Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Health

Two-Fold Patch Perturbation for Efficient Self-Supervised Learning in 3D Medical Imaging

Tirthajit Baruah[1] , **Kabir Jamadar**[2] , **Punit Rathore**[1,3]
[1]RBCCPS, Indian Institute of Science, Bengaluru, India
[2]Indian Institute of Technology Madras, India
[3]CiSTUP, Indian Institute of Science, Bengaluru, India
tirthajitb@iisc.ac.in, kabir.swish@gmail.com

Abstract

Self-supervised pre-training has become a key paradigm for reducing annotation costs in 3D medical imaging, yet many recent approaches rely on complex objectives or incur substantial computational overhead. We propose a simple and efficient self-supervised pre-training framework for 3D medical images based on a two-fold patch-wise perturbation strategy. The method applies Bernoulli patch masking and discrete rotations, and trains a shared encoder with a three-head objective for reconstruction, perturbation localization, and rotation prediction. This design encourages spatially aware and transferable representations while remaining computationally lightweight. Experiments across diverse segmentation and classification benchmarks, including modality-shift scenarios, demonstrate consistent improvements over general self-supervised baselines and competitive or superior performance compared to recent medical self-supervised methods, while requiring substantially less memory, computation, and training time than the state-of-the-art pre-training pipelines.

1 Introduction

Self-supervised learning (SSL) has become a key strategy for reducing annotation costs in 3D medical imaging, where dense voxel-level labels are expensive and scarce. This is relevant especially for volumetric segmentation and classification, where model performance typically scales with both data size and label quality, yet curated annotations remain a major bottleneck.

Prior works explore proxy objectives that encourage the development of transferable volumetric representations without labels. Broadly, these objectives emphasize either appearance-based reconstruction, such as masked modeling [Xie *et al.*, 2022b] and reconstruction, or geometric pretexts, such as rotation [Chen *et al.*, 2023], jigsaw [Chen *et al.*, 2021a], and Rubik-style transformations [Tao *et al.*, 2020], that induce spatial reasoning and orientation sensitivity. In parallel, recent methods also exploit volumetric priors such as anatomical layout and contextual position to inject higher-level structure into SSL representations [Wu *et al.*, 2024]. Despite strong progress, these approaches are often developed in isolation, and their combination is typically ad hoc or entails substantial computational cost.

Designing SSL for 3D volumes also introduces practical constraints. Compared to 2D images, volumetric inputs are high-dimensional and memory-intensive. This makes it difficult to scale sophisticated pretext pipelines. At the same time, reconstruction-only objectives can under-emphasize geometric consistency. Also, transformation-prediction objectives may become unreliable when local appearance is ambiguous, such as near-isotropic texture, and the applied transformation cannot be inferred without a broader spatial context. These observations motivate a simple question: can we unify appearance and geometry supervision in a way that remains explicit, tunable, and computationally lightweight?

We instead view augmentation as an explicit perturbation process, where each component can be controlled and ablated. For volumetric data, patch-wise perturbation is particularly appealing. As it is natural for 3D inputs, it aligns with transformer-style tokenization and provides a direct handle on how much local information is removed or perturbed. If perturbation is explicitly defined, the resulting SSL task can be systematically harder or easier, and the model can be supervised not only to reconstruct, but also to reason about *where* perturbation occurred and *what* was applied.

In this work, we propose a simple and principled SSL framework that unifies appearance and geometric supervision through a two-fold perturbation process. Specifically, we model perturbation as the composition of two distributions: (i) a Bernoulli selection over 3D patches and (ii) a discrete rotation applied only to the selected patches. This formulation is controlled by the mask (selection) rate p, the rotation set size K, and the patch grid resolution n, allowing the perturbation strength to be systematically varied and ablated from weak to strong.

To exploit this structured perturbation, we introduce a three-head pre-training objective that jointly supervises reconstruction, patch selection, and rotation prediction. Beyond reconstruction alone, predicting where perturbations occur and how they are applied encourages the encoder to integrate local and contextual information, yielding spatially aware representations. This is particularly beneficial for medical volumetric data, with its rich anatomical semantics and localized information. Experiments across diverse segmenta-

tion and classification benchmarks, including modality-shift scenarios, show consistent gains over general SSL baselines and competitive or superior performance to recent medical SSL methods, while substantially reducing memory, computation, and training time compared to heavier state-of-the-art pre-training pipelines.

