SaSaMIM: Synthetic Anatomical Semantics-Aware Masked Image Modeling for Colon Tumor Segmentation in Non-contrast Abdominal Computed Tomography

Pengyu Dai[1], Yafei Ou[1(✉)], Yuqiao Yang[1], Dichao Liu[1], Masahiro Hashimoto[2], Masahiro Jinzaki[2], Mototaka Miyake[3], and Kenji Suzuki[1]

[1] Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, Japan
ou.y.ac@m.titech.ac.jp
[2] Department of Radiology, Keio University School of Medicine, Tokyo, Japan
[3] Department of Diagnostic Radiology, National Cancer Center Hospital, Tokyo, Japan

Abstract. Colorectal cancer is a critical global concern, despite advancements in computer-aided techniques, the development of early-stage computer-aided segmentation holds substantial clinical potential and warrants further exploration. This can be attributed to the challenge for localizing tumor-related information within the colonic region of the abdomen when doing segmentation and that cancerous tissue remains indistinguishable from surrounding tissue even with contrast enhancement. In this work, a task-oriented Synthetic anatomical Semantics-aware Masked Image Modeling (SaSaMIM) method is proposed that leverages both existing and synthesized semantics for efficient utilization of unlabeled data. We first introduce a novel fine-grain synthetic mask modeling strategy that effectively integrates coarse organ semantics and synthetic tumor semantics in a label-free manner. Thus, tumor location perception in the pretraining phase is achieved by means of integrating both semantics. Next, a frequency-aware decoding branch is designed to achieve further supervision and representation of the Gaussian noise-based tumor semantics. Since the intensity of tumors in CT follows Gaussian distribution, representation in the frequency domain solves the difficulty in distinguishing cancerous tissues from surrounding healthy tissues due to their homogeneity. To demonstrate the proposed method's performance, a non-contrast CT (NCCT) colon cancer dataset was assembled, aiming at early tumor diagnosis in a broader clinical setting. We validate our approach on a cross-validation of these 110 cases and outperform the current SOTA self-supervised method for 5% Dice score improvement on average. Comprehensive experiments have confirmed the efficacy of our proposed method. To our knowledge, this is the first study to apply task-oriented self-supervised learning methods on NCCT to achieve end-to-end early-stage colon tumor segmentation. Our codes are available at https://github.com/Da1daidaidai/SaSaMIM.

Keywords: Semantic Segmentation · Self-Supervised Learning · Colon Tumor · Non-contrast Computed Tomography

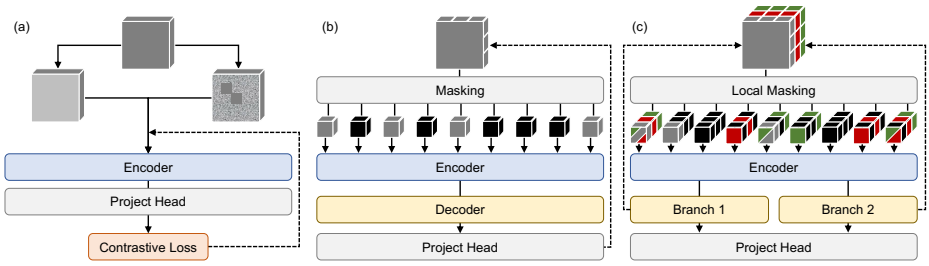

Fig. 1. An illustration of the difference between the proposed self-supervised method and the other methods. Both of the contrastive learning-based method (a) and the mask-based method (b) treat the unified image information only, whereas the proposed method (c) considers multiple types of information including the image itself and the task-related semantics.

1 Introduction

Colorectal cancer represents a major global health challenge, being the third most commonly diagnosed cancer type and the second leading cause of cancer-related deaths worldwide [12]. In addition to colonoscopy, MRI, and CT colonography-based computer-aided diagnosis (CAD), contrast-enhanced CT (CECT)-based CAD has been developed [23]. However, CECT is generally performed for diagnosis purposes after detection of polyps [13], which limits the practical use of CAD systems to diagnosis after detection. The development of automated, non-contrast CT (NCCT)-based colorectal cancer segmentation allows for more applications in diverse clinical scenarios, thus presenting significant clinical potential and research values. This segmentation process faces two main challenges: 1). Lack of NCCT data and annotations due to the difficulty of localizing colon tumor-related information in the abdomen constrains the development of segmentation methods. 2). Differentiating between cancerous and normal tissue is very difficult due to the homogeneity of CT values between them and the absence of contrast agent for enhancing cancerous tissue in NCCT.

