Behavioral Analysis of Semantic-Guided Self-Supervised Learning for Low-Labeled Medical Image Segmentation

Jihan Fan

{23208217@stu.nchu.edu.cn}

School of Software, Nanchang Hangkong University, Nanchang, Jiangxi, China

Abstract. This paper focuses on brain tumor segmentation via the BraTS 2018 dataset, reproducing and analyzing typical self-supervised pre-training methods (e.g., SimMIM) under low-label conditions. It then constructs Teacher-Guided SimMIM (TG-SimMIM) as an analytical tool (not a performance optimizer): using foreground probability maps from limited labeled data, it softly guides mask sampling in pre-training to explore its low-label behaviors. Experiments (varying label proportions/random seeds) show: moderate labeling brings limited but stable self-supervised pre-training gains; extremely low labeling makes teacher-guided semantics introduce cumulative bias, reducing model performance. This study notes mask-level semantic guidance risks in low-label medical segmentation, offering evidence for related methods' applicable boundaries.

Keywords: Medical image segmentation, Self-supervised learning, Semantic-guided.

1 Introduction

Medical image segmentation is particularly important in clinical tasks such as tumor radiotherapy, surgical navigation and curative effect evaluation. But the problem is that there are too few high-quality annotation datasets. Labeling is time-consuming and expensive, and the labeling results of different experts are not the same, which leads to noise in the data and poor generalization ability of the model. Therefore, how to achieve high-precision segmentation with a small number of labels is a big problem [1].

Self-supervised learning can help us reduce the dependence on labeled data. Its idea is to design some proxy tasks first, learn useful features from unlabeled data, and then fine-tune them with a small amount of labeled data to adapt to downstream tasks. Among them, Masked Image Modeling is widely used in medical image analysis because it is very successful in the field of natural images. Typical methods, such as MAE and SimMIM, randomly cover a part of the input image, and then let the model restore the covered content, so that the model can learn the context structure and potential semantic information[2, 3].

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365099

Previous studies have found that directly moving the MIM strategy from natural images to medical image segmentation will ignore the sparseness and particularity of lesion semantics, so it is impossible to achieve high-precision segmentation. AMLP method reduces semantic deviation by introducing lesion patch selection and dynamic masking. SACL, on the other hand, enhances the generalization ability under the condition of insufficient data by improving semantic stability and cross-domain consistency. However, these methods usually rely on complex multi-module frameworks and multiple loss functions to optimize together, which limits their practical application in small-scale experiments and clinical scenarios. In a word, although the self-monitoring methods based on masked image modeling show good potential in medical image segmentation, their effectiveness and stability in labeling data and the introduction of semantic guidance strategies still lack systematic analysis. At present, most studies only focus on how to design new methods and improve performance, but ignore the impact of different positions of semantic information injection in the self-supervised learning stage. Especially when there is very little labeled data, the problem of "teacher bias" that may occur has not been fully discussed. In order to solve this problem, this paper takes the brain tumor segmentation task as an example and mainly studies medical image segmentation with a few labeled data points. The core goal is to find out the applicable boundaries and potential risks of this self-monitoring strategy with semantic awareness in the task of low-label medical segmentation, and provide an experimental basis for designing new methods in the future.

2 Related work

2.1 Self-supervised learning and masked image modeling

Self-supervised learning will create some tasks by itself. These tasks don't need us to label data manually. It can make computers learn useful things from a lot of unlabeled data. This has become an important research direction because it can reduce the dependence of supervised learning on labeled data. In recent years, masked image modeling (MIM) has made great progress in the field of natural images, such as MAE and SimMIM.

The core assumption of the above method is that the semantic information in the picture is evenly distributed in space, and random occlusion can make the model learn the contextual semantic relationship better. This assumption is usually true in natural image scenes, but whether it is applicable in medical images needs further verification[4].