Contributions. (1) We propose a simple and tunable two-fold patch perturbation algorithm for volumetric SSL, parameterized by (n, p, K) to control pre-training difficulty systematically; (2) we introduce a lightweight three-head objective (reconstruction, mask prediction, rotation prediction) that promotes spatially aware and transferable representations with minimal additional decoder capacity; and (3) we provide extensive evaluation on diverse medical imaging benchmarks (segmentation and classification), including cross-modality transfer, demonstrating strong downstream performance while substantially reducing pre-training memory, FLOPs, and wall-clock time compared to recent state-of-the-art SSL methods.

2 Related Work

Self-supervised learning for 3D medical imaging. Most SSL methods for volumetric medical images can be broadly grouped into two families relevant to this work: (i) masked reconstruction (masked modeling) and (ii) geometric/spatial pretext tasks. Our method draws on both directions while emphasizing an explicit, tunable perturbation process.

(i) Masked reconstruction and modeling. Masked reconstruction is a dominant paradigm for learning 3D representations without labels, where models reconstruct missing or perturbed voxels from partial observations, as in MAE-style pre-training and its medical variants [He *et al.*, 2022; Chen *et al.*, 2023; Zhuang *et al.*, 2025]. GL-MAE, in particular, tailors masked autoencoding to volumetric data and highlights the role of global context under masking. While effective, reconstruction-centric objectives primarily target appearance recovery and may provide weaker supervision for geometric consistency.

(ii) Geometric and spatial pretext tasks. A complementary line of work [Taleb *et al.*, 2020; Chen *et al.*, 2021a] uses transformation-prediction objectives, such as rotation classification and patch arrangement, to encourage spatial reasoning and geometric awareness. Rubik-style objectives [Tao *et al.*, 2020] extend this idea by applying discrete permutations and training the network to predict the applied discrete transformation. Although conceptually related, these pretexts typically apply constrained transformations and do not explicitly couple *where* perturbation occurs with *what* transformation is applied at the patch level, thus making the proxy task less informative.

Position-aware 3D SSL and relation to VoCo. Recent work exploits anatomical regularities in 3D medical volumes by learning representations sensitive to spatial context. VoCo [Wu *et al.*, 2024] introduces a volume-contrast formulation that leverages contextual position prediction using global-local views and prototypes. In contrast, our approach does not rely on contrastive sampling or prototype construction; instead, it injects spatial and geometric supervision directly through structured patch-wise perturbation and corresponding prediction heads.

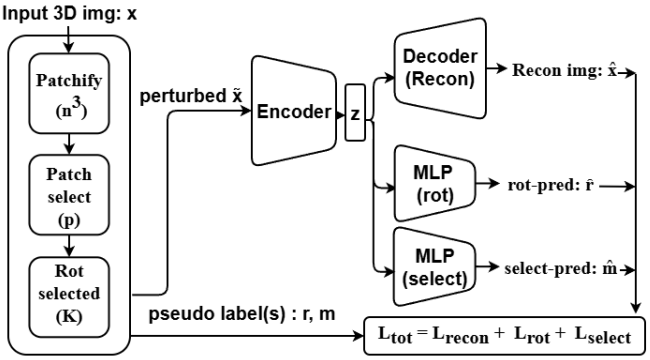

Figure 1: The overall pre-training framework.

Relation to our work. Our approach differs from prior work in how reconstruction and transformation supervision are integrated. Rather than combining objectives in a loosely coupled manner, we define a structured two-fold patch-wise perturbation process that explicitly models both *where* perturbation occurs and *how* it is applied. This yields a unified, tunable pre-training objective that encourages spatial awareness and orientation sensitivity, while remaining computationally lightweight compared to contrastive or similar SSL methods.