Recent studies on the segmentation of colon cancer using CECT scans show promise, but the previously mentioned issues remain. Researchers [23,22] had considered the spatial relationship between colon and colon tumors but introduced additional annotation burdens and the robustness of tumor segmentation depends on the quality of the colon segmentation. Recent studies [24,11,2] addressing colon cancer segmentation within universal segmentation frameworks underscore the substantial reliance on extensive annotated datasets as a notable limitation. AG-CRC [25] tries to address these problems by utilizing imperfect information for region-of-interest identification, though it also requires extensive preprocessing and substantial training data for tumor localization.

To this end, a natural inclination is to consider self-supervised learning [8] as a solution when facing data limitations. As shown in Fig. 1, although a wide range of self-supervised learning methods have been implemented for different

modalities of medical images [20,3,19], unified self-supervised algorithms tend to be more effective on the organ segmentation task. However, lesion segmentation tasks, where the relative target is a much more disunified semantic modality within CT, further widen the gap between self-supervised pre-training and downstream tasks. How to effectively represent goal-related semantics in the self-supervised training phase to minimize such a gap between the pre-training task and the segmentation task remains an important problem to be explored.

In this paper, we present SaSaMIM, a novel masked image modeling method that aims to efficiently facilitate the practical unification of self-supervision and segmentation tasks by effectively embedding tumor semantic synthesis and perception of target semantics in self-supervised pre-training. SaSaMIM employs a novel approach by leveraging imperfect colon localization and Gaussian noise-based tumor semantic synthesis, achieving label-free task semantic synthesis. To deepen the perception of Gaussian noise-based tumor semantics, an additional frequency decoder branch was introduced for fine-grained frequency-aware supervision. Moreover, local masking [20] is integrated within the aggregation process of multi-semantic tokens, enabling the generation of detailed and semantically dense reconstructed representations. Our contributions are mainly three-fold:

- 1) We proposed a task-oriented masked image modeling framework for fine-grained synthetic anatomical semantic perception to achieve high-performance colon tumor segmentation in NCCT.
- 2) We designed a spatial-frequency dual-branch decoder to enhance the model's perception of Gaussian noise-based target semantics.
- 3) We demonstrated through extensive cross-validation experiments the effectiveness and high performance of the proposed model, as well as the limitations of existing uniform semantic masked image modeling models.

2 Method

2.1 Synthetic Semantic-guided Masking

Target Semantic Generation Given the pre-training abdominal CTs $\mathcal{X}_{\mathcal{U}}$, to localize the colon, TotalSegmentator [21] is applied to the CT scans. With a coarse location of the colon available, morphological operations are developed to synthesize correlation semantics $\mathcal{X}_{\mathcal{C}}$. Specifically, let $\hat{\mathcal{X}}_{\mathcal{C}}$ be denoted as the coarse colon location. The process can be formulated as Eq. 1, where $\oplus$ and $\ominus$ represent dilation and corrosion operations, $S(r)$ represent the structural elements with radius r. Here, we configured the radius to 3 voxel points.

$$\mathcal{X}_{\mathcal{C}} = (\hat{\mathcal{X}}_{\mathcal{C}} \oplus S(r)) - (\hat{\mathcal{X}}_{\mathcal{C}} \ominus S(r)) \tag{1}$$

There is substantial evidence [15] that the intensity distribution of colon tumors follows Gaussian distribution, and drawing inspiration from SynthesisTumor [7], the synthesis of colon tumors is viewed as a process encompassing localization, deformation, and texture simulation of Gaussian noise in CT images. First, recognizing the strong correlation between colon tumors and the bowel wall [25],

