AnatoMask: Enhancing Medical Image Segmentation with Reconstruction-guided Self-masking

Yuheng Li[1,2], Tianyu Luan[3], Yizhou Wu[2], Shaoyan Pan[1], Yenho Chen[2], and Xiaofeng Yang[1,2]

[1] Emory University, Atlanta GA 30322, USA
[2] Georgia Institute of Technology, Atlanta GA 30332, USA
[3] State University of New York at Buffalo, Buffalo, NY, 14260
xiaofeng.yang@emory.edu

Abstract. Due to the scarcity of labeled data, self-supervised learning (SSL) has gained much attention in 3D medical image segmentation, by extracting semantic representations from unlabeled data. Among SSL strategies, Masked image modeling (MIM) has shown effectiveness by reconstructing randomly masked images to learn detailed representations. However, conventional MIM methods require extensive training data to achieve good performance, which still poses a challenge for medical imaging. Since random masking uniformly samples all regions within medical images, it may overlook crucial anatomical regions and thus degrade the pretraining efficiency. We propose AnatoMask, a novel MIM method that leverages reconstruction loss to dynamically identify and mask out anatomically significant regions to improve pretraining efficacy. AnatoMask takes a self-distillation approach, where the model learns both how to find more significant regions to mask and how to reconstruct these masked regions. To avoid suboptimal learning, Anatomask adjusts the pretraining difficulty progressively using a masking dynamics function. We have evaluated our method on 4 public datasets with multiple imaging modalities (CT, MRI, and PET). AnatoMask demonstrates superior performance and scalability compared to existing SSL methods. The code is available at https://github.com/ricklisz/AnatoMask.

Keywords: Medical image segmentation · Self-supervised learning · Masked image modeling

1 Introduction

While supervised learning shows great promise in 3D medical image segmentation, its potential in clinical settings is hindered by a lack of labeled data. As an alternative, self-supervised learning has shown great promise in medical image analysis by extracting semantic representations from unlabeled data [23,35,48,50]. Among various SSL methods, masked image modeling has achieved promising results in natural images [16] and has been applied in medical image

analysis [8, 9, 49]. MIM utilizes a simple task in which the model is asked to reconstruct parts of the masked image, thereby learning detailed representations of the image [16, 45].

However, MIM's success in natural images is heavily dependent on large-scale training data [46], which poses a significant challenge in medical imaging. While many MIM methods utilize random masking strategy and achieved success in natural images [3, 16], we argue that this strategy does not utilize the inherent properties of anatomical regions within medical images and leads to less effective pretraining. In medical images, anatomical regions such as organs and tumors, are deemed more important than the less relevant air-filled and fluid-filled regions. Since anatomical regions are characterized by varying shapes and contrasts, these regions are intuitively more difficult to reconstruct when masked out, requiring more guidance during pretraining. Thus, the previous methods require more training data to achieve the expected results, which greatly limits their application in medical-related tasks. Therefore, to design an effective MIM framework for medical images, it is crucial to identify these complex anatomical regions and create a more challenging task.

To address these challenges, we propose AnatoMask, a reconstruction-guided MIM framework where the model learns from more informative masks to enhance pretraining efficacy. Observing that anatomical regions often featured higher reconstruction losses (Fig. 1 a and b), if we can increase the probability of masking these regions, it would also encourage the network to extract information from these regions more efficiently. Hence, we propose to identify anatomically-significant regions by reconstruction losses. Instead of using a pretrained model to identify these regions, we propose a self-distillation approach where the model can first act as a teacher to generate anatomically-significant masks, and then act as a student to learn from these masks. To ensure that the teacher is adaptive to the student's performance, we update the teacher model with the student's exponential moving average (EMA) weights. In this way, we not only encourage the model to learn how to reconstruct these regions but also where to find such important regions. This dual-learning approach is crucial in developing a deeper anatomical understanding, especially in data-constrained scenarios. Finally, we propose an easy-to-hard masking strategy, where the teacher continuously increases the difficulty of the MIM task as the student progresses, avoiding the model overly focusing on difficult regions at first. We conduct self-supervised learning on TotalSegmentator [40], a comprehensive CT dataset for volumetric medical image segmentation. We adopt a previous SOTA architecture STU-Net [20] and evaluate the pretraining efficacy by finetuning on TotalSegmentator [40], FLARE22 [26], AMOS22 [22] and AutoPETII [11]. We first compare the differences between random masking and our AnatoMask in (Fig. 1 c and d). We then compare the training efficiency of AnatoMask vs random masking (SparK [37]) in (Fig. 1 e). Our method consistently improves the training efficiency, outperforming random masking. We also compare our method with previous state-of-the-art (SOTA) SSL approaches. We find that our method demonstrates superior performance and scaling capabilities.

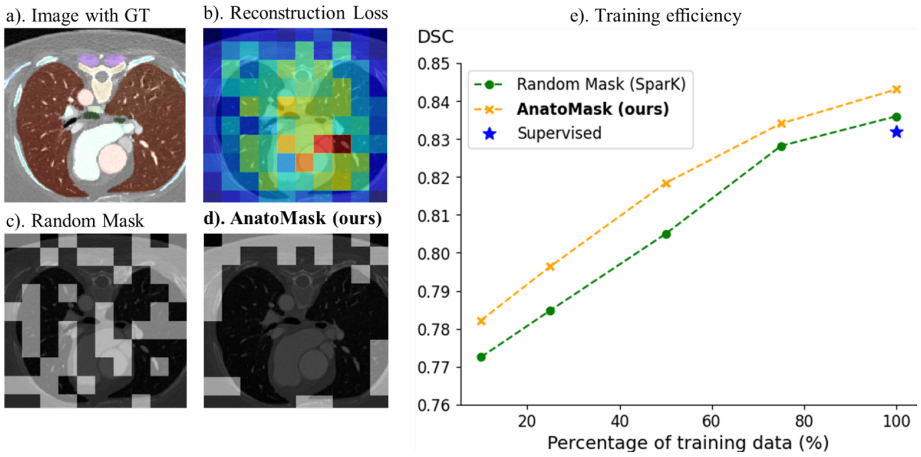


Fig. 1: Comparison of random masking (SparK) vs AnatoMask. We visualize: a). input with anatomical ground truth (GT); b). reconstruction losses obtained by averaging over 2 random masks; c). a random mask ; d). our AnatoMask generated from b. In e), we also compared the training efficiency of AnatoMask and SparK. Transparent areas indicate unmasked regions.

