masked image modeling methods, and 70 for fine-tuning. The experiments are conducted on 8 GeForce RTX2080 GPUs.

### 3.1.3. Evaluation

We employ three widely used metrics to evaluate our method, including positive predictive value (PPV), sensitivity (Sen), and dice similarity coefficient (DSC). PPV is defined as the ratio of correctly segmented positive pixels to all pixels classified as positive in the segmentation result. Sen represents the ratio of correctly segmented positive pixels to all pixels annotated as positive in the ground truth. DSC is the harmonic mean of PPV and Sen, providing a more comprehensive assessment of model performance.

### 3.1.4. Baselines

To evaluate the performance of our proposed MPS-AMS, we choose randomly initialized U-Net without self-supervised pretraining as the full-supervised baselines, leveraging 5% and 10% annotations ratio. Besides, several state-of-the-art self-supervised learning methods are chosen as the self-supervised learning baselines in our experiments, including SimCLR [5], BYOL [6], SwAV [7], MAE [9], and SimMIM [10]. We evaluate the quality of the learned representations by transferring the weight from different self-supervised learning methods to the medical image segmentation task, and then we evaluate their downstream performances.

All baselines are implemented and run leveraging similar procedures and settings as those in their original papers, and additional parameter adjustments are made to our best efforts.

## 3.2. Main results

To investigate the effectiveness of MPS-AMS, we conduct experiments on three datasets and compare the performance with two state-of-the-art baselines: Fully Supervised Baseline (i.e., Fully Supervised) and Self-Supervised Baselines (i.e.,

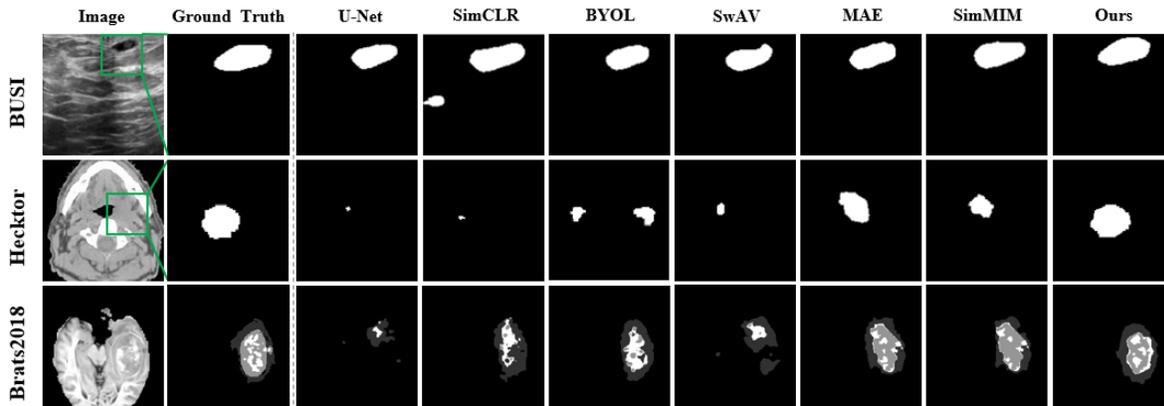


Fig. 2. Visualized segmentation results on the BUSI, Hecktor and BraTS2018 datasets with 10% labeled data.

Table 3. Results of different clustering methods on Brats2018 with 10% labeled data.

methods	k-means	hierarchical	t-SNE	DBSCAN
DSC	0.3633	0.3474	0.3716	0.3592
complexity	$O(n)$	$O(n^2)$	$O(n^2)$	$O(n^2)$

SimCLR, MAE). For a fair comparison, we use the same backbone network (U-Net) with 5% and 10% annotations across all methods. The experimental results are shown in Table 1 and the segmentation results are shown in Figure 2.

As shown in Table 1 and Table 2, MPS-AMS generally outperforms all baselines, which proves it achieves better segmentation performance with limited annotations. Besides, our proposed strategies are also well demonstrated.

#### Compare with Fully Supervised Learning from Scratch.

Specifically, MPS-AMS generally outperforms the baseline model trained from scratch by a large margin with 5% and 10% annotations. Furthermore, when using 10% annotations, we can generally outperform the fully supervised method with 50% annotations.

#### Compare with Self-Supervised Learning Baselines.

Then, we compared our MPS-AMS with state-of-the-art self-supervised methods on the BUSI, Hecktor, and BraTS2018 datasets with 5% and 10% labeled data. Firstly, we find that self-supervised methods generally outperformed fully supervised learning from scratch using partial annotations. This suggests that, in addition to limited labeled data, self-supervised methods also learn useful information from a large amount of unlabeled data. Secondly, when comparing MPS-AMS with SimCLR, BYOL, SwAV, MAE, and SimMIM, we observe that MPS-AMS significantly outperformed these methods on all datasets. Specifically, MPS-AMS achieves 5.18%, 4.23%, 7.51%, 2.75%, and 3.77% higher DSC index scores than other self-supervised baselines under the 10% annotation ratio. This improvement can be attributed to the use of masked patches selection and adaptive masking strategy in MPS-AMS, which reduces the uncertainty of masked patches and improves the upper bound of conditional mutual information, thereby learning more comprehensive and fruitful representation information.