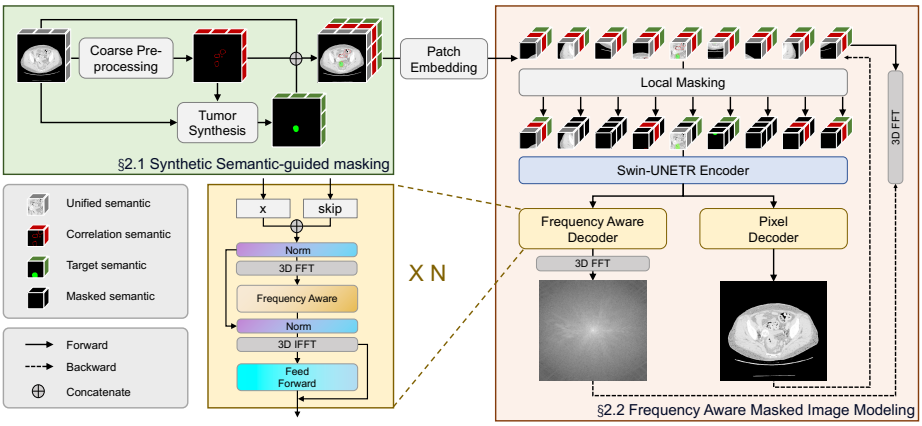

Fig. 2. Overview of our proposed SaSaMIM pre-training method. Our method can be understood in three steps. (1) Data preprocessing and tumor synthesis are employed to generate unified semantics (original image) $\mathcal{X}_{\mathcal{U}}{}^{\in C\times H\times W\times D}$, correlation semantics (bowel wall) $\mathcal{X}_{\mathcal{C}}{}^{\in C\times H\times W\times D}$, and the target semantics (tumor) $\mathcal{X}_{\mathcal{T}}{}^{\in C\times H\times W\times D}$, which are then integrated. (2) The integrated semantic features are randomly and separately masked at the channel level to achieve more fine-grained masking. The masked tokens are concurrently decoded by two distinct branches: a frequency decoder and a pixel decoder. (3) The pre-trained encoder is leveraged for the task of segmentation.

the potential tumor locations t are designated to be adjacent to both the outer and inner parts of the intestinal wall, which can be denoted as,

$$t(x,y,z) \sim \text{Uniform}\left(\{(x,y,z) \mid \mathcal{X}_{\mathcal{C}}[x,y,z] = 1 \wedge (x,y,z) \in \mathcal{X}_{\mathcal{U}}\}\right) \tag{2}$$

Then, as most tumors grow from the centers and gradually swell. The region delineation of the ellipsoid is performed based on selected potential points t, denoted as $E_t \cap \mathcal{X}_{\mathcal{C}} \neq 0$. After elastic deformation [4], E_t can simulate more irregular tumor structures in real scenarios. Following [7], a predefined Gaussian noise texture $T(x,y,z) \sim \mathcal{N}(\mu_t, \sigma_p^2)$ is added to the E_t region after being blurred by the Gaussian filter, which is set to further simulation of the tumor intensity distribution. Thus, the target semantics (tumor) $\mathcal{X}_{\mathcal{T}}$ was generated, which can be formulated as Eq. 3, where $g(x,y,z;\sigma_b)$ denotes the Gaussian filter.

$$\mathcal{X}_T = T(x,y,z) \otimes g(x,y,z;\sigma_b) \tag{3}$$

Fine-grained Semantic Masking The fusion semantics is denoted as Eq. 4, which is defined as the input of our pre-training model.

$$\mathcal{X}_F = \text{Concat}(\mathcal{X}_{\mathcal{U}}, \mathcal{X}_{\mathcal{C}}, \mathcal{X}_{\mathcal{T}}) \in \mathbb{R}^{3C\times H\times W\times D} \tag{4}$$

Instead of creating and implementing more agent tasks for detailed pre-training, we aim for the pre-training task itself to be more aligned with the semantic segmentation task. Thus, local masking [20] is used to partially mask each channel

with different semantic features, which can be denoted as Eq. 5, where $T(\cdot)$ denotes the tokenization operation in Vision Transformer, γ is the mask ratio set to be 0.6 and $\mathcal{X}_M$ denotes the mask.

$$T(\hat{\mathcal{X}_F}) = T(\mathcal{X}_F) \in \mathbb{R}^{(1-\gamma)\cdot 3C\times HWD} + T(\mathcal{X}_M) \in \mathbb{R}^{(\gamma)\cdot 3C\times HWD} \tag{5}$$

2.2 Frequency Aware Masked Image Modeling

Architecture The architecture of our proposed SaSaMIM is depicted in the orange block of Fig. 2, constructed upon a foundational masked autoencoder (MAE) baseline. We adopt the encoder design from the MAE, utilizing a Swin-UNETR [5] to map unmasked patches to discernible markers.

