

Article

ARM: Active Region Masking for 3D Medical Image Analysis

Can Wang and Zesheng Cheng *

School of Computer Science Technology, Qingdao University, Qingdao 266071, China; wangcan@qdu.edu.cn

* Correspondence: czs_110@hotmail.com

Featured Application

This work provides an adaptive self-supervised pre-training method whose learned encoder can initialize downstream Swin-UNETR segmentation models and support label-efficient 3D medical image segmentation, disease classification, and cross-modal medical image analysis.

Abstract

Self-supervised learning can reduce the dependence of 3D medical image analysis on expensive voxel-level annotations, but masked image modeling remains inefficient when uniform random masking is applied to volumetric data with large redundant backgrounds. Random masks often generate easy reconstruction targets that can be recovered through local interpolation, limiting the learning of anatomical semantics. Therefore, an effective masking strategy should adapt to image-specific structural difficulty rather than sample regions uniformly. We propose Active Region Masking (ARM), a self-supervised pre-training method that treats 3D mask generation as a patch-wise actor–critic decision process. Patch-level decision units share an actor–critic policy and use reconstruction error with a masking-ratio constraint as an intrinsic reward to identify regions that are difficult to reconstruct and potentially informative for anatomical representation learning. The reconstructor is trained with an asymmetric 3D Swin Transformer encoder–decoder, encouraging global anatomical reasoning from visible context. For segmentation tasks, the ARM pre-trained encoder is used to initialize the downstream Swin-UNETR framework; classification is evaluated with a matched downstream protocol. Across 12 task-level downstream evaluations, including BTCV, MSD, MM-WHS, AMOS22, FLARE22, CC-CCII, and BraTS21, ARM consistently improves over the evaluated contrastive and masked-modeling baselines. With 1k scans as pre-training data, ARM achieves an average Dice score of 89.80% on the BTCV-and-unseen-dataset benchmark, and scaling to 10k scans increases the average Dice score to 90.66%. These results indicate that active region masking improves label efficiency, segmentation robustness, and CT-to-MRI transfer in 3D medical image analysis.

Keywords: self-supervised learning; masked image modeling; medical image analysis; reinforcement learning; 3D image segmentation



Academic Editor: Thomas Lindner

Received: 9 May 2026

Revised: 7 June 2026

Accepted: 12 June 2026

Published: 29 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Deep learning has become central to medical image analysis, including organ segmentation, lesion delineation, and disease classification. However, training reliable 3D models still depends heavily on dense manual annotations, which are expensive, time-consuming, and difficult to standardize across institutions and imaging protocols. This annotation bottleneck is particularly severe for volumetric CT and MRI because expert labels must be produced at the voxel level. Self-supervised learning (SSL) provides a practical route

to reduce this dependence by pre-training representation models on large collections of unlabeled scans before fine-tuning them on downstream tasks [1].

Among SSL paradigms, masked image modeling (MIM) has attracted increasing attention because it trains a network to infer missing image content from visible context [2,3]. In 3D medical imaging, however, directly transferring random masking strategies from natural-image pre-training is suboptimal. Medical volumes contain large regions of homogeneous tissue and background, while clinically important structures, such as small organs, lesions, and anatomical boundaries, occupy only a limited part of the volume. As a result, uniform random masking can select many low-information patches and produce reconstruction targets that are solvable through local interpolation. This weakens the pressure on the encoder to learn long-range anatomical dependencies and structure-aware semantics. In addition, slice-wise or spatially uncoordinated masks can disrupt volumetric continuity, which is important for dense 3D prediction.

Existing medical SSL methods address this problem only partially. Contrastive approaches improve global semantic consistency but may underemphasize voxel-level details required by segmentation [4,5]. MIM-based methods, including 3D Swin Transformer pre-training and masked autoencoder variants, provide stronger dense representations but still often rely on random or hand-designed masking policies [6–9]. Deep reinforcement learning (DRL) has shown promise in medical image analysis for sequential anatomical decision-making tasks, such as landmark detection [10,11], but most applications rely on supervised rewards and are therefore not directly scalable to unlabeled pre-training. These limitations suggest the need for a self-supervised masking strategy that can adapt to image-specific structural difficulty without requiring manual annotations.

To address these limitations, we propose Active Region Masking (ARM), a self-supervised pre-training method for 3D medical image analysis. ARM recasts mask generation as a patch-wise actor–critic masking problem, where parameter-shared patch-level decision units operate over volumetric patches and learn to select informative regions for masking. The decision units receive intrinsic rewards from reconstruction error and a target mask-ratio regularization term, while the reconstructor learns to recover masked content from global anatomical context. This adversarial interaction turns pre-training from passive random filling into active hard-region mining, encouraging the encoder to model spatial dependencies that are useful for segmentation, classification, and cross-modal transfer.

The main contributions of this work are summarized as follows:

- We introduce an RL-driven adaptive masking framework for 3D medical MIM. The masking policy is learned with shared actor–critic parameters and intrinsic reconstruction rewards, reducing the inefficiency caused by redundant background regions.
- We design a volumetric active masking mechanism that targets hard-to-reconstruct regions that may contain informative anatomical structures while preserving 3D contextual continuity, thereby encouraging structure-aware representation learning.
- We evaluate ARM on 12 task-level downstream evaluations covering multi-organ segmentation, lesion-related segmentation, disease classification, and CT-to-MRI transfer, using the ARM pre-trained encoder to initialize downstream Swin-UNETR for segmentation. The results show consistent gains over the evaluated contrastive and masked-modeling baselines, especially in data-scarce settings.