2.2 The Application of Mask Modeling in Medical Image Analysis

With the success of Masked Image Modeling (MIM) in the field of natural images, related research began to introduce it into medical image analysis tasks, such as organ segmentation, lesion detection and classification [5]. The existing research shows that self-supervised pre-training can enhance the feature expression ability of the model to a certain extent under the condition of appropriate labeling, thus improving the performance of downstream segmentation tasks [6].

However, medical images are quite different from ordinary pictures in information distribution and structural characteristics. Those pathological areas that are important for diagnosis usually only occupy a very small part of the picture, and their positions are particularly scattered and

irregular. In this way, if the random masking method is still used, the model may only focus on restoring a large background area, while the really key small lesions are rarely covered and practiced, so the model can't learn the knowledge related to the lesions well. Previous studies have found that it may be difficult to give full play to the advantages of self-supervised learning without labels if the MIM strategy in the field of natural images is directly copied to the task of medical image segmentation.

2.3 Semantic-aware self-supervised methods and their limitations

In order to solve the problem of random masking in medical image segmentation, some studies add a semantic perception mechanism to self-supervised learning, so that the model can pay more attention to key areas. For example, the AMLP method improves the semantic expression by screening lesion blocks and adjusting the mask ratio, while SACL uses semantic consistency and comparative learning to enhance the stability and generalization ability of features in the case of few samples [7, 8]. However, these methods usually require complex multi-stage training and joint optimization of multiple loss functions, and it is very troublesome to realize and adjust parameters. In addition, they default that semantic priors are useful in all annotation scales, but they do not analyze the possible deviation accumulation problem in extremely low annotation scenes. This paper mainly studies the behavior stability and failure mode of semantic awareness strategy under the condition of few labeled data, not just the performance improvement. By comparing TG-SimMIM and the random masking strategy, this paper wants to analyze the actual effect and potential risks of semantic guidance under different labeling scales.

3 Method

3.1 Overview of framework

In order to systematically study the behavior characteristics of intelligent masking strategies that can "understand" the image content in the task of medical image segmentation with few labeled data, this paper builds an analysis framework to study this self-monitoring masking modeling method based on semantic guidance. This framework mainly takes TG-SimMIM as the core research object. By controlling the way that the "teacher" signal affects the masking distribution, this paper analyzes the effect of semantic guidance on the dynamic process and stable performance of model learning under different labeled data. This framework does not aim to pursue the optimal segmentation performance, but rather observes the impact of explicitly controlling the injection method of semantic information during the self-supervised pre-training stage on the dynamic and stability of model learning under different labeling scales.

The overall process follows a two-stage paradigm of "self-supervised pre-training - supervised fine-tuning - multiple random seed evaluation". In the self-supervised stage, different forms of semantic-guided masking strategies are introduced to compare their differences with traditional random masking in feature learning; in the downstream segmentation stage, the network structure and training configuration remain the same, only the pre-training initialization method is changed to eliminate other interfering factors. Through this analytical framework, this paper focuses on examining whether the semantic perception mechanism

always has a positive effect under low annotation conditions, as well as its potential bias accumulation risks.

3.2 Semantic-guided Mask Generation Strategy

In order to simulate the semantic perception masking behavior without introducing complex supervisory information, this paper adopts an approximate semantic partitioning strategy based on unsupervised clustering to perform coarse-grained semantic distinction on image patches. It should be emphasized that this strategy is only used as an analytical tool, and the semantic accuracy of this strategy itself is not the focus of this paper.

Simplified Semantic Classification Strategy (MPS). Specifically, this paper divides the two-dimensional slices of the input medical images into fixed-sized non-overlapping patches (16× 16) and uses the k-means method based on the patch features extracted by the encoder to cluster them, in order to construct a simplified semantic prototype structure. The number of clusters is set to k=2, to simulate the binary distribution of "potential lesion area - background area".
Considering that the lesion areas in medical images are usually significantly fewer in number than the background areas, this paper regards the clustering categories with fewer samples as the potential lesion patch set. Subsequently, all patches are sorted according to the Euclidean distance from the patch to the center of the cluster to construct the priority distribution for mask sampling.