the tumor semantics had its own HU intensity set slightly higher than the surrounding tissue. In traditional mask modeling, such fine-grained semantics are difficult to perceive and reconstruct effectively in the spatial domain. Thus, a frequency-aware branch is designed to perceive Gaussian noise-based semantics in the frequency domain. The core structure of this decoder is the 3D Fast Fourier transform and Inverse Fast Fourier Transform (FFT), and the learnable parameter matrix Ω between them, which can be denoted as Eq. 6, where $\mathcal{X}_f$ represents the latent frequency semantics, x, and skip represent the input for the current layer and input for skip connections in U-shaped structures. $\mathcal{LN}$ denotes layer normalization operation, $\mathcal{F}$ and $\mathcal{IF}$ represent the FFT and IFFT, respectively. Following this, feed-forward networks equipped with residual connections and Layer Normalization (LN) are appended.

$$\mathcal{X}_f = \mathcal{IF}\{\mathcal{F}\{\mathcal{LN}(\Omega \odot (\mathrm{x} + \mathrm{skip}))\}\} \tag{6}$$

Loss Function To effectively capture both spatial and frequency domain discrepancies between the reconstructed and original images, the loss function is composed of three components: pixel domain reconstruction loss, frequency domain weighted loss and regularization term.

Pixel Domain Reconstruction Loss The pixel domain reconstruction loss is computed using the Mean Squared Error (MSE) between the original image and the reconstructed image, defined as Eq. 7, where P and $\hat{P}$ represent the pixels set of the ground truth and the output, respectively.

$$\mathcal{L}_{pix} = MSE(P, \hat{P}) \tag{7}$$

Frequency Domain Weighted Loss A weighted loss is introduced for the frequency domain. The loss is computed as the weighted sum of the Euclidean distance between the frequency representations of the original and reconstructed images, modulated by a dynamic weight matrix, as shown in Eq. 8, where P and $\hat{P}$ represent the frequency band set of the ground truth and the output, respectively. Same as focal frequency loss [9] the weight matrix W is derived from the normalized difference between the reconstructed and original frequency components, ensuring an adaptive focus on critical frequency mismatches.

$$\mathcal{L}_{fre} = \mathrm{Mean}(W \odot \|F - \hat{F}\|) \tag{8}$$

The overall loss function integrates the aforementioned components, as shown in Eq. 9, where L_{pix}^{L1} and L_{pix}^{L2} denote the L1-paradigm and L2-paradigm of the pixel domain reconstruction, respectively, to contribute to reconstructed feature stability and smoothness.

$$\mathcal{L} = \mathcal{L}_{fre} + \mathcal{L}_{pix} + \alpha \cdot \mathcal{L}_{pix}^{L1} + \beta \cdot \mathcal{L}_{pix}^{L2} \tag{9}$$

2.3 Downstream Segmentation

The pre-trained encoder weights are subsequently transferred for downstream tasks and utilize the standard Swin-UNETR [5] architecture for colon tumor segmentation. Specifically, to maintain input dimensional consistency and incorporate additional anatomical insights, the colon location data from TotalSegmentator is also concated to the input image. Finally, the Dice loss [18] is employed to facilitate network convergence.

3 Experiments

3.1 Datasets and Evaluation Metrics

The MSD-Task10 [1] abdominal CECT dataset (portal-venous phase) was chosen as the pre-training dataset, 116 out of 126 were used for training and the remaining 10 were used for validation. Subsequently, to assess the effectiveness of our proposed method, we compiled a dataset of early-stage colorectal cancer CT scans from Keio University Hospital. This dataset encompasses 110 patients diagnosed with colorectal cancer. For each patient, abdominal NCCT scans were obtained, with the tumor regions annotated by a seasoned gastroenterologist and subsequently verified by a senior radiologist. All images were scanned with an in-plane resolution of 512 × 512 pixels, and the z-axis dimensions varied from 249 to 1421 with a median of 474. The voxel spacing is fixed to 1.5mm, 1.5mm, and 1.5 mm. We employed a five-fold cross-validation approach to evaluate the dataset.

We adopt the four standard medical segmentation metrics Dice Score Coefficient (DSC), the 95th percentile Hausdorff Distance (HD95), Precision, and Recall to evaluate the performance of our proposed approach.