2. Related Work

Self-supervised learning (SSL) has become an important strategy for reducing annotation dependence in 3D medical image analysis. Existing methods can be broadly grouped into contrastive and reconstruction-based paradigms, each of which addresses representation learning from a different perspective. However, contrastive methods may learn representations that are too coarse for dense prediction, whereas masked image modeling (MIM) methods can be inefficient when masks are generated without considering the structural distribution of medical volumes. ARM is motivated by these limitations and introduces an intrinsic-reward-based adaptive masking policy to make reconstruction-based pre-training more structure-aware.

2.1. Self-Supervised Learning for Medical Image Analysis

Self-supervised medical image learning aims to exploit unlabeled data to improve downstream performance when annotations are scarce. Contrastive learning frameworks, such as PCRL [5] and global–local contrastive learning [4], encourage invariant or aligned representations across augmented views. These methods can capture semantic consistency but may overlook fine-grained local structures required for dense segmentation. Reconstruction-based SSL methods, especially MIM approaches such as SimMIM [3], local multi-scale reconstruction [7], SwinMM [8], global–local masked autoencoders [9], and mask-in-mask pre-training [12], train models to recover missing image content from visible context. Recent volume-level contrastive learning methods, such as VoCo [13], also show that large-scale 3D pre-training benefits from volumetric structural cues. However, when random masking is applied to volumes with large low-information regions, the reconstruction objective can be dominated by easy patches. ARM addresses this issue by integrating reinforcement learning into the generative objective, enabling image-specific masking policies that preferentially select hard-to-reconstruct regions that may contain informative anatomical structures.

2.2. Deep Reinforcement Learning in Medical Imaging

Deep reinforcement learning (DRL) has been explored for sequential decision-making problems in medical imaging, including anatomical navigation and landmark localization [10,11]. These applications usually require task-specific supervised rewards based on expert annotations, which limits their use in large-scale unlabeled pre-training. ARM differs from these methods by using reconstruction error as an intrinsic reward. This design allows the masking policy to be optimized without manual labels and makes DRL compatible with self-supervised representation learning.

3. Materials and Methods

This section describes the proposed 3D Active Region Masking (ARM) framework. Figure 1 compares the proposed adaptive masking behavior with uniform random masking, and Figure 2 illustrates the overall framework. ARM reformulates mask generation as a collaborative minimax game between a masking policy and a reconstruction network. The masking policy learns to assign higher masking probabilities to hard-to-reconstruct regions, while the reconstructor learns to recover the missing content from visible 3D context.

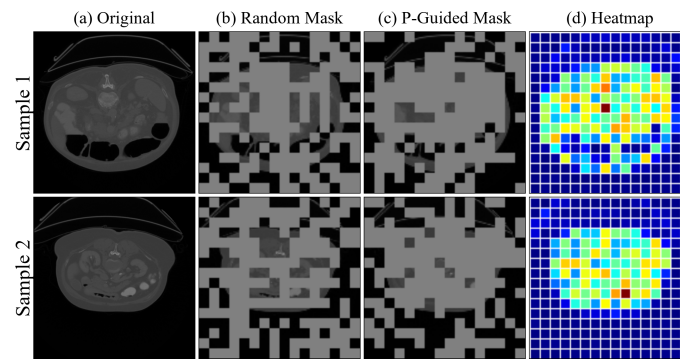


Figure 1. Comparison of masking strategies on two representative volumetric medical samples (rows). Columns: (a) original image; (b) uniform random mask; (c) the ARM policy mask; and (d) the per-patch difficulty heatmap. At a 0.7 masking ratio, ARM concentrates masking probability on high-uncertainty regions across both dense structures and soft tissues, encouraging global anatomical reasoning rather than local interpolation.