Moreover, we have found that in some cases, the PPV of our method is not always the best. This may be attributed to marginal background patches that bear some resemblance to lesion patches. To address this issue, we plan to explore atten-

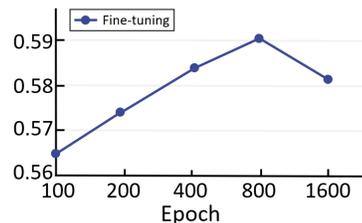


Fig. 3. Training schedules on BUSI with 10% labeled data.

tion mechanisms [24, 25] to make model pay more attention to foreground patches.

**Analysis of Visualized Segmentation Results.** Moreover, all the aforementioned findings are further supported by the visual results presented in Figure 2. The results show that MPS-AMS outperforms all other self-supervised medical image segmentation methods, with segmentation results more similar to the ground truth. These visual observations provide further evidence that the effectiveness of our proposed masked patches selection and adaptive masking strategy.

**Analysis of Clustering Methods.** We have tried different kinds of clustering methods. As shown in Table 3, we can see that under the situation of Brats2018 with 10% labeled data, the result of k-means is only 0.83% different from the result of the best method t-SNE, but it can bring a significant reduction in computational complexity, thereby improving the efficiency of model operations, which indicates k-means can achieve good performance with lower complexity and faster efficiency. Considering actual hardware and time requirements, k-means is the best choice.

**Analysis of Training Schedules.** We also test the performance of different epochs and the results are shown in Figure 3. The vary trend before convergence is the same as MAE and it achieves the best when epoch reaches 800 as mentioned before. When the epoch continues to increase, performance begins to decline due to an excessive masking ratio.

### 3.3. Ablation study

To verify the effectiveness of our proposed MPS-AMS, we conducted further experiments to evaluate the segmentation performance of three intermediate models: *base*, *base+AMS*, and *base+MPS*. The *base* is a masked autoencoder based on U-Net, which aims to reconstruct masked patches of the input image and output an image of the same size as input image. Its masking ratio is 75%. The *base+AMS* model can be seen as the *base* model with an adaptive masking strategy, while *base+MPS* is the *base* model based on the masked patches

selection strategy. We conduct the above ablation studies on BUSI, Hecktor, and BraTS2018 datasets using 5% ratios of annotations. The corresponding results are shown in Table 2.

As shown in Table 2, different strategies contribute differently to the model performance on segmenting tasks. Concretely, *base+AMS* outperforms *base*, indicating that fixed high masking ratio limits the upper bound of representation learning capacity and AMS can solve the problem effectively, *base+MPS* outperforms *base*, indicating the efficiency of masked patches selection strategy. Besides, MPS-AMS achieves the best performance when AMS and MPS are all utilized. For example, the DSC, PPV, and Sen increase 4.03%, 2.65%, and 6.85% for the BUSI dataset. The results demonstrate that our proposed strategies are highly effective.

#### 4. CONCLUSION

In this paper, we propose a self-supervised medical image segmentation method named MPS-AMS, which is based on masked patches selection and adaptive masking strategy. The proposed method can not only alleviate the limitations of current MIM methods in medical images, but also improve the upper bound of conditional mutual information, and reduce gradient noise, thus learning more representation information and achieving better performance. To evaluate the effectiveness of our method, we conduct extensive experiments on three public medical image datasets, and the results demonstrate that our method is effective in self-supervised medical image segmentation tasks.

Considering that there is abundant mutual information in multimodal medical image data and the imbalanced data problem, it is worth of conducting further investigations to apply MIM methods in multimodal [26] and imbalanced [27] medical image analysis tasks to extract more representation information and enhance the deep models' performances.

#### 5. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under the grants 62276089, 61906063 and 62102265, by the Natural Science Foundation of Hebei Province, China, under the grant F2021202064, by the "100 Talents Plan" of Hebei Province, China, under the grant E2019050017, by the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under the grant GML-KF-22-29, and by the Natural Science Foundation of Guangdong Province of China under the grant 2022A1515011474.