3.2 Experiment Setup

All structures for downstream segmentation of self-supervised networks follow the standard setup of Swin-UNETR [5]. The crop size is fixed to (96, 96, 96). The training and testing phases of the proposed pre-training and segmentation network were executed end-to-end on four NVIDIA Tesla V100 GPUs. The batch size and iteration step were configured to 1 and 10,000, respectively. The initial learning rate and learning rate decay were set to 3e-4 and 0.1 every 2,500 steps. Adam [10] was employed as the optimizer.

Table 1. Comparison with the State-of-the-art Medical Image Segmentation Method

Methods	DSC (%)↑	HD (mm)↓	Precision (%)↑	Recall (%)↑
UNet [4]	57.99±2.44	N/A*	66.05±8.92	58.25±2.15
SegResNet [14]	57.47±2.13	144.06±14.72	60.28±6.14	57.43±1.62
Swin-UNETR [5]	58.24±2.47	159.60±30.95	60.95±5.61	59.44±1.37
UNETR [6]	52.19±1.17	175.14±7.62	52.46±2.09	54.51±2.38
UNETR++ [17]	56.80±3.08	151.68±17.41	58.63±4.47	57.37±2.78
MedNeXt [16]	57.61±3.13	153.20±17.55	60.43±6.82	58.17±3.24
SaSaMIM (Ours)	**64.73±1.37**	**141.35±22.84**	**67.07±4.04**	**70.00±5.24**

* Anomalous pixels in some results that lead to extreme HDs (e.g. infinite)

Table 2. 5-fold Comparison on State-of the-art Self-supervised and Universal Methods

	Fold-1		Fold-2		Fold-3		Fold-4		Fold-5		Mean	
Method	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓
Inner-Cutout [19]	58.7	188	53.2	170	58.4	162	64.3	200	60.1	188	58.9	182
SimMIM [3]	55.2	163	52.1	202	56.2	169	62.8	174	56.5	**152**	56.6	172
Universal [11]	58.5	176	51.7	191	58.2	169	63.8	176	62.9	197	59.0	182
DAE [20]	57.4	183	53.9	176	53.0	192	56.7	185	**65.4**	166	59.3	180
SaSaMIM (Ours)	**65.2**	**127**	**64.3**	**143**	**66.5**	**123**	**65.0**	**126**	62.3	184	**64.7**	**141**

3.3 Comparison with the State-of-the-art Methods

For supervised methods, as shown in Table 1, we compare our approach with six state-of-the-art 3D medical image segmentation models, including UNet [4], SegResNet [14], UNETR [6], UNETR++ [17], MedNeXt [16], and Swin-UNETR [5], where UNETR++ are purely transformer-based; UNETR and Swin-UNETR are hybrid models that employ different transformer as encoder and a CNN as decoder; UNet,SegResNet and MedNeXt are purely CNN models. From the quantitative results, it can be found that it is challenging to achieve fine-grained segmentation with a simple transformer structure (e.g. UNETR, UNETR++) under data constraints, which illustrates the demand and necessity of self-supervised pre-training in the current scenario.

For the self-supervised method, as shown in Table 2, we conducted comparisons with the leading-edge self-supervised pre-training method for 3D medical images, including Inner-Cutout [19]. This method represents a self-supervised method that leverages contrast learning, explicitly tailored for Swin-UNETR. Additionally, we compared with SimMIM [3] and DAE [20], both of which utilize uniform and low-level features of the image itself for medical image mask reconstruction. Notably, the models pre-trained using SimMIM demonstrated lower performance across multiple metrics than Swin-UNETR models without pre-training. This outcome supports our hypothesis that an excessive focus on uniform semantics may not effectively enhance the network's ability to segment specific semantics, which is particularly true in scenarios where the feature rep-