3 Methodology

3.1 Overview

We aim to learn transferable representations from unlabeled 3D medical volumes using complementary self-supervised pretext tasks. Given a fixed-size ROI volume, we generate a *two-fold* perturbed view via patch-wise masking and rotation, encode it with a Swin-UNETR-style backbone, and jointly optimize reconstruction, mask localization, and rotation prediction objectives (Figure 1) . The resulting encoder is transferred and finetuned for downstream supervised tasks.

3.2 Input Preprocessing and Notation

After standard preprocessing, each input volume is cropped or padded to a fixed ROI of size $64 \times 64 \times 64$. We denote the resulting tensor as

$$x \in R^{B \times C \times 64 \times 64 \times 64}, \tag{1}$$

where B is the batch size and C the number of input channels ($C = 1$ in our experiments).

3.3 Two-Fold Patch-wise Perturbation

We partition the ROI volume into a regular grid of $n \times n \times n$ non-overlapping cubic patches ("patchification"), yielding $P = n^3$ patches, each of spatial size $s \times s \times s$ with $s = 64/n$. Let $\mathcal{S}_n(\cdot)$ denote the patchification operator:

$$X = \mathcal{S}_n(x) \in R^{B \times P \times C \times s \times s \times s}, \tag{2}$$

and let $\mathcal{C}_n(\cdot)$ denote the inverse operation that reassembles patches into a full volume.

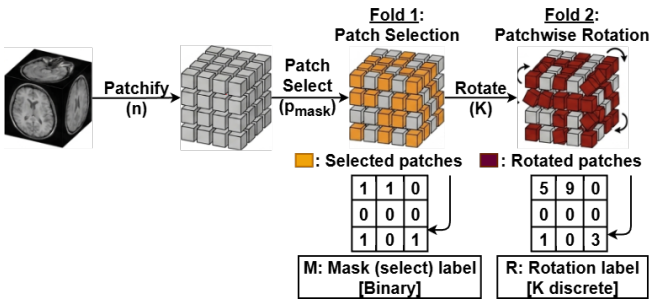

Figure 2: Two-fold patch-wise perturbation strategy.

Fold 1: Patch selection (masking). We sample a Bernoulli mask over patches:

$$m_{b,i} \sim \text{Bernoulli}(p_{\text{mask}}), \;\; b = 1, .., B, \;\; i = 1, .., P \quad (3)$$

where p_{mask} is the masking rate. Intuitively, $m_{b,i} = 1$ indicates that patch i in sample b will be perturbed.

Fold 2: Patch-wise rotation. For each selected patch, we sample a discrete rotation label from a predefined set of K 3D rotation transformations:

$$r_{b,i} \sim \begin{cases} \text{Unif}\{1, \ldots, K-1\}, & \text{if } m_{b,i} = 1, \\ 0, & \text{if } m_{b,i} = 0, \end{cases} \quad (4)$$

where $r = 0$ denotes the identity (no rotation). Let $\{\mathcal{R}_k\}_{k=0}^{K-1}$ be the set of 3D rotation operators implemented by axis permutations and flips (90°-multiple rotations without interpolation). We perturb patches as:

$$\tilde{X}_{b,i} = \begin{cases} \mathcal{R}_{r_{b,i}}(X_{b,i}), & \text{if } m_{b,i} = 1, \\ X_{b,i}, & \text{if } m_{b,i} = 0, \end{cases} \quad (5)$$

and reassemble the perturbed patches into perturbed volume:

$$\tilde{x} = \mathcal{C}_n(\tilde{X}) \in R^{B \times C \times 64 \times 64 \times 64} \quad (6)$$

This two-fold perturbation (Figure 2) introduces both spatial uncertainty and orientation inconsistency, requiring the encoder to localize perturbed regions and leverage surrounding anatomical context to resolve ambiguities.

3.4 Model Architecture

Our network consists of (i) a Swin-UNETR-style encoder, (ii) a UNETR decoder for volumetric reconstruction, and (iii) two lightweight MLP heads for mask and rotation prediction.