## 6. REFERENCES

- [1] Zhenghua Xu, Shijie Liu, Di Yuan, Lei Wang, Junyang Chen, Thomas Lukasiewicz, Zhigang Fu, and Rui Zhang, “ $\omega$ -net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution,” *Neurocomputing*, vol. 500, pp. 177–190, 2022.
- [2] Di Yuan, Yunxin Liu, Zhenghua Xu, Yuefu Zhan, Junyang Chen, and Thomas Lukasiewicz, “Painless and accurate medical image analysis using deep reinforcement learning with task-oriented homogenized automatic pre-processing,” *Computers in Biology and Medicine*, vol. 153, pp. 106487, 2023.
- [3] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert, “Self-supervised learning for medical image analysis using image context restoration,” *Medical Image Analysis*, vol. 58, pp. 101539, 2019.
- [4] Son T. Ly, Bai Lin, Hung Q. Vo, Dragan Maric, Badri Roysam, and Hien V. Nguyen, “Student collaboration improves self-supervised learning: dual-loss adaptive masked autoencoder for brain cell image analysis,” *arXiv preprint arXiv:2205.05194*, 2022.
- [5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, “A simple framework for contrastive learning of visual representations,” *arXiv preprint arXiv:2002.05709*, 2020.
- [6] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al., “Bootstrap your own latent—a new approach to self-supervised learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271–21284, 2020.
- [7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.
- [8] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang, “Context autoencoder for self-supervised representation learning,” *arXiv preprint arXiv:2202.03026*, 2022.
- [9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, “Masked autoencoders are scalable vision learners,” in *CVPR*, 2022, pp. 16000–16009.
- [10] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu, “SimMIM: A simple framework for masked image modeling,” in *CVPR*, 2022, pp. 9653–9663.
- [11] Jihao Liu, Xin Huang, Yu Liu, and Hongsheng Li, “MixMIM: Mixed and masked image modeling for efficient visual representation learning,” *arXiv preprint arXiv:2205.13137*, 2022.
- [12] Xiang Li, Wenhai Wang, Lingfeng Yang, and Jian Yang, “Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality,” *arXiv preprint arXiv:2205.10063*, 2022.
- [13] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin, “Unsupervised learning of visual features by contrasting cluster assignments,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.
- [14] Zhenda Xie, Zheng Zhang, Xizhou Zhu, Gao Huang, and Stephen Lin, “Spatially adaptive inference with stochastic feature sampling and interpolation,” in *ECCV*, 2020, pp. 531–548.
- [15] Guoyin Wang, Hong Yu, DC Yang, et al., “Decision table reduction based on conditional information entropy,” *Chinese Journal of Computers*, vol. 25, no. 7, pp. 759–766, 2002.
- [16] Walid Al-Dhabyani, Mohammed Goma, Hussien Khaled, and Aly Fahmy, “Dataset of breast ultrasound images,” *Data in Brief*, vol. 28, pp. 104863, 2020.
- [17] Yading Yuan, “Automatic head and neck tumor segmentation in PET/CT with scale attention network,” in *3D Head and Neck Tumor Segmentation in PET/CT Challenge*, 2020, pp. 44–52.
- [18] Vincent Andrearczyk, Valentin Oreiller, Martin Vallières, Joel Castelli, Hesham Elhalawani, Mario Jreige, Sarah Boughdad, John O. Prior, and Adrien Depeursinge, “Automatic segmentation of head and neck tumors and nodal metastases in PET-CT scans,” in *MIDL*, 2020, pp. 33–43.
- [19] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burten, Nicole Porz, Johannes Slotboom, Roland Wiest, et al., “The multimodal brain tumor image segmentation benchmark (BRATS),” *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.
- [20] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos, “Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features,” *Scientific Data*, vol. 4, no. 1, pp. 1–13, 2017.
- [21] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al., “Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge,” *arXiv preprint arXiv:1811.02629*, 2018.
- [22] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-Net: convolutional networks for biomedical image segmentation,” in *MICCAI*, 2015, pp. 234–241.
- [24] Zhenghua Xu, Tianrun Li, Yunxin Liu, Yuefu Zhan, Junyang Chen, and Thomas Lukasiewicz, “PAC-Net: Multi-pathway FPN with position attention guided connections and vertex distance IoU for 3D medical image detection,” *Frontiers in Bioengineering and Biotechnology*, vol. 11, pp. 1049555, 2023.
- [25] Zhenghua Xu, Chang Qi, and Guizhi Xu, “Semi-supervised attention-guided CycleGAN for data augmentation on medical images,” in *IEEE BIBM*, 2019, pp. 563–568.
- [26] Shuo Zhang, Jiaojiao Zhang, Biao Tian, Thomas Lukasiewicz, and Zhenghua Xu, “Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation,” *Medical Image Analysis*, vol. 83, pp. 102656, 2023.
- [27] Jianfeng Wang, Thomas Lukasiewicz, Xiaolin Hu, Jianfei Cai, and Zhenghua Xu, “RSG: A simple but effective module for learning imbalanced datasets,” in *CVPR*, 2021, pp. 3784–3793.