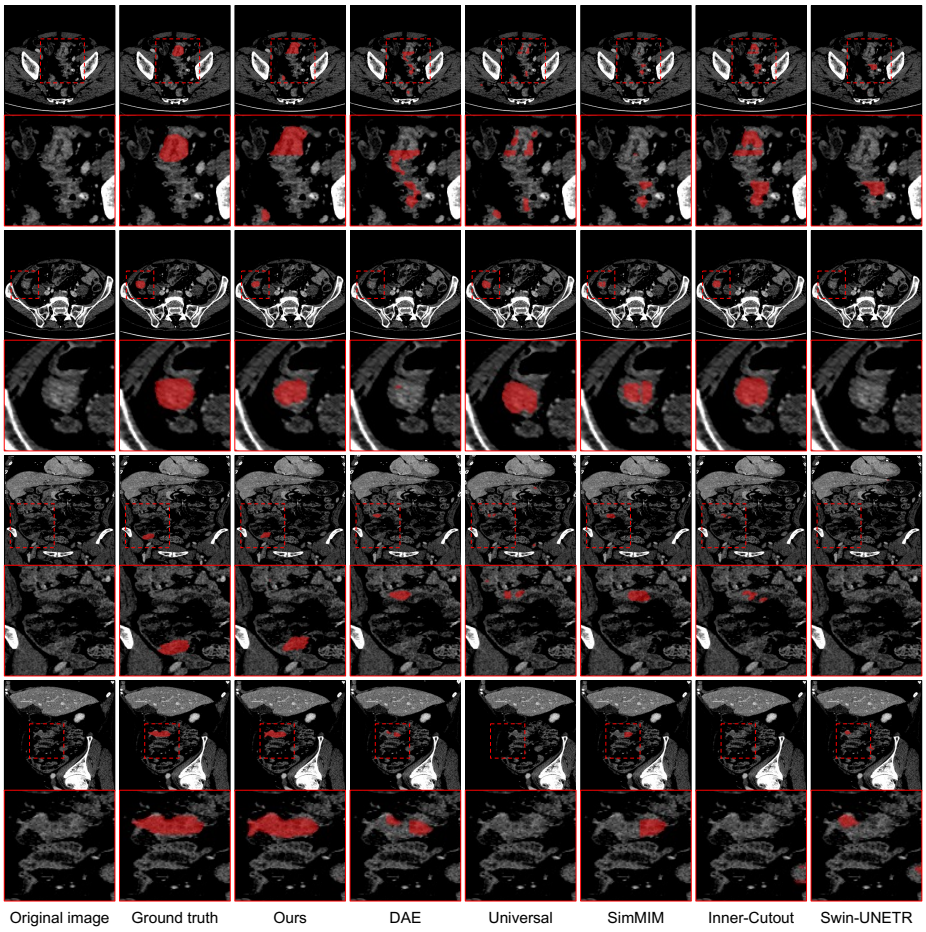

Fig. 3. Visualization on qualitative segmentation results of several exemplar cases. We visualized and analyzed the segmentation results from axial view (row 1,2), coronal view (row 3), and sagittal view (row 4). The red dotted rectangle area in the image is enlarged and placed in the lower half to show the details.

resentations of specific semantics closely resemble those of uniform semantics. To further argue that large-scale unified semantics cannot effectively contribute to task semantic awareness, we additionally compared the Universal Model [11], and we directly transferred the weights of its supervised training on nearly 2,000 3D medical images. Overall, our proposed method improves several critical metrics for medical image segmentation. Also, from the visualization of the qualitative results in Fig. 3, it can be found that our proposed method is closer to the golden standard and achieves effective segmentation in some problematic cases, which are missed by other methods.

Table 3. 5-fold Ablation Experiments of the Proposed Method

	Fold-1		Fold-2		Fold-3		Fold-4		Fold-5		Mean	
Method	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓	DSC↑	HD↓
BNet	59.4	173	54.5	184	55.9	141	61.3	199	58.5	183	57.9	176
BNet+§2.1	64.3	148	60.4	178	63.2	144	**65.9**	133	60.9	**181**	62.9	157
BNet+§2.1+§2.2	**65.2**	**127**	**64.3**	**143**	**66.5**	**123**	65.0	**126**	**62.3**	184	**64.7**	**141**

3.4 Ablation Study

To further validate the effectiveness of our primary components in SaSaMIM, including Synthetic Semantic-guided masking (Sec.2.1) and Frequency Aware Masked Image Modeling (Sec.2.2). All alternative networks were cross-validated on the Keio-colon dataset with five folds, and we show the DSC and HD on all folds for comparison. As depicted in Tab 3, BNet indicates the baseline self-supervised strategy SimMIM. The introduction of Sec.2.1 brings about 5% improvement of Dice score. Furthermore, the adding of Sec.2.2 further brings about 1.8% of Dice score and 16mm Hausdorff Distance improvement.