This strategy aims to construct a controllable semantic bias mechanism, so as to analyze the trend of the model's feature learning behavior after introducing weak semantic guidance in the self-supervised pre-training stage.

Dynamic Masking Rate Strategy (AMR). In terms of mask ratio, this paper uses a dynamic mask ratio strategy, so that the fixed mask ratio will not affect the analysis results [9]. At the beginning, the mask rate is set to be relatively low (25%), so that the model can learn the basic structural information steadily at the initial stage of training; After that, with the increase of training rounds, the mask rate will gradually rise to a higher level (70) to simulate more difficult context reconstruction tasks [10].
In each round of training, will calculate how many patches to cover according to the current mask rate, and then give priority to selecting samples from those patches with higher semantic priority. Through this design can introduce some semantic biases while keeping the overall mask size unchanged, so that can analyze the influence of the change of mask distribution on the model learning process.

It should be noted that this dynamic strategy is not to make the model behave better, but to avoid drawing accidental conclusions because only one mask rate is selected.

3.3 Reconstruction target and relative reconstruction loss

In the aspect of self-monitoring training objectives, this paper is based on the pixel-level reconstruction task, and the reconstruction loss is based on the Mean Squared Error (MSE). In order to further analyze the learning differences of the model on different image blocks, Relative Reconstruction Loss (RRL) is introduced to describe the differences in reconstruction difficulty between different occlusion blocks.

Let the reconstruction loss of the i-th masked patch be L_i, and the average reconstruction loss of all masked patches is $\bar{L}$. Then, the relative reconstruction loss is defined as:

$$\boldsymbol{L}_{\mathrm{RRL}} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \frac{\boldsymbol{L_i}}{\bar{\boldsymbol{L}}} \tag{1}$$

3.4 Overall optimization objective

The total loss function during the self-supervised pre-training stage is composed of a linear combination of the basic reconstruction loss and the relative reconstruction loss:

$$\boldsymbol{L} = \boldsymbol{L}_{\mathrm{rec}} + \boldsymbol{\lambda} \boldsymbol{L}_{\mathrm{RRL}} \tag{2}$$

Here, $\boldsymbol{\lambda}$ represents the loss weight, which is used to control the influence degree of the relative reconstruction term during the optimization process. In the experiment, $\boldsymbol{\lambda}$ is set to a fixed value to avoid the interference of additional hyperparameter adjustments on the analysis results.

4 Experimental design

4.1 Experiment settings

This study selected the commonly used public datasets in the field of medical image segmentation, covering different modalities (MRI, PET/CT) and tumor types. The experiments were conducted on the BraTS2018 brain tumor segmentation dataset.

To systematically analyze the behavioral characteristics of semantic-aware self-supervised strategies under different annotation scales, this paper constructed various annotation ratio settings, including three scenarios: 100%, 10%, and 5%. Among them, 100% annotation serves as a fully supervised reference, 10% represents a moderately low annotation condition, and 5% is used to simulate an extremely low annotation scenario. In the low annotation experiments, only the annotation samples corresponding to the respective ratio were used for downstream segmentation fine-tuning, while the remaining samples were only used for the self-supervised pre-training stage.

This study comprehensively evaluated the segmentation performance using authoritative and universal indicators in the field of medical image segmentation, including three items: First, the Dice similarity coefficient (DSC), which is used to measure the overlap of the segmented regions, with a value range of [0,1], and the larger the value, the better the segmentation effect; second, the 95% Hausdorff distance (HD95), which is used to measure the consistency of the segmentation boundary, in voxels, with a smaller value representing a higher boundary matching degree; third, the boundary intersection-over-union (BIoU), which is used to measure the overlap of the boundary, with a value range of [0,1], and the larger the value, the better the boundary segmentation effect. The specific experimental parameter settings are shown in Table 1.

Table 1. Basic Settings of the Experiment

Item	Description
Dataset	BraTS 2018 (2D slice-based)
Input Modalities	T1, T1ce, T2, FLAIR
Task	Multi-class tumor segmentation (WT / TC / ET)
Labeled Ratios	5%, 10%
Backbone	UNet-2D
Pretraining	SimMIM, TG-SimMIM
Metrics	Dice ↑, HD95 ↓, Boundary IoU ↑
Seeds	3 (42, 2026, 3208)

4.2 Experimental process

The experimental process of this study is divided into two stages: self-supervised pre-training and downstream segmentation fine-tuning.