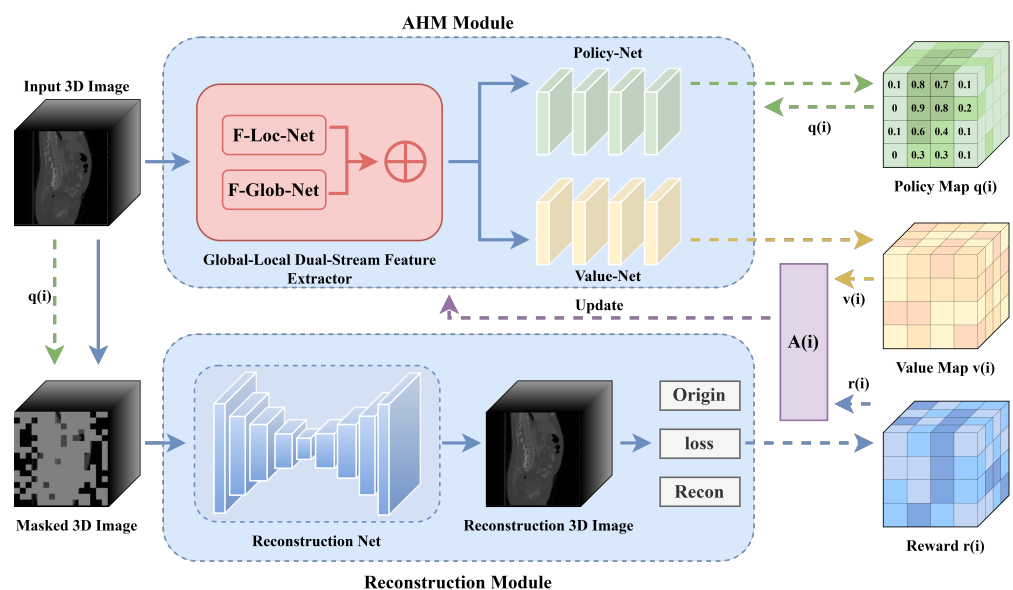


Figure 2. Overview of the proposed ARM framework. The system contains two modules: (1) Adaptive Hard Masking (AHM), the ARM masking submodule that uses a shared actor–critic policy to predict masking maps and (2) a reconstruction module based on a 3D Swin Transformer. Solid arrows indicate the forward data and feature flow, dashed arrows indicate reward or optimization feedback, and different arrow colors distinguish the masking-policy branch from the reconstruction branch. The adversarial training strategy increases reconstruction difficulty by selecting hard-to-reconstruct patches.

3.1. Problem Formulation: Adversarial Masking Game

We model self-supervised pre-training as a minimax game between a masking policy π_θ and a reconstructor G_ϕ . Given a 3D volume $X \in \mathbb{R}^{D \times H \times W}$ partitioned into N patches $\mathcal{B} = \{b_i\}_{i=1}^N$, the policy generates a binary mask $m \in \{0, 1\}^N$ to produce the masked view $\tilde{X} = X \odot (1 - m)$. The objective is

$$\min_{\phi} \max_{\theta} \mathbb{E}_{m \sim \pi_\theta} \left[\mathcal{L}_{\text{rec}}(G_\phi(\tilde{X}), X) - \lambda_{\text{reg}} \left| \frac{1}{N} \sum_{i=1}^N m_i - \rho \right| \right], \quad (1)$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruction loss, ρ is the target masking ratio, and the regularization term constrains the generated mask density. This formulation encourages G_ϕ to learn robust representations under challenging but ratio-controlled masking patterns.

3.2. Adaptive Hard Masking Module

The Adaptive Hard Masking (AHM) module formulates mask generation as a patch-wise reinforcement learning problem. Each patch b_i is treated as a decision unit that observes a local state vector s_i , takes a binary action $m_i \in \{0, 1\}$ indicating whether the patch is masked, and receives the intrinsic reward defined in Equation (5). All decision units share the same actor–critic parameters to maintain translation invariance and parameter efficiency. They do not exchange explicit messages; instead, coordination is induced by the shared global–local state representation and by the volume-level mask-ratio term in the reward. Thus, ARM is best interpreted as a parameter-shared patch-wise actor–critic formulation rather than a setting with independently parameterized communicating agents.

3.2.1. Global–Local Dual-Stream Feature Extractor

To capture both fine-grained texture and global anatomical context, we design a dual-stream encoder that maps the input volume X to a state representation S . The local stream F_{loc} uses a truncated 3D ResNet-18 to extract fine-grained local features at $8\times$ downsampling. The global stream F_{glob} uses cascaded pooling and large-kernel convolutions to model long-range dependencies. The two streams are fused with 3D sinusoidal positional encoding P_{pos} through a $1 \times 1 \times 1$ projection:

$$S = \sigma\left(\text{Conv}_{1 \times 1 \times 1}\left([F_{\text{loc}}(X); F_{\text{glob}}(X)]\right) + P_{\text{pos}}\right). \quad (2)$$

The resulting state s_i integrates local details and global structural priors, improving policy stability under partial observations.

3.2.2. Policy and Value Networks

Based on s_i , a parameter-shared actor–critic module with two parallel multilayer perceptrons jointly generates the masking policy and estimates state values. The actor π_θ outputs a confidence score q_i , representing the probability that the i -th patch is selected for masking:

$$q_i = \pi_\theta(m_i = 1 \mid s_i) = \sigma\left(W_\pi^{(2)} \cdot \text{ReLU}\left(W_\pi^{(1)} s_i\right)\right). \quad (3)$$

The critic V_ψ predicts the expected reconstruction difficulty, denoted as the state value v_i :

$$v_i = V_\psi(s_i) = W_v^{(2)} \cdot \text{ReLU}\left(W_v^{(1)} s_i\right). \quad (4)$$

Here, W_* denotes learnable parameters. To reduce policy-gradient variance, we compute the advantage function using v_i as a baseline: $A_i(s_i, m_i) = r_i - v_i$. The binary mask is sampled as $m_i \sim \text{Bernoulli}(q_i)$ and supplemented by an ϵ -greedy strategy to balance exploration and exploitation.

3.3. Asymmetric Masked Reconstruction

To reduce the computational burden of 3D voxel processing and avoid shortcut learning, we adopt an asymmetric masked-autoencoder-style encoder–decoder architecture. The 3D Swin Transformer encoder [6] operates only on the visible patch set $\mathcal{B}_{\text{vis}} = \{b_i \mid m_i = 0\}$, reducing the attention complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}((N(1 - \rho))^2)$ under a target masking ratio ρ . The encoded features are concatenated with learnable mask tokens to restore the original sequence length and are passed to a lightweight three-layer Transformer decoder. This heavy-encoder and light-decoder design imposes an information bottleneck, encouraging the encoder to learn global anatomical structure rather than relying on decoder-based texture interpolation.

3.4. Optimization Objectives

ARM stabilizes the adversarial interaction between the masking policy and reconstruction network through alternating optimization with intrinsic rewards.