We summarize our contribution as follows: 1). We propose a reconstruction-guided masking strategy, in which the model learns the anatomically significant regions through reconstruction losses and dynamically adjusts the masking to formulate a more challenging and specific MIM objective, thereby improving pretraining efficacy. 2). We propose to leverage self-distillation to perform self-masking. The teacher network first identifies important regions to mask and generates a more difficult mask for the student to solve. The student will then update the teacher with EMA weights. 3). To prevent the network from converging to a suboptimal solution early during training, we propose a masking dynamics function controlling the difficulty of the MIM objective, facilitating an easy-to-hard strategy

2 Related Works

Contrastive learning. Self-supervised learning leverages unlabeled data to learn meaningful representations without explicit supervision, demonstrating significant advancements in pattern recognition [2, 10, 12, 28, 29, 39, 43, 47]. In this paradigm, contrastive learning first gains much attention in computer vision [6, 13, 14, 17, 18, 27] and medical imaging [1, 4, 33, 35]. Contrastive learning extracts powerful semantic representation by minimizing the distance of similar image pairs while maximizing dissimilar ones. However, contrastive learning relies on tailored data augmentations to create such image pairs [7]. Inappropriate augmentations could distort important clinical semantics leading to misrepresentations in the learned features. Also, since segmentation tasks require precise

pixel-level embeddings, contrastive methods could overlook the nuanced features critical for detecting small organs or anomalies.

Masked image modeling for ViTs, ConvNets. Motivated by the success of masked language modeling in natural language processing, masked image modeling has attracted much attention in computer vision. MIM learns image representations by reconstructing their masked parts, a concept that resembles early techniques like context encoders [30]. Naturally, MIM is first applied to vision transformers due to its sequence-to-sequence modeling. Both He et al. [16] and Xie et al. [45] propose to mask random image patches and regress the RGB pixels. More recent advancements in MIM focus on new reconstruction targets [3, 19, 31, 36, 41] or masking strategies [24, 25, 32, 38]. For ConvNets, recent works such as SparK [37] and ConvNeXt V2 [42] propose to leverage sparse convolution to address the information leakage during pretraining. However, novel MIM strategies for medical image segmentation remain largely unexplored.

SSL for medical image segmentation. Previous works propose various pretext tasks to solve medical image segmentation with promising performance. Similar to the context encoder, image inpainting has been applied to various ConvNet backbones to generalize the anatomical contents of medical images [5]. Taleb et al. explore a variety of pretext tasks such as rotation, solving Rubik's cube, and predicting patch locations [34]. Tang et al. propose a combination of pretext tasks and contrastive learning for self-supervising transformer [35]. Fatemeh et al. combine discriminative, restorative, and adversarial learning strategies to extract rich semantic information from medical images [15]. More recent works aim at developing a universal SSL framework for medical images without tailoring pretext tasks. Xie et al. propose a universal pretraining method UniMiSS, leveraging transformer's sequence-to-sequence modeling to pretrain on both 2D and 3D medical images [44]. As another universal framework, MIM also gains attention in medical image analysis. Chen et al. first explore using SimMIM and MAE for 3D medical image segmentation [8]. However, MIM still remains data-intensive and requires large datasets for scaling up model parameters [46]. Furthermore, the success of MIM heavily relies on the mask strategies due to the spatial information redundancy in images [16]. Intuitively, a universal medical MIM task should guide the model to focus on anatomically rich areas for reconstruction. Consequently, the model gains a deeper understanding of the anatomical information to generate masks for more effective learning.

3 Method

3.1 Overview

Figure 2 illustrates the overall pipeline of AnatoMask, which is composed of a student network and a teacher network sharing the same underlying architecture. Each module has an image encoder and an image decoder. Given an input CT volume at epoch t, we first generate a random mask with masking ratio γ similar to SimMIM [45]. Then, we feed the randomly masked input into the teacher network, whose encoder E_T abstracts the input into embeddings and