Swin-UNETR-style Encoder

We use a 3D Swin Transformer backbone with hierarchical representations. Denote the encoder by $E(\cdot)$; given the perturbed volume $\tilde{x}$ it produces multi-scale feature maps:

$$\{h^{(\ell)}\}_{\ell=1}^{L} = E(\tilde{x}), \quad (7)$$

where $h^{(\ell)}$ corresponds to a spatial scale in the Swin hierarchy. Following the Swin-UNETR wiring, selected transformer features are refined using convolutional blocks to produce multi-resolution encoder feature maps

$$\mathcal{E}(\tilde{x}) = \{e^{(0)}, e^{(1)}, e^{(2)}, e^{(3)}, e^{(4)}\}, \quad (8)$$

where $e^{(0)}$ captures low-level appearance and $e^{(4)}$ provides the highest-level spatial representation.

Multi-scale pooled representation. We construct a global representation by applying global average pooling to each encoder scale and concatenating the results:

$$z_{\text{global}} = \text{concat}\Big(\text{GAP}(e^{(0)}), .., \text{GAP}(e^{(4)})\Big) \in R^{d_g}, \quad (9)$$

where d_g denotes the concatenated feature dimension. This compact representation is used to drive patch-level prediction heads.

Reconstruction Decoder

We attach a UNETR-style decoder $D_{\text{rec}}(\cdot)$, reusing the Swin-UNETR decoder pathway that combines the highest-level feature map with skip connections from intermediate encoder resolutions:

$$\hat{x} = D_{\text{rec}}\Big(\{h^{(\ell)}\}_{\ell=1}^{L}, \mathcal{E}(\tilde{x})\Big) \in R^{B \times C \times 64 \times 64 \times 64}. \quad (10)$$

Mask and Rotation Heads

We predict (i) which patches were perturbed and (ii) the rotation applied to each perturbed patch, using lightweight MLP heads on the pooled multi-scale representation z_{global}:

$$\hat{m} = D_{\text{mask}}(z_{\text{global}}) \in R^{B \times P}, \quad (11)$$

$$\hat{r} = D_{\text{rot}}(z_{\text{global}}) \in R^{B \times P \times K}, \quad (12)$$

where $\hat{m}$ are patch-wise logits for perturbation localization and $\hat{r}$ are patch-wise logits over K discrete rotation classes. Using lightweight heads biases representational capacity toward the encoder rather than task-specific decoders.

3.5 Pre-training Objective

We jointly optimize reconstruction, mask prediction, and rotation prediction losses:

$$\mathcal{L} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}} + \lambda_{\text{rot}} \mathcal{L}_{\text{rot}}, \quad (13)$$

with scalar weights $\lambda_{\text{rec}}, \lambda_{\text{mask}}, \lambda_{\text{rot}}$.

Masked-only Charbonnier Reconstruction Loss

We use the Charbonnier penalty (a smooth, robust variant of ℓ_1 [Barron, 2019]):

$$\rho_\epsilon(u) = \sqrt{u^2 + \epsilon^2}, \qquad \epsilon = 10^{-3}. \quad (14)$$

To align reconstruction learning with our perturbation process, we compute reconstruction error only over voxels in perturbed patches. To do so, we convert the patch mask $m \in \{0,1\}^{B \times P}$ into a voxel-level mask $w \in \{0,1\}^{B \times 1 \times 64 \times 64 \times 64}$ by reshaping m into a $n \times n \times n$ grid and upsampling with nearest-neighbor interpolation (preserving exact patch boundaries). The masked-only reconstruction loss is:

$$\mathcal{L}_{\text{rec}} = \frac{\sum_v w_v \, \rho_\epsilon(\hat{x}_v - x_v)}{\sum_v w_v + \delta}, \qquad \delta = 10^{-6}, \quad (15)$$

where v indexes voxels (and channels). This term encourages the model to "undo" local patch perturbation by leveraging the surrounding anatomical context.

Mask Prediction Loss

The mask head predicts which patches were perturbed. We use a class-balanced binary cross-entropy on patch logits:

$$\mathcal{L}_{\text{mask}} = \text{BCEWithLogits}(\hat{m}, m; \alpha), \;\; \alpha \approx \frac{\#\text{neg}}{\#\text{pos}}, \quad (16)$$

where α is a positive-class reweighting factor to counter class imbalance induced by sparse masking. In practice, we estimate α per mini-batch, with a safe fallback to unweighted BCE when the number of $\text{positives} = 0$.