3.4.1. Intrinsic Reward Design

To guide the policy toward hard-to-reconstruct patches, the intrinsic reward r_i is defined using reconstruction error and a mask-ratio constraint:

$$r_i = \left\| b_i - \hat{b}_i \right\|_2^2 - \lambda_{\text{reg}} \left| \frac{1}{N} \sum_{j=1}^N m_j - \rho \right|. \quad (5)$$

The first term assigns higher rewards to patches that are difficult to reconstruct, while the second term encourages the overall mask density to remain close to the target ratio.

3.4.2. Iterative Minimax Optimization

To avoid instability during joint optimization, we use a two-stage alternating update. In the reconstruction phase, θ is fixed and the reconstructor G_ϕ is updated by minimizing the masked reconstruction error:

$$\phi \leftarrow \phi - \eta_\phi \nabla_\phi \mathbb{E}_{m \sim \pi_\theta} \left[\sum_{i=1}^N m_i \cdot \left\| b_i - (G_\phi(\tilde{X}))_i \right\|_2^2 \right]. \quad (6)$$

In the policy phase, ϕ is fixed and the masking policy is updated through A3C [14] to maximize $J(\theta)$:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^N (\nabla_\theta \log \pi_\theta(m_i | s_i) \cdot A_i + \beta \nabla_\theta H(\pi_\theta(\cdot | s_i))). \quad (7)$$

This procedure forms a closed-loop curriculum: as the reconstructor improves, the masking policy searches for increasingly difficult patches, and stronger policies push the reconstructor toward more robust representation learning.

3.5. Loss Function Configuration

For the reconstruction network, we use masked mean squared error [15] so that learning focuses on masked regions:

$$\mathcal{L}_{\text{rec}} = \frac{1}{\sum_{i=1}^N m_i + \delta} \sum_{i=1}^N m_i \cdot \left\| b_i - (G_\phi(\tilde{X}))_i \right\|_2^2, \quad (8)$$

where δ is a small constant for numerical stability. For the policy update, the objective combines policy gradient, value regression, and entropy regularization:

$$\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{policy}} + \lambda_{\text{val}} \mathcal{L}_{\text{value}} + \lambda_{\text{ent}} \mathcal{L}_{\text{entropy}}. \quad (9)$$

Here, $\mathcal{L}_{\text{policy}}$ uses the advantage $A_i = r_i - v_i$ to maximize expected rewards, $\mathcal{L}_{\text{value}}$ minimizes critic regression error, and $\mathcal{L}_{\text{entropy}}$ encourages diversity to avoid premature policy collapse. We fix $\lambda_{\text{val}} = 0.5$ and $\lambda_{\text{ent}} = 0.01$ for all experiments. The value-loss coefficient follows the common actor–critic practice of keeping the critic sufficiently weighted to stabilize advantage estimates, whereas the entropy coefficient is kept small so that exploration improves mask diversity without dominating the reconstruction-error reward.

4. Experiments and Results

This section describes the datasets, implementation settings, baseline methods, downstream results, and ablation studies used to evaluate ARM.

4.1. Datasets and Evaluation Metrics

Pre-training uses two cohorts. ARM-1k consists of BTCV and TCIA COVID-19 scans, while ARM-10k aggregates 10,502 CT scans from eight datasets following the large-scale pre-training setting in prior work [13]. Downstream validation covers 12 task-level evaluations: BTCV, MM-WHS, AMOS22, FLARE22, six task-specific subsets of MSD (Spleen, Liver, Lung, Pancreas, Hepatic Vessels, and Colon), CC-CCII classification, and cross-modal CT-to-MRI transfer on BraTS 21. We follow official splits when available, and the non-BTCV datasets are not used during pre-training, providing an out-of-distribution evaluation of the pre-trained representations. Dice similarity coefficient (DSC) and normalized surface Dice (NSD) are reported for segmentation, while accuracy (ACC) and area under the receiver operating characteristic curve (AUC) are reported for classification [12].

4.2. Implementation Details

For pre-training, input volumes are reoriented to a canonical RAS orientation, clipped to $[-1000, 1000]$ HU, normalized to $[0, 1]$, and resampled to $128 \times 128 \times 128$. The architecture consists of a 3D Swin Transformer encoder and a 3D ResNet policy encoder with dimension 64, followed by two-layer MLP actor–critic heads with dimension 32. We use alternating optimization: AdamW with a learning rate of 1×10^{-4} for reconstruction and Adam with a learning rate of 5×10^{-4} for the masking policy. Training lasts 1600 epochs on eight NVIDIA A100 GPUs with cosine annealing, a target masking ratio of 0.7, and $\epsilon = 0.3$. For downstream segmentation tasks, the ARM pre-trained Swin Transformer encoder initializes Swin-UNETR; for classification, the pre-trained encoder is evaluated by linear probing under the matched downstream protocol. Inference uses an overlapping sliding-window strategy without post-processing.

4.3. Baseline Comparisons

We compare ARM with two categories of representative SSL baselines. The first category includes general SSL methods, including SimCLR [16], MoCo v3 [17], localMIM [7], and MCMAE [18]. The second category includes methods that are widely used in 3D medical self-supervised learning, including SimMIM [3], MAE3D [19], GL-MAE [9], SwinMM [8], GVSL [20], and Swin-UNETR pre-training [6]. We also cite Swin-Unet as a representative transformer-based medical segmentation architecture [21]. Recent large-scale 3D SSL methods, including VoCo [13] and MiM [12], are also discussed as current reference points for volumetric pre-training. For re-implemented baselines, we use the same input resolution, backbone family, pre-training data, fine-tuning split, optimizer schedule, and training budget as ARM whenever the method design permits. Hyperparameters are taken from the original papers when available and otherwise tuned under the same validation protocol used for ARM. A compact comparison of these representative methods is provided in Table 1.