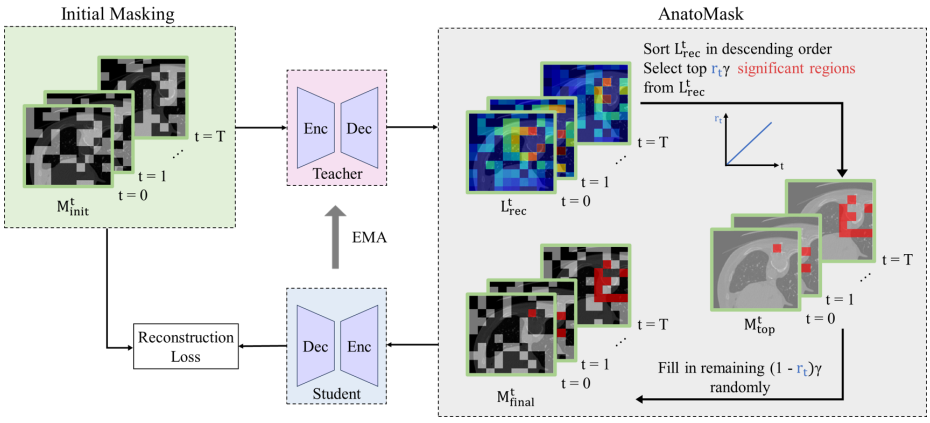


Fig. 2: Overview of proposed AnatoMask. During SSL pretraining at epoch t, the teacher network receives randomly masked inputs and computes the patch-level reconstruction loss L^t_{rec}. The top r_t regions with the highest reconstruction losses are selected to form a binary mask M^t_{top}. Then, the remaining $(1-r_t)\gamma$ areas are randomly filled with binary values to form our final mask M^t_{final}. The student network is trained to reconstruct input masked by M^t_{final}.

its decoder D_T generates the reconstruction. We then compute the reconstruction loss L^t_{rec} on the masked regions by subtracting with input. Next, we sort L^t_{rec} in descending order and use a masking dynamics function r_t to select top anatomical regions with high losses to form a binary mask M^t_{top}. Then, we generate M^t_{final} by randomly masking the remaining regions to maintain an overall masking ratio γ. Finally, we apply the newly generated mask to the input and train the student network on reconstruction loss. To ensure a consistent mask generation between student and teacher, we update the teacher network with an exponential moving average of the student's network weights:

$$\theta_T^{t+1} = m\theta_T^t + (1-m)\theta_S^{t+1}, \tag{1}$$

where m represents the weight decay, θ_T^t and θ_T^{t+1} represent the teacher's parameters at time t and $t+1$ respectively and θ_S^t represents the student's parameter at time t.

3.2 Hierarchical Image Encoder-decoder

Conventional ViT-based MIM backbones such as MAE or SimMIM utilize a simple decoder design with a few linear layers. However, Tian et al. propose that a hierarchical encoder-decoder design performed better when pretraining on CNNs [37]. To this end, we also adopt a multi-scale encoder-decoder backbone for our AnatoMask, which is well-suited to spatially aggregate feature maps. We find that this approach enables the progressive reconstruction of the masked image, effectively incorporating both local and global features. Given an input

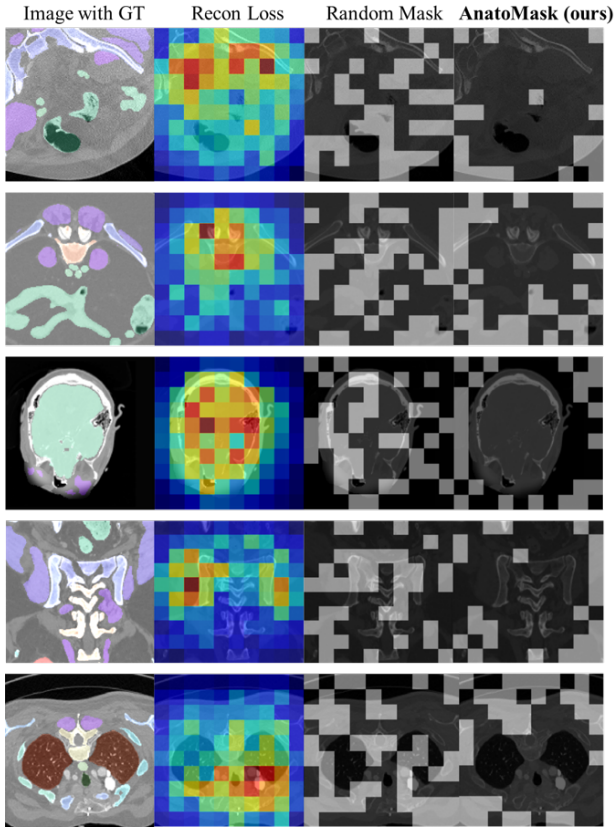


Fig. 3: Visual comparison between segmentation ground truths and reconstruction losses. For each row, we show a). an image with organ ground truths, b). reconstruction losses by averaging over 2 masks, c). a random mask (60%), and d). our AnatoMask obtained from b). Red means higher loss values while blue indicates lower ones. Red areas tend to overlap with organ regions. Transparent areas indicate unmasked regions.

volume V of $H \times W \times D$, we first divide V input patches of $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$ and randomly masked 60%. Then, we utilize STU-Net to downsample the masked input into 4 scales: $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$, $\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}$, $\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}$, $\frac{H}{16} \times \frac{W}{16} \times \frac{D}{16}$. Similar to SparK [37], we adopt the sparse submanifold convolution for our encoder to generate the multi-scale feature maps of the input. A hierarchical decoder is used to densify and upscale the feature maps for reconstruction. Concretely, let M^t_{final} denote the newly generated mask, the reconstruction loss can be formulated as:

$$\mathcal{L}^t_{rec} = \frac{1}{HWD} \sum (D_S(E_S(M^t_{final} \odot V)) - (1 - M^t_{final}) \odot V)^2, \tag{2}$$

where E_S and D_S represent the student network's encoder and decoder.

3.3 Reconstruction-guided Self-masking

While recent studies in MIM have demonstrated remarkable successes in natural images [16, 45], its effectiveness in medical imaging is worth further exploration due to the large domain gap between natural images and medical images. Conventional MIM takes a random masking approach, which ignores the anatomical region's heterogeneity by uniformly sampling all areas. Given the significant feature variations of human organs, a uniform random masking approach may not be an effective pretraining task for subsequent segmentation tasks. Thus, an effective MIM strategy should prioritize these complex anatomical regions to enhance diagnostic accuracy and efficacy. Intuitively, these complex regions are more difficult to reconstruct than background regions. We then take the regions with high reconstruction losses as the highly informative areas to guide the masking process. As demonstrated in Fig. 3, we find that anatomical regions often demonstrate high reconstruction losses, indicative of rich semantic information for the network to learn.