Rotation Prediction Loss (Selected-only)

For masked patches, the rotation head predicts the applied rotation label. We compute cross-entropy only over the masked subset $\Omega = \{(b, i) \mid m_{b,i} = 1\}$:

$$\mathcal{L}_{\text{rot}} = \frac{1}{|\Omega|} \sum_{(b,i)\in\Omega} \text{CE}(\hat{r}_{b,i}, r_{b,i}) \,, \quad (17)$$

where $\hat{r}_{b,i} \in R^K$ are logits over K rotation classes. We employ label smoothing to reduce overconfident predictions and improve the generalization of the learned representation. If $|\Omega| = 0$ for a mini-batch, we set $\mathcal{L}_{\text{rot}} = 0$ for that update.

3.6 Training Procedure

During pre-training, a perturbed view $(\tilde{x}, m, r)$ is generated for each unlabeled input x using the two-fold perturbation in Sec. 3.3. The network predicts $(\hat{x}, \hat{m}, \hat{r})$, and all parameters are jointly optimized using the combined objective in Eq. (13). The encoder is subsequently finetuned on downstream tasks.

4 Experiments

4.1 Experimental Protocol

We adopt the experimental protocol of recent state-of-the-art methods to ensure a fair and directly comparable evaluation against a broad range of prior self-supervised learning methods. Unless otherwise stated, we follow VoCo for preprocessing, training budget, backbone family, and finetuning recipe, and we compare against the results reported in [Wu *et al.*, 2024; Zhuang *et al.*, 2025], which re-implements most baselines under a unified setup.

Implementation details We use a Swin-UNETR [Hatamizadeh *et al.*, 2021] architecture as the model backbone for both self-supervised pre-training and downstream finetuning, in accordance with the previous baselines. For training, we use AdamW [Loshchilov and Hutter, 2017] optimization and a warmup-cosine learning rate schedule over 500 epochs to maintain consistency with previous methods. The finetuning setup is mostly consistent with the pre-training, differing mainly in the number of training epochs across datasets. We apply slicing-window inference to compare with SOTA methods. We do not use foundation models or post-processing, aiming to evaluate our method's pure effectiveness. Details of the training settings and preprocessing are provided in our GitHub repository.

4.2 Pre-training Data

We pre-train on three publicly available CT datasets: BTCV [Landman *et al.*, 2015], TCIA COVID-19 [Clark *et al.*, 2013], and LUNA [Setio *et al.*, 2017], which totals approximately 1.6k CT scans. For downstream experiments on BTCV [Landman *et al.*, 2015], WORD [Luo *et al.*, 2022], and AMOS22 [Ji *et al.*, 2022], we pre-train only on BTCV and TCIA COVID-19, consistent with the protocol of [Wu *et al.*, 2024; Zhuang *et al.*, 2025; Chen *et al.*, 2023]. For the other 4 downstream tasks on MM-WHS [Zhuang, 2018], MSD-spleen [Antonelli *et al.*, 2022], CC-CCII [Zhang *et al.*, 2020], and BraTS [Baid *et al.*, 2021], we pre-train on the union of all three datasets.

The two-fold perturbation parameters. Unless otherwise specified, we use a patch grid of $n \times n \times n$ with $n = 4$ (thus $P = n^3 = 64$ patches), masking rate $p_{\text{mask}} = 0.3$, and a rotation set of size $K = 9$ for the pre-training task.

4.3 Downstream Tasks and Datasets

We evaluate transfer performance on seven downstream benchmarks spanning segmentation and classification across CT and MRI data. Details on the source, modality, size, and split of the dataset are provided in the supplementary materials in the GitHub repository and align with the standard SOTA methods reported in previous works. We will also provide the JSON files containing the sample lists for each dataset to ensure reproducibility and fair comparison.

Segmentation on CT. We finetune on BTCV, AMOS, WORD, MSD-Spleen, and MM-WHS to benchmark against all medical segmentation baselines, and report the Dice score (%) on the test set.

Segmentation on MRI. To evaluate robustness under modality shift, we finetune on BraTS (MRI data). Since pretraining is conducted exclusively on CT, BraTS constitutes an out-of-distribution (OOD) modality transfer setting.