Table 1. Comparison with recent 3D self-supervised pre-training methods on representative downstream tasks. Bold rows denote the proposed ARM variants.

Method	Year	PT	BTCV Dice	MM-WHS Dice	BraTS21 Avg.	CC-CCII ACC
VoCo * [13]	2024	1.6k	83.85	90.54	78.53	90.83
MiM * [12]	2025	1k	84.66	91.04	79.28	93.63
ARM (Ours)	2026	1k	83.68	90.68	79.62	92.01
ARM (Ours)	2026	10k	84.65	91.15	80.25	92.55

* Officially reported results from the corresponding methods. PT: pre-training dataset size.

4.4. Results on Downstream Tasks

4.4.1. BTCV and Unseen-Dataset Benchmark

Table 2 compares methods across five segmentation tasks. ARM improves the average Dice score by replacing random masking with adaptive hard-region mining. With 1k scans as pre-training data, ARM reaches an average Dice score of 89.80%, outperforming SimCLR by 5.47 percentage points and improving over strong MIM baselines such as MAE3D and GL-MAE. Scaling to 10k scans as pre-training scans increases the average Dice score to 90.66%. The qualitative examples in Figure 3 show clearer boundaries and more complete anatomical structures.

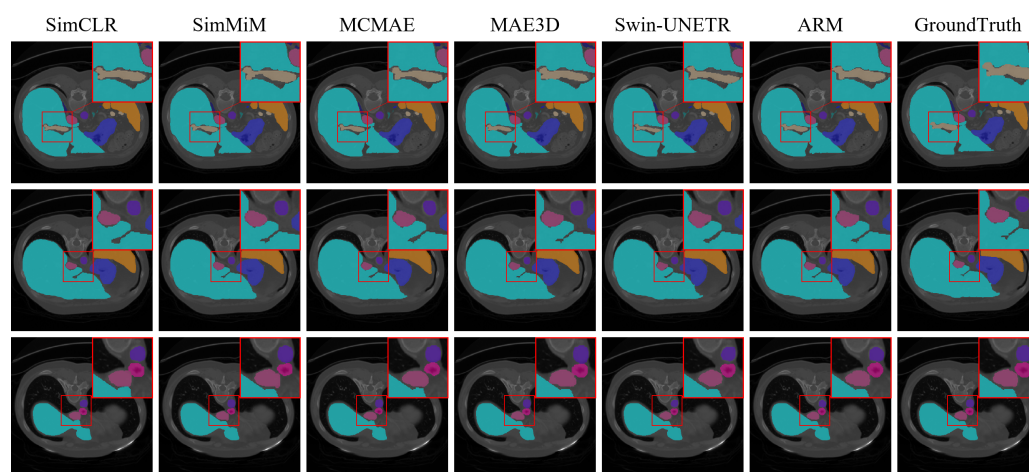


Figure 3. Qualitative comparison of segmentation results. For readability, representative baselines are shown from the full quantitative comparison. The top row shows small organs, the middle row shows large organs, and the bottom row highlights organ boundaries.

Table 2. Comparison with representative methods on BTCV and unseen datasets (Dice score, %). Bold rows denote the proposed ARM variants.

Method	PT	BTCV	MM-WHS	MSD Sple.	AMOS22	FLARE22	Avg.
SimCLR	1k	79.85	86.00	94.11	84.66	77.03	84.33
SimMIM	1k	81.41	90.22	94.94	89.74	86.05	88.47
MCMAE [†]	1k	82.18	90.32	96.08	86.60	84.35	87.91
SwinMM [†]	1k	81.86	90.44	95.52	88.80	87.66	88.86
GL-MAE	1k	82.33	88.88	95.72	87.40	87.26	88.32
MAE3D	1k	82.48	90.03	95.50	89.71	87.31	89.01
Swin-UNETR [†]	1k	82.62	90.35	95.05	89.80	87.20	89.00
ARM (Ours)	1k	83.68	90.68	96.45	90.22	87.98	89.80
Swin-UNETR [†]	10k	83.25	90.50	96.22	89.92	88.08	89.59
MAE3D [†]	10k	83.55	90.72	96.95	88.52	88.52	89.65
ARM (Ours)	10k	84.65	91.15	97.05	91.25	89.20	90.66

[†] Experiments re-implemented by us. PT: pre-training dataset size.

4.4.2. MSD Multi-Organ Segmentation

Table 3 evaluates five MSD tasks under the unified 1k pre-training protocol. ARM achieves an average DSC of 69.3% and an average NSD of 77.9%, outperforming the Swin-UNETR baseline by 2.1 percentage points in both metrics. The gains are prominent for challenging structures with ambiguous boundaries, such as pancreas and colon. Scaling ARM to 10k pre-training data further improves the average DSC to 70.1% and the average NSD to 78.7%.

Table 3. Performance on MSD multi-organ segmentation. D: DSC (%); N: NSD (%). Bold rows denote the proposed ARM variants.