To identify such regions, a naive approach is to use a pretrained model as a teacher to extract high-loss areas from its predictions, which remains resource-intensive. Nevertheless, this frozen teacher cannot adaptively refine the representations of masked regions, which could hinder the student from learning more diverse features. Alternatively, we propose a self-distillation approach where the model first serves as a teacher to generate anatomically-significant masks and then acts as a student to perform self-supervised learning. In this way, we can train the model to self-mask anatomically informative regions, subsequently guiding the MIM toward more efficient pretraining.

Given an initial mask, our AnatoMask teacher can generate an initial reconstruction loss L^t_{rec}, from which we can employ anatomical priors in its masking process to enhance pretraining. We perform an *argsort* operation on L^t_{rec} to rank the masked regions in *descending* reconstruction losses. This ranking guides the model to focus on more challenging regions for reconstruction, which are more valuable for mining anatomical information. In the initial stages of training, high reconstruction loss may not always correlate with anatomical importance, as the model is still learning to reconstruct the image effectively. We propose a dynamic mask generation strategy, where the portion of anatomically significant regions in the masked regions gradually increases, providing some reasonable hints that guide the model to reconstruct anatomical regions step by step. By incrementally guiding the model, we ensure it learns effective representations, aligning more closely with anatomical significance as it advances in the training process. Concretely, we define a masking dynamics function based on current epoch t (with total epoch T):

$$r_t = r_0 + \frac{t}{T}(r_T - r_0), \tag{3}$$

where r_t represents anatomical-significance ratio of M^t_{final}. To retain the original masking ratio γ, the remaining $(1-r_t)$ of the masked regions in M^t_{final} are "filled" by a newly generated random mask. r_T, and r_0 are two hyperparameters which

we empirically set to 0.5 and 0 respectively. Given input volume V, masking ratio γ, and $M^t_{initial}$ being the initial mask fed into the teacher model, we select $r_t \cdot \gamma V$ voxels with the highest L^t_{rec} from $M^t_{initial}$ and then randomly generate $(1 - r_t) \cdot \gamma V$ voxels from $V - r_t \cdot \gamma V$ to form the final mask M^t_{final}

Algorithm 1 Algorithm for AnatoMask M^t_{final}

input: $G = (g_1, g_2, \ldots, g_n)$: All regions in the image.
$L_{rec} = (l_{g_1}, l_{g_2}, \ldots, l_{g_n})$: Reconstruction losses from the initial mask.
γ: Masking ratio.
r_t: Anatomical-significance ratio.
output: M^t_{final}: Mask with anatomically-significant regions.
$i \leftarrow 1$
for k in argsort(L_{rec}) : ▷ Sort regions based on loss
  $g'_i \leftarrow g_k$
  $i \leftarrow i + 1$
endfor
$p \leftarrow \lceil r_t \times n \rceil$ ▷ Compute selection threshold p
$M_{\text{top}} \leftarrow \{\}$
for i in $\{1, \ldots, p\}$: ▷ Select top p regions
  $M_{\text{top}} \leftarrow M_{\text{top}} \cup g'_i$
endfor
$n_{\text{random}} \leftarrow \lceil (\gamma - r_t) \times n \rceil$ ▷ Compute how many regions to randomly select.
$M_{\text{random}} \leftarrow \{\}$
for i in $\{1, \ldots, n_{\text{random}}\}$: ▷ Randomly sample n_{random} regions
  $M_{\text{random}} \leftarrow M_{\text{random}} \cup \text{sample}\ (G \setminus (M_{\text{top}} \cup M_{random}))$
endfor
$M^t_{\text{final}} \leftarrow M_{\text{top}} \cup M_{\text{random}}$
Return M^t_{final}

4 Result

4.1 Dataset and Evaluation

We conduct self-supervised pretraining for all methods on TotalSegmentator [40] dataset containing 1204 CT images with 104 anatomical structures (27 organs, 59 bones, 10 muscles, and 8 vessels). For preprocessing, all the images are resampled to $1.5 \times 1.5 \times 1.5\ mm^3$ and z-score normalization is used to rescale image intensity to zero mean and unit variance. Image patch size is set to $112 \times 112 \times 128$. We first use the following ratio to randomly split the dataset: 70% train, 10% validation, and 20% test. The training set is used for SSL pretraining. We then finetune the pretrained networks on the training set and use the validation set for early stops.

We further conduct extensive evaluations by finetuning on 3 public datasets to test the transferability of our models. We select FLARE22 (50 CT cases

Table 1: Ablation on the encoder-decoder design of AnatoMask.

Architecture	Masking	DSC (%)
Simple Decoder	Random	82.9
Simple Decoder	AnatoMask	83.4
Hierarchical Decoder	Random	83.6
Hierarchical Decoder	**AnatoMask**	**84.3**

of 13 organs) [26], AMOS22 (240 CT cases and 120 MR cases with 15 organ annotations) [22] and AutoPETII (398 PET/CT cases with lesion annotations) [11]. For each dataset, we utilize 70% train, 10% validation, and 20% test for evaluation.

For evaluation metrics, we calculated the Dice Coefficient (DSC) and Normalized Surface Dice (NSD) based on segmentation results.