Classification on CT. We further evaluate on CC-CCII, a CT-based classification benchmark. We report classification accuracy (%) following the evaluation protocol of previous methods.

4.4 Baselines

We compare against three baseline categories, using results reported in [Wu *et al.*, 2024; Zhuang *et al.*, 2025] under a matched setup. Not all methods are reported for every dataset; we follow the available comparisons per benchmark.

Training from scratch. We include randomly initialized baselines using standard 3D architectures, including 3D U-Net [Ronneberger *et al.*, 2015], UNETR [Hatamizadeh *et al.*, 2022], and Swin-UNETR [Hatamizadeh *et al.*, 2021].

General-purpose SSL. We compare with representative general SSL methods adapted to 3D data, including MAE3D [He *et al.*, 2022], SimCLR [Chen *et al.*, 2020], SimMIM [Xie *et al.*, 2022b], MoCo v3 [Chen *et al.*, 2021b], Jigsaw [Chen *et al.*, 2021a], Position-Label [Zhang and Gong, 2023], and DINO [Caron *et al.*, 2021].

Method	Network	Dice Score(%)	
		AMOS22	WORD
From Scratch[1]			
UNETR	–	83.94	77.05
General SSL			
MoCo v3	UNETR	84.25	77.05
Jigsaw	UNETR	84.60	77.57
DINO	UNETR	85.11	78.12
3D Medical SSL			
Swin-UNETR	UNETR	85.64	78.26
Medical MAE	UNETR	86.13	78.43
GL-MAE	UNETR	87.14	79.44
Our method	Sw-UNETR	**89.78**	**84.20**

Table 1: Experimental results on AMOS22 and WORD.

Medical SSL. We compare with medical SSL methods, including Model Genesis [Zhou *et al.*, 2019], Rotation [Taleb *et al.*, 2020], VICRegK [Bardes *et al.*, 2022], UniMiSS [Xie *et al.*, 2022a], Rubik++ [Tao *et al.*, 2020], PCRLv2 [Zhou *et al.*, 2023], Med-MAE [Chen *et al.*, 2023], SwinMM [Wang *et al.*, 2023], JointSSL [Nguyen *et al.*, 2023], GVSL [He *et al.*, 2023], GLMAE [Zhuang *et al.*, 2025], and VoCo [Wu *et al.*, 2024].

4.5 Downstream Results

We report downstream transfer performance in Tables 1–5, covering challenging multi-organ CT segmentation (BTCV, AMOS22, WORD), cardiac whole-body CT segmentation (MM-WHS), single-organ CT segmentation (MSD-Spleen), MRI brain tumor segmentation under modality shift (BraTS21), and CT-based classification (CC-CCII).

Large gains on AMOS22 and WORD. As shown in Table 1, our method achieves the best results on both AMOS22 and WORD, reaching 89.78 and 84.20 Dice, respectively. These scores substantially exceed the strongest prior baseline in the table (GL-MAE) by $+2.64$ Dice on AMOS22 (89.78 vs. 87.14) and $+4.76$ Dice on WORD (84.20 vs. 79.44), indicating that the proposed method's objective transfers effectively to complex multi-organ segmentation.

Robust transfer under CT→MRI modality shift. Table 2 reports BraTS21 results, where pre-training is performed on CT while finetuning is conducted on MRI, making this an out-of-distribution modality transfer setting. Our method achieves the best overall performance with an average Dice of 88.50, improving upon VoCo by $+9.97$ Dice (88.50 vs. 78.53). Notably, the gain is driven by a large improvement on the enhancing tumor (ET) region (83.56 vs. 59.87), suggesting that the learned representation captures transferable cues that remain useful under substantial appearance shifts.

Improved CT classification on CC-CCII. On CC-CCII classification (Table 3), our method achieves 94.25% accuracy, outperforming VoCo (90.83%) and all other reported SSL baselines. This result indicates that the learned encoder features generalize beyond segmentation and remain effective for image-level decision making.