Method	PT	Liver		Lung		Panc.		Hepatic		Colon		Avg.	
		D	N	D	N	D	N	D	N	D	N	D	N
MoCo v3	1k	94.1	93.2	52.1	48.1	74.1	78.1	58.1	78.1	43.9	55.1	64.9	72.5
SimMIM	1k	94.5	92.4	56.9	52.4	74.9	79.7	60.4	78.9	42.4	54.3	66.2	73.6
MCMAE	1k	95.4	94.3	57.0	57.1	75.7	78.1	60.4	78.9	45.6	55.1	67.2	74.9
GL-MAE	1k	94.6	93.4	53.2	49.5	72.9	87.9	60.2	78.6	45.5	59.4	65.5	74.2
MAE3D	1k	95.5	96.6	58.3	52.3	77.3	91.9	60.4	78.8	45.4	56.0	66.7	75.3
GVSL	1k	95.6	95.7	59.4	54.5	77.8	92.6	57.7	77.9	50.5	61.4	68.2	76.7
Swin-UNETR	1k	95.5	96.6	60.2	57.2	76.9	90.8	60.4	79.0	43.2	55.4	67.2	75.8
ARM (Ours)	1k	95.8	96.8	60.8	57.9	78.2	92.8	60.8	79.5	50.9	62.5	69.3	77.9
ARM (Ours)	10k	96.1	97.2	61.5	58.8	78.8	93.1	61.3	80.0	52.6	64.2	70.1	78.7

4.4.3. CC-CCII Disease Classification

To test cross-paradigm transfer from segmentation-oriented pre-training to classification, we evaluate ARM on CC-CCII. As shown in Table 4, ARM with 1k pre-training scans achieves 92.01% ACC and 99.15% AUC, outperforming both contrastive and generative baselines. With 10k pre-training scans, ARM reaches 92.55% ACC and 99.45% AUC, indicating that the learned features preserve not only local anatomical structures but also disease-relevant global semantics.

Table 4. Experiment results on the CC-CCII dataset. Bold rows denote the proposed ARM variants.

Method	Pre-Training Dataset	ACC (%)	AUC
SimCLR	1k	87.12	96.21
MoCo v3	1k	87.01	95.49
SimMIM	1k	89.62	98.16
MCMAE	1k	90.26	97.12
GL-MAE	1k	88.00	96.97
MAE3D	1k	91.30	98.13
SwinMM	1k	89.99	98.77
Swin-UNETR	1k	91.81	98.97
ARM (Ours)	1k	92.01	99.15
ARM (Ours)	10k	92.55	99.45

4.4.4. BraTS 21 CT-to-MRI Transfer

To validate cross-modal robustness, we transfer CT-pre-trained features to MRI brain tumor segmentation on BraTS 21. In this experiment, the CT-pre-trained Swin Transformer encoder initializes the downstream Swin-UNETR segmentation model, and all segmentation parameters are then fine-tuned on BraTS 21. No BraTS 21 images are used during CT pre-training. Despite different imaging physics, ARM achieves an average Dice score of 79.62% with 1k pre-training and 80.25% with 10k pre-training, as shown in Table 5. These results indicate that active region masking can learn 3D geometric features that transfer across imaging modalities.

Table 5. Experiment results on the BraTS 21 dataset (Dice score, %). Bold rows denote the proposed ARM variants.

Method	Pre-Training Dataset	TC	WT	ET	Avg.
MoCo v3	1k	82.60	88.89	57.69	76.39
localMIM [†]	1k	82.74	88.58	59.04	76.79
SimMIM [†]	1k	83.66	90.73	58.87	77.75
MCMAE [†]	1k	84.15	90.65	59.18	77.99
MAE3D [†]	1k	82.84	90.05	59.38	77.42
Swin-UNETR	1k	82.51	89.08	58.15	76.58
SwinMM	1k	83.48	90.47	58.72	77.56
ARM (Ours)	1k	85.35	92.15	61.35	79.62
ARM (Ours)	10k	86.25	92.65	61.85	80.25

[†] Experiments re-implemented by us.

4.5. Ablation Studies

4.5.1. Impact of Labeled Data Ratios

Figure 4 compares ARM with MAE3D and Swin-UNETR on BTCV and MM-WHS using labeled-data ratios from 25% to 100%. ARM consistently outperforms the baselines at all label ratios, with the largest gain in the low-label regime. On BTCV with 25% labeled data, ARM improves over the baselines by approximately 4.75 percentage points, demonstrating improved label efficiency.

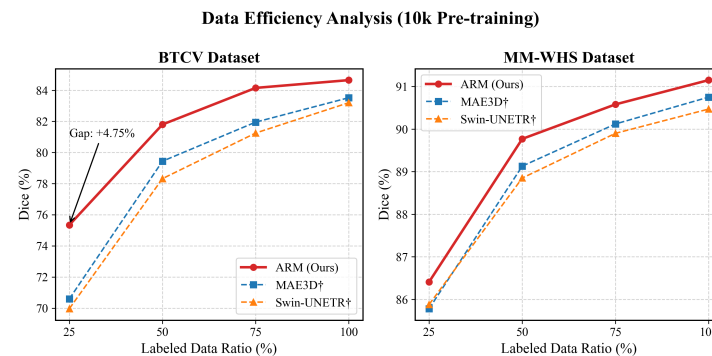


Figure 4. Performance comparison under different labeled-data ratios. ARM shows clear gains, especially in low-data regimes. The dagger symbol denotes experiments re-implemented by us.