4.2 Implementation Details

All experiments in this study are conducted based on Python 3.9, Ubuntu 22.04.3, Pytorch 2.0.1, and nnU-Net 2.2 [21]. For finetuning on TotalSegmentator, we utilize default data pre-processing and data augmentations in nnU-Net. For SSL pretraining, we follow the same data pre-processing as in TotalSegmentator but do not use additional data augmentations other than random cropping and random flipping. For finetuning on FLARE22, AMOS22, and AutoPETII, we use the nnUNet's default data pre-processing and data augmentation pipelines for each dataset. However, we use the same network configuration as in TotalSegmentator and in SSL pretraining. A more detailed description of hyperparameters is shown in the supplementary material. Models are trained on NVIDIA Tesla A100 cards with 80 GB VRAM.

4.3 Ablation Study

In this section, we investigate the decoder design, different masking strategies, and masking ratios of AnatoMask. We use STU-Net-B as the backbone with 1000 epochs for pretraining and finetuning on TotalSegmentator.

Decoder design. First, we verify the effectiveness of using a hierarchical decoder for AnatoMask. We compare the U-Net style architecture with a simple decoder architecture, which only uses two convolution blocks and does not receive hierarchical feature maps from the encoder. As shown in Tab. 1, a U-Net style encoder-decoder performed better than a simple decoder for MIM. Nevertheless, our AnatoMask still improves finetuning performance on a simple decoder design compared to random masking.

Effectiveness of Self-distillation. Then, we verify the importance of using self-distillation to maintain consistent mask generations between teacher and student. In our proposed approach, the teacher network is updated with student's

Table 2: Ablation study on the masking strategies of AnatoMask. We study the importance of our easy-to-hard masking dynamics, and self-distillation scheme and select the highest or lowest anatomically significant regions.

Self-distillation	Anatomical Significance	Masking Dynamics	DSC (%)
×	-	-	83.6
×	High	easy-to-hard	83.7
✓	**High**	**easy-to-hard**	**84.3**
✓	Low	easy-to-hard	83.6
✓	High	hard-to-easy	84.0

EMA weights at the end of each iteration. Without this, the teacher is simply a randomly initialized model without explicit knowledge on where to mask. As shown in Tab. 2 rows 2 and 3, self-distillation ensures that the teacher consistently generates meaningful MIM tasks for the student to solve.

Incorporating anatomically-significant regions. We study the effectiveness of reconstruction guidance to incorporate anatomically-significant regions. To obtain regions of high significance for masking, we apply *argsort* operations on the reconstruction losses in *descending order* to obtain the top r_t regions. Conversely, if we sort the reconstruction losses in *ascending* order, we then obtain regions with low significance, possibly making the final mask easier than a random mask to restore. As shown in Tab. 2 rows 3 and 4, we find that models did not benefit from learning low-significance regions, demonstrating the importance of incorporating anatomically rich regions.

Effectiveness of masking dynamics function. Additionally, we study the importance of our proposed masking dynamics function which controls the MIM difficulties over time. Specifically, the relationship between r_0 and r_T determines whether the pretraining difficulties increase or decrease with time. When $r_0 > r_T$, more anatomically complex components are included in the final mask, and fewer components are generated from random masking. This leads to an easy-to-hard objective which is more intuitive. Conversely, if $r_0 < r_T$, we obtain a hard-to-easy MIM objective. As shown in Tab. 2, we observe that an easy-to-hard masking strategy leads to better finetuning performance. This aligns with our hypothesis that starting with a low masking difficulty can offer valuable cues to the model and prevent training collapse.

Masking ratio γ. As shown in Tab. 3, we find that a lower masking ratio outperformed a higher masking ratio, different from traditional MIM methods. We hypothesize that a lower masking ratio leads to better preservation of the original anatomical information, resulting in more effective pretaining.

4.4 Comparison with previous SSL methods

We compare the finetuning performances of our method against previous SOTA SSL methods: 1). Tang et al. propose a combination of pretext tasks to self-supervise a transformer model SwinUNETR [35]; 2). Zhou et al. propose Self-

Table 3: Ablation study on masking ratio γ.

Masking ratio γ	DSC (%)
0.6	**84.3**
0.8	84.1
0.9	84.2

Table 4: Comparison of AnatoMask and existing SSL methods on TotalSegmentator. Data is shown as patient-wise mean ± SD. **Bold** means best method

Methods	Backbone	DSC (%)	NSD (%)
Tang et al. [35]	SwinUNETR	79.2 ± 11.2	75.3 ± 12.8
SelfMedMAE [49]	UNETR	78.8 ± 12.5	74.0 ± 15.0
UniMiSS [44]	MiTnet	82.2 ± 10.4	86.4 ± 13.0
MoCov2 [7]	STU-Net-B	81.8 ± 10.6	85.6 ± 12.3
BYOL [13]	STU-Net-B	78.8 ± 11.1	82.8 ± 12.7
SparK [37]	STU-Net-B	83.6 ± 10.5	86.7 ± 12.8
AnatoMask	STU-Net-B	$\mathbf{84.3 \pm 9.9}$	$\mathbf{87.8 \pm 11.8}$
AnatoMask	STU-Net-L	$\mathbf{87.8 \pm 9.8}$	$\mathbf{90.3 \pm 10.7}$
AnatoMask	STU-Net-H	$\mathbf{89.0 \pm 10.3}$	$\mathbf{91.9 \pm 10.0}$