[1] 'From Scratch' denotes the supervised baseline

Method	Net.	Dice Score(%)			Avg.
		TC	WT	ET	
From Scratch					
UNETR	–	81.62	87.81	57.34	75.58
Swin-UNETR	–	81.28	88.67	57.73	75.89
With General SSL					
MAE3D	UNETR	82.34	90.35	59.18	77.29
SimMIM	UNETR	84.06	90.43	59.07	77.85
SimCLR	UNETR	83.13	89.44	58.42	76.99
MoCo v3	UNETR	82.60	88.89	57.69	76.39
Jigsaw	Sw-UNE.	81.62	89.45	59.10	76.72
PosLabel	Sw-UNE.	81.35	89.62	58.73	76.64
With Medical SSL					
PCRLv1	Sw-UNE.	81.96	88.83	57.58	76.12
PCRLv2	Sw-UNE.	82.13	90.06	57.70	76.63
Rubik++	Sw-UNE.	84.32	90.23	58.01	77.51
Swin-UNETR	Sw-UNE.	82.51	89.08	58.15	76.58
SwinMM	Sw-UNE.	83.48	90.47	58.72	77.56
VoCo	Sw-UNE.	85.27	90.45	59.87	78.53
Our method	Sw-UNE.	**89.24**	**92.71**	**83.56**	**88.50**

Table 2: Experimental results on BraTS 21. WT, TC, and ET denote the whole tumor, tumor core, and enhancing tumor, respectively.

Method	Network	Accuracy(%)
From Scratch		
UNETR	–	88.92
Swin-UNETR	–	88.04
With General SSL		
MAE3D	UNETR	89.47
MoCo v3	UNETR	84.95
Jiasaw	Swin-UNETR	86.88
PositionLabel	Swin-UNETR	87.54
With Medical SSL		
PCRLv1	Swin-UNETR	88.72
PCRLv2	Swin-UNETR	89.15
Rubik++	Swin-UNETR	89.23
Swin-UNETR	Swin-UNETR	89.45
SwinMM	Swin-UNETR	89.61
VoCo	Swin-UNETR	90.83
Our method	Swin-UNETR	**94.25**

Table 3: Experimental results of CC-CCII classification.

Competitiveness on BTCV organ-wise segmentation. Table 4 shows organ-wise results on BTCV. Our method achieves an average Dice of 83.57, which is close to the best-performing method (VoCo at 83.85) while surpassing other strong baselines such as GL-MAE (82.01) and SwinMM (81.81). Qualitatively, our approach yields particularly significant improvements on smaller or shape-variable structures

such as the gallbladder and the esophagus, while remaining comparable on large organs.

Strong performance on MM-WHS and MSD-Spleen. Table 5 summarizes results on MM-WHS and MSD-Spleen. Our method achieves the best performance on MM-WHS with a Dice score of 90.67, slightly improving over VoCo (90.54). On MSD-Spleen, our method achieves 96.11 Dice, remaining close to VoCo (96.34) and outperforming other SSL baselines. These results suggest that the proposed pre-training is effective both for challenging multi-structure settings and for high-accuracy single-organ segmentation.

Across datasets. The proposed two-fold SSL objective delivers consistent gains over training from scratch and general SSL methods, and it is competitive with, and often superior to, specialized medical SSL pre-training. The greatest improvements are observed on AMOS22, WORD, BraTS21, and CC-CCII, while performance on BTCV and MSD-Spleen remains close to the best reported baseline.

4.6 Computational Efficiency

We evaluate the computational efficiency of the proposed pre-training strategy against the current state-of-the-art baseline, VoCo. To ensure a controlled comparison, we pre-train both methods on a single GPU (AMD Instinct MI300X) each, under identical experimental protocol. We report three practical efficiency indicators: (i) average wall-clock time per epoch, (ii) peak GPU memory allocation, and (iii) forward-pass FLOPs (reported per sample).

As summarized in Table 6, our method is consistently more efficient across all metrics while maintaining strong downstream transfer performance (Sec. 4.5). Specifically, our approach reduces peak GPU memory usage from 53.11 GiB to 10.41 GiB (an $\sim$80% reduction), decreases forward FLOPs from 1499.03 to 104.59 GFLOPs per sample (over $14\times$ fewer FLOPs), and shortens average epoch time from 723.19 s to 489.85 s (approximately $1.5\times$ faster). These gains make the proposed two-fold SSL method substantially more practical for large-scale 3D pre-training, enabling faster iteration and deployment under constrained compute budgets.