In the self-supervised pre-training stage, the original MRI data are uniformly preprocessed, including resampling, cropping, and intensity normalization, and two-dimensional slices are extracted to avoid the interference of empty slices in the training process. Then, the images are divided into fixed-sized patches, and corresponding mask images are generated according to different methods. The model is trained through the reconstruction task for self-supervised training. After training, the encoder weights are retained.

In the downstream segmentation stage, the pre-trained encoder weights are transferred to the segmentation network, and an encoder-decoder structure similar to nnU-Net is used for supervised fine-tuning. During the fine-tuning process, only the corresponding proportion of labeled data is used, and the model performance is evaluated on the test set.

To reduce the influence of randomness on the experimental conclusions, all experiments are repeated using multiple random seeds, and the average performance and standard deviation are reported.

4.3 Quantitative result analysis

Comparison of segmentation performance under different annotation ratios. As shown in Table 2, under the condition of a 5% extremely low annotation ratio, the overall segmentation performance of all methods was significantly limited. The Baseline method achieved an average Dice of 0.744 ± 0.004, serving as a no-pretraining control, reflecting the basic performance level of the model under very little supervision information.

Table 2. Segmentation Performance under 5% Labeling Ratio

Method	Dice ↑	HD95 ↓	Boundary IoU ↑
Baseline	0.744 ± 0.004	36.89 ± 1.06	0.612 ± 0.008
SimMIM	0.738 ± 0.007	35.34 ± 1.35	0.615 ± 0.003
TG-SimMIM	0.729 ± 0.004	38.60 ± 1.84	0.609 ± 0.005

The SimMIM with random mask self-supervised pre-training achieves a Dice score close to that of the Baseline (0.738 ± 0.007), but significantly lowers the HD95 score, indicating that this method helps improve the spatial consistency of the segmentation boundaries. In contrast,

TG-SimMIM shows a certain degree of decline in both Dice and HD95 scores, but achieves the highest value in the Boundary IoU score, suggesting that the teacher-guided semantic information can enhance the overlap degree of local boundaries, but may introduce additional biases under extremely low annotation conditions, thereby affecting the overall segmentation performance.

As shown in Table 3, the performance comparison of different methods under a 10% annotation ratio is presented. The performance of different methods has improved, and the specific comparison results are as follows.

Table 3. Comparison Table of Segmentation Performance of Each Method under 10% Annotation Ratio

Method	**Dice ↑**	**HD95 ↓**	**Boundary IoU ↑**
Baseline	0.763 ± 0.002	33.62 ± 0.90	0.618 ± 0.004
SimMIM	0.770 ± 0.004	30.46 ± 2.13	0.621 ± 0.003
TG-SimMIM	0.762 ± 0.003	32.03 ± 0.66	0.624 ± 0.004

As the proportion of labeled data increases, all the methods show improvements in all evaluation metrics. Among them, SimMIM performs the best in the Dice (0.770 ± 0.004) and HD95 metrics, indicating that random mask self-supervised pre-training can effectively enhance the overall segmentation accuracy and boundary stability of the model in medium-low labeled scenarios. TG-SimMIM is close to the Baseline in Dice, but achieves the highest value in the Boundary IoU metric, suggesting that the teacher-guided mechanism is more conducive to learning fine-grained boundary consistency features when the supervisory information is relatively sufficient.

Stability analysis under multiple random seed conditions. As shown in Table 4, under the 5% marking condition, all three methods exhibited a certain degree of performance fluctuation in the multiple random seed experiments.