MedMAE, a masked autoencoder method to pretrain transformer models [49]; 3). Xie et al. propose UniMiSS to extract 2D information from 3D volumes using self-distillation pretraining [44]; 4). He et al. propose MoCov2, a momentum contrastive method that overcomes the limitations of large batch size [7]; 5). Grill et al. propose BYOL, a contrastive method with dual-network to learn bootstrapped representations [13]; 6). Spark, a U-Net design for BERT-style pretraining of CNNs [37]. For network backbones, since certain methods are built on specific transformer networks, we maintain their backbones for fair comparison. For MoCov2 and BYOL, we use STU-Net-B as the backbone. For all methods, we fix pretraining epochs to 1000. The input patch size for SwinUNETR, UNETR, and MiTnet is set to (128, 128, 128) and the batch size is set to 4. As shown in Tab. 4, AnatoMask with STU-Net-B as backbone significantly outperforms all previous SSL methods. With 1000 epochs for pretraining, AnatoMask is able to achieve 84.3% DSC and 87.8% SDC, surpassing the best method (UniMiSS) by 2.1% and 1.4% respectively. Notably, UniMiSS extracts both 2D information and 3D information from medical images, whereas our AnatoMask focuses only on 3D information. This further demonstrates AnatoMask's superior efficiency and effectiveness in leveraging 3D data for segmentation tasks. We visualize segmentation results on TotalSegmentator CT images in Fig. 4. Our Anatomask demonstrates promising segmentation results compared with other SSL methods. To demonstrate the scalability of AnatoMask, we further test it on large segmentation backbones such as STU-Net-L and STU-Net-H (with 440 million and 1.5 billion parameters respectively) which improve performance to 87.8%

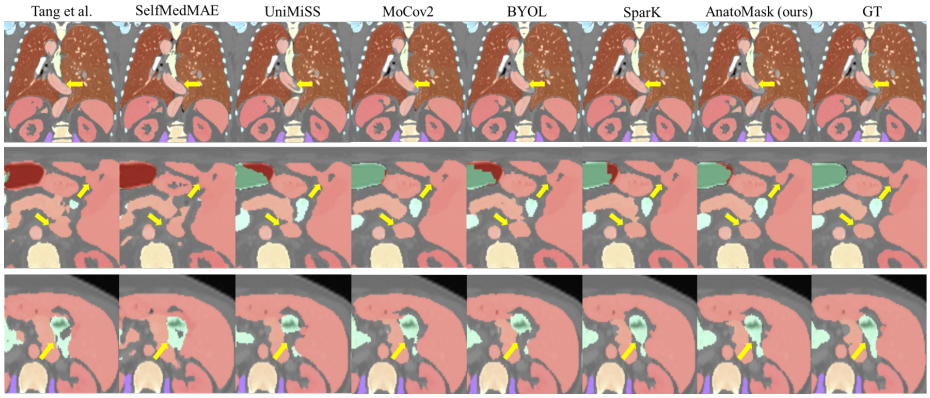


Fig. 4: Visualization of multi-organ CT segmentation results on TotalSegmentator. Yellow arrows indicate improvements in segmentation.

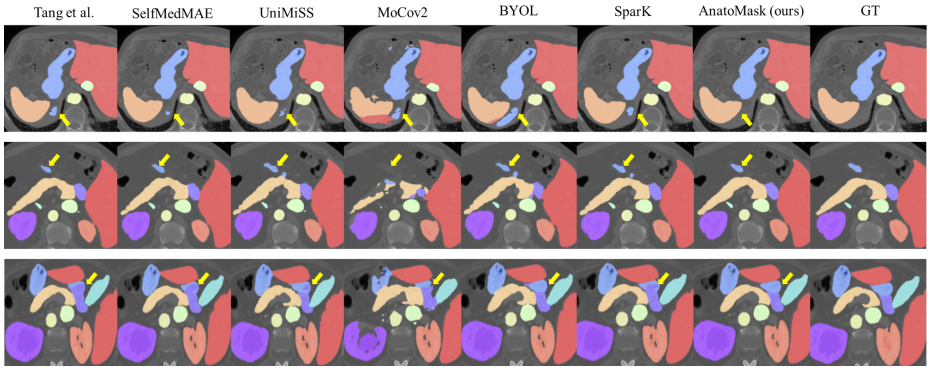


Fig. 5: Visualization of multi-organ CT segmentation results on FLARE22. Yellow arrows indicate improvements in segmentation.

DSC and 89.0% DSC. Our results indicate that AnatoMask is also beneficial for network parameter scaling, paving the way for self-supervising large foundation models for medical image segmentation.

Finally, we conduct transfer learning on three additional datasets by finetuning each model trained on TotalSegmentator. To demonstrate the transferability of our method, we choose three segmentation tasks: 1). multi-organ segmentation on CT (FLARE22); 2). cross-modality segmentation on MRI and CT (AMOS22 Task 2); 3). tumor segmentation on PET/CT (AutoPETII). We initialized all models with weights trained from TotalSegmentator and finetuned for 1000 epochs. As shown in Tab. 5 and Tab. 6, AnatoMask with STU-Net-B as backbone consistently outperforms comparable methods. We also visualize multi-organ segmentation results on CT in Fig. 5, cross-modality segmentation results (CT and MRI) in Fig. 6, and PET lesion segmentation results in Fig. 7.

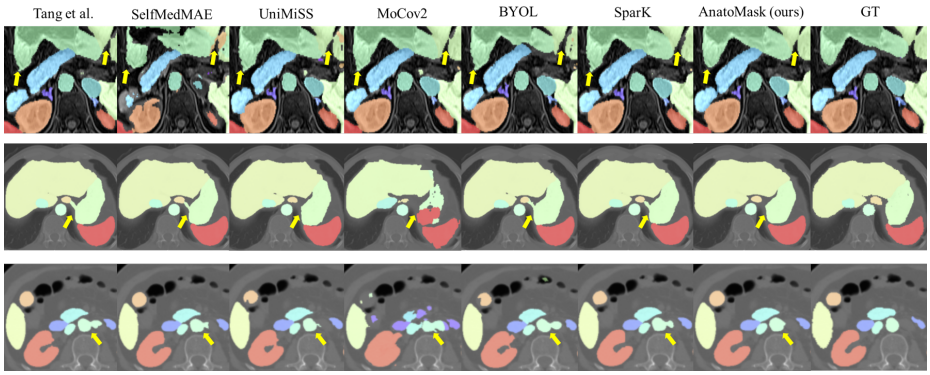


Fig. 6: Visualization of cross-modality organ segmentation on MRI (first row) and CT (second and third rows) from AMOS22. Yellow arrows indicate improvements in segmentation.