4.5.2. Non-RL Top-K Hard Masking

To examine whether the actor–critic component is necessary, we additionally compare ARM with a non-RL hard-masking baseline, denoted as Top-K Error Mask. This baseline first estimates patch-wise reconstruction errors and then masks the top ρN patches with the largest errors, where $\rho = 0.7$ is the same target masking ratio used by ARM. Unlike ARM, Top-K Error Mask does not use actor–critic policy learning, value estimation, entropy regularization, or alternating policy optimization. Therefore, it serves as a direct control for separating the effect of hard-region mining from the effect of the learned masking policy.

Table 6 shows that Top-K Error Mask improves over Random MIM, suggesting that hard-to-reconstruct regions provide more informative pre-training signals than uniformly sampled patches. However, ARM further improves over Top-K Error Mask on both BTCV and MM-WHS. This indicates that the actor–critic policy does not merely select high-error patches, but learns a more stable and adaptive masking distribution that balances hard-region mining with contextual preservation.

Table 6. Ablation study of non-RL Top-K hard masking. All methods use the same backbone, masking ratio, pre-training data, fine-tuning protocol, and training budget. The bold row denotes the proposed ARM method.

Method	BTCV Dice (%)	MM-WHS Dice (%)	2-Task Avg. (%)	Mask Ratio
Random MIM	81.98	89.33	85.66	0.7
Top-K Error Mask	82.60	89.80	86.20	0.7
ARM (Ours)	83.68	90.68	87.18	0.7

4.5.3. Impact of Masking Strategy and Randomness

Table 7 evaluates the exploration index ϵ . Deterministic masking ($\epsilon = 0$) degrades performance, suggesting that overly hard masking can remove excessive informative context and create an unstable or nearly impossible reconstruction task. Fully random masking ($\epsilon = 1$) also performs poorly because it returns to easy, non-adaptive reconstruction. The setting $\epsilon = 0.3$ provides the best average trade-off across BTCV and MM-WHS, while

$\epsilon = 0.4$ produces the highest MM-WHS score alone. This supports the need to balance hard example mining with contextual preservation.

Table 7. Effect of the exploration index ϵ on downstream segmentation performance. Bold values indicate the best result in each metric column.

Exploration ϵ	Difficulty	BTCV	MM-WHS
0 (Deterministic)	Hard	81.84	88.21
10%		82.71	89.90
20%		83.25	90.41
30%	Balanced	83.68	90.68
40%		83.06	91.01
50%		82.41	90.24
100% (Random)	Easy	81.98	89.33

5. Discussion

The experimental results support the hypothesis that adaptive hard-region masking can be more effective than uniform random masking for 3D medical MIM in the evaluated settings. The strongest improvements are observed in label-scarce settings and in anatomically challenging tasks involving small structures or ambiguous boundaries. This suggests that ARM shifts the reconstruction objective away from background interpolation and toward learning structure-aware contextual representations. The CC-CCII and BraTS 21 experiments further indicate that the learned features are not limited to segmentation, but can also transfer to disease classification and CT-to-MRI segmentation.

The Top-K Error Mask ablation further separates the effect of hard-region mining from the effect of policy learning. Although Top-K Error Mask improves over Random MIM, ARM remains better on both BTCV and MM-WHS, indicating that the learned masking policy provides additional benefit beyond simply selecting patches with the largest instantaneous reconstruction errors. This supports the use of actor–critic policy learning to balance reconstruction difficulty, mask diversity, and contextual preservation.

The pattern of improvements is internally consistent across the 12 task-level evaluations. The largest absolute gains appear precisely where uniform random masking is most under-supplied with informative content: small or boundary-ambiguous structures (e.g., approximately +7.7 DSC points on MSD Colon and +1.3 DSC points on MSD Pancreas over Swin-UNETR), out-of-distribution segmentation benchmarks such as AMOS22 and FLARE22, and the 25% labeled-data regime where ARM exceeds the baselines by approximately 4.75 DSC points on BTCV. Conversely, on benchmarks where informative content is comparatively dense—such as MSD Liver, where Dice scores already exceed 95% across all baselines—the margin shrinks to less than one percentage point. This is the expected behavior of an adaptive masking method: ARM cannot create supervision that is not latent in the data; rather, it routes the existing pre-training signal toward patches where it is most useful.

ARM also has limitations. The alternating minimax optimization adds overhead relative to vanilla MIM at pre-training time, although this cost is paid only once and downstream fine-tuning loads only the Swin Transformer encoder, incurring no additional inference cost. In addition, reconstruction error is an imperfect proxy for anatomical informativeness: high-error patches may correspond not only to meaningful anatomical boundaries or small structures but also to image noise, artifacts, or sharp intensity transitions. The current policy operates at the patch level and does not incorporate explicit anatomical priors such as organ-level connectivity or symmetry, which may limit gains for tasks where global anatomical structure is itself diagnostic. Finally, the ablation on ϵ

shows that the masking policy must preserve sufficient contextual information; purely deterministic hard masking is counterproductive because it can make the reconstruction task unstable or overly difficult. Future work will focus on amortizing the policy update into a single-pass masking head, integrating anatomical priors into the state representation, adding systematic anatomical-overlap analysis of selected masks, and scaling the framework to large 3D medical foundation models.