Table 4. Statistical Table of Stability of Multiple Random Seed Experiments at 5% Labeling Ratio

Method	**Mean**	**Std**	**Max**	**Min**	**Seeds**
Baseline	0.777	0.005	0.782	0.772	3
SimMIM	0.778	0.006	0.782	0.772	3
TG-SimMIM	0.765	0.007	0.771	0.758	3

Under the 5% annotation condition, all three methods exhibited certain performance fluctuations in the multi-random seed experiments. Compared to Baseline and SimMIM, the average Dice of TG-SimMIM was the lowest, and its standard deviation reached 0.007, indicating that the introduction of semantic guidance did not improve the stability of performance in extremely low annotation scenarios; instead, it might amplify the performance fluctuations caused by different random initializations. In contrast, SimMIM achieved a slightly higher average performance under the 5% condition, but its standard deviation was still relatively large, suggesting that the benefits of self-supervised pre-training in extremely low annotation scenarios have some uncertainty.

Full supervision condition. Under the 100% annotation condition, all methods were trained under the support of sufficient supervision signals, and the results can be regarded as the upper

limit reference for the model performance under different initialization strategies. The experimental results are shown in Table 5.

Table 5. Comparison Table of Dice Performance of Each Method under Full Supervision ConditionSimMIM

Method	**Dice ↑**
Baseline	0.835
SimMIM	0.851
TG-SimMIM	0.843

Pre-training achieved the highest Dice (0.851) in this setting, indicating that random mask self-supervised learning still provides certain initial advantages for feature learning when supervision is sufficient. In contrast, the performance of TG-SimMIM is slightly lower than SimMIM, but still better than the randomly initialized Baseline. This result suggests that when the supervision signal is sufficient, the teacher-guided semantic information no longer significantly leads to performance degradation, and its potential bias effect is alleviated by strong supervision. Combined with the results of the low-labeling experiments, it can be seen that the effectiveness of the semantic guidance strategy is closely related to the labeling scale. The risk of this strategy under extremely low labeling conditions mainly stems from insufficient supervision rather than the irrationality of the method itself.

5 Conclusion

This study systematically explores the behavior and risks of the semantic awareness self-monitoring strategy in low-label medical image segmentation. Although self-monitoring methods, especially those based on masked image modeling, are expected to improve the segmentation performance, they face the challenge of semantic sparseness and the specificity of lesions. By introducing TG-SimMIM, this paper reveals the potential deviation and stability problems in low-label scenes. The research results emphasize the importance of understanding these strategic boundaries, especially the risks brought by semantic guidance under extremely low label conditions. These findings provide valuable insights for designing a more stable and effective self-monitoring model of medical segmentation in the future.

References

[1] Isensee, F., Jaeger, P. F., Kohl, S. A. A., et al.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods. pp. 203–211 (2021)

[2] He, K., Chen, X., Xie, S., et al.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)

[3] Xie, Z., Zhang, Z., Cao, Y., et al.: SimMIM: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9653–9663 (2022)

[4] Menze, B. H., Jakab, A., Bauer, S., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging. pp. 1993–2024 (2014)

[5] Zhou, Z., Sodha, V., Rahman Siddiquee, M. M., et al.: Models genesis: Generic autodidactic models for 3D medical image analysis. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 384–393 (2019)
[6] Taleb, A., Loetzsch, W., Danz, N., et al.: 3D self-supervised methods for medical imaging. Advances in Neural Information Processing Systems. pp. 18158–18172 (2020)
[7] Wang, X., Wang, R., Lukasiewicz, T., et al.: AMLP: Adjustable masking lesion patches for self-supervised medical image segmentation. IEEE Transactions on Medical Imaging. (2025)
[8] Li, T., Nie, K., Pei, Y.: SACL: Semantics-aware contrastive learning for one-shot medical image segmentation. In: Proceedings of the IEEE International Symposium on Biomedical Imaging. pp. 1–5 (2025)
[9] Bengio, Y., Louradour, J., Collobert, R., et al.: Curriculum learning. In: Proceedings of the International Conference on Machine Learning. (2009)
[10] Arazo, E., Ortego, D., Albert, P., et al.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: Proceedings of the International Joint Conference on Neural Networks. pp. 1–8 (2020)