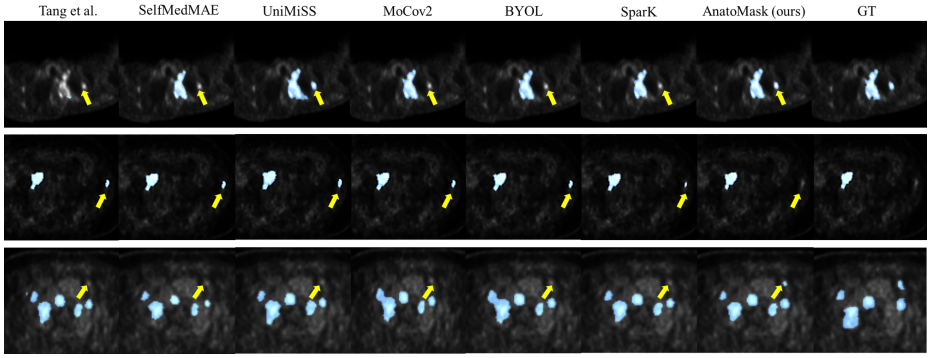


Fig. 7: Visualization of PET lesion segmentation results on AutoPETII. Yellow arrows indicate improvements in segmentation.

These results further demonstrate that our SSL strategy possesses promising transferability to other modalities, such as MRI and PET. When using larger backbones STU-Net-L and STU-Net-H, AnatoMask further improves segmentation performance on each task, demonstrating the scalability of our method.

5 Conclusion and Discussion

In this paper, we propose AnatoMask, a novel SSL approach to augment the existing MIM workflow for enhancing medical image segmentation. Compared to conventional random masking techniques, AnatoMask aims at mining anatomical information through an iterative masking process, guided by the reconstruction loss which signifies the anatomically significant regions. We find that AnatoMask not only augments the quality of representations learned but also

Table 5: Comparison of AnatoMask and existing SSL methods when transferring to FLARE22 and AMOS22. Data is shown as patient-wise mean ± SD. **Bold** means best method.

Methods	Backbone	FLARE22		AMOS22	
		DSC	NSD	DSC	NSD
Tang et al. [35]	SwinUNETR	91.2 ± 2.1	94.1 ± 2.6	85.4 ± 5.8	91.3 ± 6.1
SelfMedMAE [49]	UNETR	91.0 ± 3.4	93.7 ± 3.2	76.1 ± 7.7	86.7 ± 6.0
UniMiSS [44]	MiTnet	91.7 ± 2.2	95.2 ± 2.2	87.7 ± 5.6	93.7 ± 5.5
MoCov2 [7]	STU-Net-B	93.6 ± 1.3	97.0 ± 1.6	88.3 ± 5.3	94.1 ± 4.9
BYOL [13]	STU-Net-B	93.7 ± 1.2	97.0 ± 1.4	88.3 ± 5.4	94.1 ± 4.7
SparK [37]	STU-Net-B	93.8 ± 1.2	97.1 ± 1.2	88.5 ± 5.3	94.2 ± 4.8
AnatoMask	STU-Net-B	$\mathbf{94.0 \pm 1.3}$	$\mathbf{97.3 \pm 1.5}$	$\mathbf{89.1 \pm 5.0}$	$\mathbf{94.8 \pm 4.5}$
AnatoMask	STU-Net-L	$\mathbf{94.2 \pm 1.1}$	$\mathbf{97.3 \pm 1.3}$	$\mathbf{89.9 \pm 5.4}$	$\mathbf{95.6 \pm 4.7}$
AnatoMask	STU-Net-H	$\mathbf{94.3 \pm 1.5}$	$\mathbf{97.3 \pm 1.5}$	$\mathbf{89.9 \pm 5.2}$	$\mathbf{95.7 \pm 5.0}$

Table 6: Comparison of AnatoMask and existing SSL methods when transferring to AutoPETII. Data is shown as patient-wise mean ± SD. **Bold** means best method.

Methods	Backbone	AutoPETII	
		DSC	NSD
Tang et al. [35]	SwinUNETR	72.3 ± 23.7	80.5 ± 22.8
SelfMedMAE [49]	UNETR	67.0 ± 23.9	74.8 ± 24.3
UniMiSS [44]	MiTnet	64.0 ± 23.7	75.0 ± 23.0
MoCov2 [7]	STU-Net-B	72.9 ± 21.3	81.2 ± 21.6
BYOL [13]	STU-Net-B	74.1 ± 21.1	82.2 ± 21.7
SparK [37]	STU-Net-B	75.3 ± 21.1	84.5 ± 21.4
AnatoMask	STU-Net-B	$\mathbf{75.8 \pm 19.7}$	$\mathbf{85.3 \pm 19.7}$
AnatoMask	STU-Net-L	$\mathbf{77.5 \pm 19.0}$	$\mathbf{85.9 \pm 18.7}$
AnatoMask	STU-Net-H	$\mathbf{78.1 \pm 19.7}$	$\mathbf{87.0 \pm 19.8}$

boosts data efficiency in the subsequent finetuning phase. Our experiments on 4 public datasets demonstrate that AnatoMask consistently improves the segmentation performance of different organs and imaging modalities. Our results also show strong scalability with larger segmentation backbones.