6. Conclusions

We presented ARM, a self-supervised pre-training method for 3D medical image analysis that formulates mask generation as a patch-wise actor–critic masking problem. By actively masking hard-to-reconstruct regions that may contain informative anatomical structures, ARM encourages the reconstruction network to learn global context and structure-aware representations instead of relying on local interpolation. Evaluations across 12 task-level benchmarks show consistent improvements over the evaluated SSL baselines, strong label efficiency in data-scarce regimes, and robust CT-to-MRI transfer. These findings demonstrate the potential of active masking as a general pre-training strategy for volumetric medical image analysis. Future work will investigate lower-cost policy optimization, explicit anatomical priors, and broader validation on large-scale multimodal medical foundation models.

Author Contributions: Conceptualization, C.W. and Z.C.; methodology, C.W.; software, C.W.; validation, C.W.; formal analysis, C.W.; investigation, C.W.; data curation, C.W.; writing—original draft preparation, C.W.; writing—review and editing, C.W. and Z.C.; visualization, C.W.; supervision, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study because only publicly available, anonymized datasets were used.

Informed Consent Statement: Patient consent was waived because this study used only publicly available, anonymized datasets, and no identifiable private information was accessed.

Data Availability Statement: The datasets analyzed in this study are publicly available from their respective repositories, including BTCV, TCIA COVID-19, MM-WHS, AMOS22, FLARE22, MSD, CC-CCII, and BraTS 21. Pre-trained model weights and the training code will be released at the corresponding GitHub repository upon acceptance.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ARM	Active Region Masking
SSL	Self-Supervised Learning
MIM	Masked Image Modeling
A3C	Asynchronous Advantage Actor–Critic
AHM	Adaptive Hard Masking
DSC	Dice Similarity Coefficient
NSD	Normalized Surface Dice
ACC	Accuracy
AUC	Area Under the Receiver Operating Characteristic Curve

References

- Shurrab, S.; Duwairi, R. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Comput. Sci.* **2022**, *8*, e1045. [[CrossRef](#)] [[PubMed](#)]
- Bao, H.; Dong, L.; Piao, S.; Wei, F. BEiT: BERT pre-training of image transformers. *arXiv* **2021**, arXiv:2106.08254.
- 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, New Orleans, LA, USA, 18–24 June 2022; pp. 9653–9663.
- Chaitanya, K.; Erdil, E.; Karani, N.; Konukoglu, E. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12546–12558.
- 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, Montreal, QC, Canada, 10–17 October 2021; pp. 3499–3509.
- 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, New Orleans, LA, USA, 18–24 June 2022; pp. 20730–20740.
- Wang, H.; Tang, Y.; Wang, Y.; Guo, J.; Deng, Z.-H.; Han, K. Masked image modeling with local multi-scale reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2122–2131.
- Wang, Y.; Li, Z.; Mei, J.; Wei, Z.; Liu, L.; Wang, C.; Sang, S.; Yuille, A.L.; Xie, C.; Zhou, Y. SwinMM: Masked multi-view with Swin Transformers for 3D medical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2023; pp. 486–496.
- Zhuang, J.; Luo, L.; Wang, Q.; Wu, M.; Luo, L.; Chen, H. Advancing volumetric medical image segmentation via global-local masked autoencoders. *IEEE Trans. Med. Imaging* **2025**, *44*, 4226–4238. [[CrossRef](#)] [[PubMed](#)]
- Alansary, A.; Oktay, O.; Li, Y.; Le Folgoc, L.; Hou, B.; Vaillant, G.; Kamnitsas, K.; Vlontzos, A.; Glocker, B.; Kainz, B.; et al. Evaluating reinforcement learning agents for anatomical landmark detection. *Med. Image Anal.* **2019**, *53*, 156–164. [[CrossRef](#)] [[PubMed](#)]
- Ghesu, F.C.; Georgescu, B.; Zheng, Y.; Grbic, S.; Maier, A.; Hornegger, J.; Comaniciu, D. Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *41*, 176–189. [[PubMed](#)]
- Zhuang, J.; Wu, L.; Wang, Q.; Fei, P.; Vardhanabhuti, V.; Luo, L.; Chen, H. MIM: Mask in mask self-supervised pre-training for 3D medical image analysis. *IEEE Trans. Med. Imaging* **2025**, *44*, 3727–3740. [[PubMed](#)]
- Wu, L.; Zhuang, J.; Chen, H. VoCo: A simple-yet-effective volume contrastive learning framework for 3D medical image analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 22873–22882.
- Mnih, V.; Badia, A.P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1928–1937.
- Wang, Z.; Bovik, A.C. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag.* **2009**, *26*, 98–117. [[CrossRef](#)]
- Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Online, 13–18 July 2020; pp. 1597–1607.
- Chen, X.; Xie, S.; He, K. An empirical study of training self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 9640–9649.
- Gao, P.; Ma, T.; Li, H.; Lin, Z.; Dai, J.; Qiao, Y. MCMAE: Masked convolution meets masked autoencoders. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 35632–35644. [[CrossRef](#)]
- Jiang, J.; Lu, X.; Zhao, L.; Dazeley, R.; Wang, M. Masked autoencoders in 3D point cloud representation learning. *IEEE Trans. Multimed.* **2023**, *27*, 820–831. [[CrossRef](#)]
- He, Y.; Yang, G.; Ge, R.; Chen, Y.; Coatrieux, J.L.; Wang, B.; Li, S. Geometric visual similarity learning in 3D medical image self-supervised pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2023; pp. 9538–9547.
- Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: UNet-like pure transformer for medical image segmentation. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 205–218.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.