4 Conclusion

We proposed SaSaMIM, an innovative masked image modeling method that efficiently integrates self-supervision with segmentation semantics, focusing on tumor semantic synthesis and perception in self-supervised pretraining. Leveraging imperfect colon location and Gaussian noise for semantic synthesis without labels, SaSaMIM enhances cancer lesion semantics perception through an additional frequency decoder branch. It also utilizes fine-grain masking in aggregated tokens for detailed semantic-level representations. Extensive experiments show SaSaMIM's superiority over both existing medical image segmentation methods and unified self-supervised pretraining methods for medical imaging.

Acknowledgments. This work was supported by JST-Mirai Program Grant Number JPMJMI20B8, Japan.

Disclosure of Interests. The authors declare no competing interests.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)

2. Chen, J., Xia, Y., Yao, J., Yan, K., Zhang, J., Lu, L., Wang, F., Zhou, B., Qiu, M., Yu, Q., et al.: Cancerunit: Towards a single unified model for effective detection, segmentation, and diagnosis of eight major cancers using a large collection of ct scans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21327–21338 (2023)
3. Chen, Z., Agarwal, D., Aggarwal, K., Safta, W., Balan, M.M., Brown, K.: Masked image modeling advances 3d medical image analysis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1970–1980 (2023)
4. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
5. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2021)
6. 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)
7. Hu, Q., Chen, Y., Xiao, J., Sun, S., Chen, J., Yuille, A.L., Zhou, Z.: Label-free liver tumor segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7422–7432 (2023)
8. Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F.: A survey on contrastive self-supervised learning. Technologies **9**(1), 2 (2020)
9. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13919–13929 (2021)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21152–21164 (2023)
12. Mangat, S., Kozoriz, M.G., Bicknell, S., Spielmann, A.: The accuracy of colorectal cancer detection by computed tomography in the unprepared large bowel in a community-based hospital. Canadian Association of Radiologists Journal **69**(1), 92–96 (2018)
13. Møller, M., Juvik, B., Olesen, S.C., Sandstrøm, H., Laxafoss, E., Reuter, S.B., Bodtger, U.: Diagnostic property of direct referral from general practitioners to contrast-enhanced thoracoabdominal ct in patients with serious but non-specific symptoms or signs of cancer: a retrospective cohort study on cancer prevalence after 12 months. BMJ open **9**(12), e032019 (2019)
14. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization (2018)
15. Ng, F., Kozarski, R., Ganeshan, B., Goh, V.: Assessment of tumor heterogeneity by ct texture analysis: can the largest cross-sectional area be used as an alternative to whole tumor analysis? European journal of radiology **82**(2), 342–348 (2013)

16. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023)
17. Shaker, A., Maaz, M., Rasheed, H., Khan, S., Yang, M.H., Khan, F.S.: Unetr++: delving into efficient and accurate 3d medical image segmentation. arXiv preprint arXiv:2212.04497 (2022)
18. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3. pp. 240–248. Springer (2017)
19. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740 (2022)
20. Valanarasu, J.M.J., Tang, Y., Yang, D., Xu, Z., Zhao, C., Li, W., Patel, V.M., Landman, B., Xu, D., He, Y., et al.: Disruptive autoencoders: Leveraging low-level features for 3d medical image pre-training. arXiv preprint arXiv:2307.16896 (2023)
21. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5) (2023)
22. Yao, L., Li, S., Tao, Q., Mao, Y., Dong, J., Lu, C., Han, C., Qiu, B., Huang, Y., Huang, X., et al.: Deep learning for colorectal cancer detection in contrast-enhanced ct without bowel preparation: A retrospective, multicentre study
23. Yao, L., Xia, Y., Zhang, H., Yao, J., Jin, D., Qiu, B., Zhang, Y., Li, S., Liang, Y., Hua, X.S., et al.: Deepcrc: Colorectum and colorectal cancer segmentation in ct scans via deep colorectal coordinate transform. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 564–573. Springer (2022)
24. Ye, Y., Xie, Y., Zhang, J., Chen, Z., Xia, Y.: Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. arXiv preprint arXiv:2304.03493 (2023)
25. Zhang, R., Bai, Z., Yu, R., Pang, W., Wang, L., Zhu, L., Zhang, X., Zhang, H., Hu, W.: Ag-crc: Anatomy-guided colorectal cancer segmentation in ct with imperfect anatomical knowledge. arXiv preprint arXiv:2310.04677 (2023)