However, our study still has some limitations. We only conduct SSL pretraining on CT images. Given the availability of a vast array of unlabeled MR images, we plan to expand our SSL approach to this modality. Also, our evaluation focuses on organ segmentation in this paper. However, given the encouraging transferability observed in PET lesion segmentation, our future study will explore the development of foundational models for more comprehensive tumor segmentation tasks in CT and MR images.

Acknowledgements

This research is supported in part by the National Institutes of Health under Award Number R01CA272991, R56EB033332 and R01EB032680.

References

1. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al.: Big self-supervised models advance medical image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3478–3488 (2021) 3
2. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. In: International Conference on Machine Learning. pp. 1298–1312. PMLR (2022) 3
3. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021) 2, 4
4. Chaitanya, K., Erdil, E., Karani, N., Konukoglu, E.: Contrastive learning of global and local features for medical image segmentation with limited annotations. Advances in neural information processing systems **33**, 12546–12558 (2020) 3
5. Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., Rueckert, D.: Self-supervised learning for medical image analysis using image context restoration. Medical image analysis **58**, 101539 (2019) 4
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 3
7. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 3, 11, 14
8. 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) 2, 4
9. Chen, Z., Du, Y., Hu, J., Liu, Y., Li, G., Wan, X., Chang, T.H.: Multi-modal masked autoencoders for medical vision-and-language pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 679–689. Springer (2022) 2
10. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE international conference on computer vision. pp. 1422–1430 (2015) 3
11. Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., Rubin, D.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. Scientific Data **9**(1), 601 (2022) 2, 9
12. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018) 3
13. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020) 3, 11, 14

14. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 297–304. JMLR Workshop and Conference Proceedings (2010) 3
15. Haghighi, F., Taher, M.R.H., Gotway, M.B., Liang, J.: Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20824–20834 (2022) 4
16. 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) 1, 2, 4, 7
17. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020) 3
18. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 (2018) 3
19. Huang, Z., Jin, X., Lu, C., Hou, Q., Cheng, M.M., Fu, D., Shen, X., Feng, J.: Contrastive masked autoencoders are stronger vision learners. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 4
20. Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., et al.: Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716 (2023) 2
21. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021) 9
22. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022) 2, 9
23. Jiang, Y., Sun, M., Guo, H., Bai, X., Yan, K., Lu, L., Xu, M.: Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15859–15869 (2023) 1
24. Kakogeorgiou, I., Gidaris, S., Psomas, B., Avrithis, Y., Bursuc, A., Karantzalos, K., Komodakis, N.: What to hide from your students: Attention-guided masked image modeling. In: European Conference on Computer Vision. pp. 300–318. Springer (2022) 4
25. Li, G., Zheng, H., Liu, D., Wang, C., Su, B., Zheng, C.: Semmae: Semantic-guided masking for learning masked autoencoders. Advances in Neural Information Processing Systems **35**, 14290–14302 (2022) 4
26. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., et al.: Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. Medical Image Analysis **82**, 102616 (2022) 2, 9
27. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6707–6717 (2020) 3
28. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. pp. 69–84. Springer (2016) 3

29. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 3
30. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544 (2016) 4
31. Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F.: Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint arXiv:2208.06366 (2022) 4
32. Shi, Y., Siddharth, N., Torr, P., Kosiorek, A.R.: Adversarial masking for self-supervised learning. In: International Conference on Machine Learning. pp. 20026–20040. PMLR (2022) 4
33. Taleb, A., Lippert, C., Klein, T., Nabi, M.: Multimodal self-supervised learning for medical image analysis. In: International conference on information processing in medical imaging. pp. 661–673. Springer (2021) 3
34. Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C.: 3d self-supervised methods for medical imaging. Advances in neural information processing systems **33**, 18158–18172 (2020) 4
35. 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) 1, 3, 4, 10, 11, 14
36. Tao, C., Zhu, X., Su, W., Huang, G., Li, B., Zhou, J., Qiao, Y., Wang, X., Dai, J.: Siamese image modeling for self-supervised vision representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2132–2141 (2023) 4
37. Tian, K., Jiang, Y., Diao, Q., Lin, C., Wang, L., Yuan, Z.: Designing bert for convolutional networks: Sparse and hierarchical masked modeling. arXiv preprint arXiv:2301.03580 (2023) 2, 4, 5, 6, 11, 14
38. Wang, H., Song, K., Fan, J., Wang, Y., Xie, J., Zhang, Z.: Hard patches mining for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10375–10385 (2023) 4
39. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: Proceedings of the IEEE international conference on computer vision. pp. 2794–2802 (2015) 3
40. 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) 2, 8
41. 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) 4
42. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16133–16142 (2023) 4
43. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018) 3
44. Xie, Y., Zhang, J., Xia, Y., Wu, Q.: Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In: European Conference on Computer Vision. pp. 558–575. Springer (2022) 4, 11, 14

45. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: 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) 2, 4, 7
46. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Wei, Y., Dai, Q., Hu, H.: On data scaling in masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10365–10374 (2023) 2, 4
47. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. pp. 649–666. Springer (2016) 3
48. Zhou, H.Y., Lu, C., Yang, S., Han, X., Yu, Y.: Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3499–3509 (2021) 1
49. Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., Prasanna, P.: Self pre-training with masked autoencoders for medical image classification and segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). pp. 1–6. IEEE (2023) 2, 11, 14
50. Zhu, J., Li, Y., Hu, Y., Ma, K., Zhou, S.K., Zheng, Y.: Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. Medical image analysis **64**, 101746 (2020) 1