4.7 Ablation Studies

We conduct ablation studies to analyze the contribution of individual objective components and to assess the effect of the perturbation strength of our proposed method. All ablations are performed using the same pre-training budget and finetuning protocol as the main experiments, ensuring that performance differences are attributable solely to the ablated factors.

Effect of Objective Components

We first examine the role of each loss component by selectively removing the rotation loss $\mathcal{L}_{\text{rot}}$, the mask prediction loss $\mathcal{L}_{\text{mask}}$, and both jointly, while keeping the masked-only reconstruction loss $\mathcal{L}_{\text{rec}}$ unchanged. Results are summarized in Table 7.

Removing either auxiliary objective leads to a consistent drop in performance, indicating that both patch localization and rotation prediction provide complementary supervision signals beyond reconstruction alone. The full objective achieves the best results across all reported datasets, confirming that jointly optimizing reconstruction, mask prediction, and rotation classification is critical to learning transferable representations. When both auxiliary losses are removed, performance degrades further, demonstrating that reconstruction alone is insufficient to capture the full benefits of the proposed two-fold pre-training strategy.

Effect of perturbation Strength

Next, we study the effect of perturbation strength by varying the patch grid size n and the masking rate p_{mask}, while keeping the rotation set size fixed at $K = 9$. Increasing n produces smaller patches and thus higher spatial granularity, whereas increasing p_{mask} increases the fraction of perturbed patches. Both factors control the difficulty of the pretext task.

As shown in Table 8, moderate perturbation strength yields the best downstream performance. Weak perturbations (small n or low p_{mask}) result in suboptimal performance, likely because the pretext task is insufficiently challenging and fails to enforce robust feature learning. Conversely, over-strong perturbations can introduce excessive noise and degrade representation quality. The default setting ($n = 4$, $p_{\text{mask}} = 0.3$) strikes a favorable balance and is used throughout the main experiments.

Discussion. These ablations confirm that the proposed two-fold SSL framework benefits from both auxiliary objectives and appropriately strong perturbation. The combination of masked-only reconstruction with patch-wise localization and rotation prediction provides richer supervision than any of the individual components. Also, carefully balancing perturbation strength is essential for maximizing downstream transfer performance.

5 Conclusion

We proposed a simple and efficient self-supervised pre-training framework for 3D medical imaging based on a two-fold patch-wise perturbation strategy that combines masking and discrete rotations. By jointly optimizing reconstruction, patch localization, and rotation prediction, the method learns spatially aware and transferable representations without annotations. Experiments across diverse segmentation and classification benchmarks show consistent gains over training from scratch and general SSL baselines, competitive or superior performance compared to recent medical SSL methods, and robust transfer under modality shift, while substantially reducing memory, computation, and training time relative to the SOTA method. Ablation studies further support the contribution of both auxiliary objectives and the role of balancing perturbation strength. Overall, this work offers a practical, scalable approach to learning 3D medical image representations with limited annotation and compute budget.

The explicit parameterization of perturbation by (n, p, K) provides a transparent handle for adjusting pre-training difficulty and systematically studying robustness. Beyond the tasks evaluated here, the same formulation can be extended to other volumetric modalities and objectives, such as richer transformation sets or additional auxiliary signals, without changing the core training methodology.

Method	Dice Score(%)													AVG
	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	RAG	LAG	
From Scratch														
UNETR	93.02	94.13	94.12	66.99	70.87	96.11	77.27	89.22	82.10	70.16	76.65	65.32	59.21	79.82
Swin-UNETR	94.06	93.54	93.80	65.51	74.60	97.09	75.94	91.80	82.36	73.63	75.19	68.00	61.11	80.53
With General SSL														
MAE3D	93.98	94.37	94.18	69.86	74.65	96.66	80.40	90.30	83.13	72.65	77.11	67.34	60.54	81.33
SimCLR	92.79	93.04	91.41	49.65	50.99	98.49	77.92	85.56	80.58	64.37	67.16	59.04	48.99	73.85
SimMIM	95.56	95.82	94.14	52.06	53.52	**98.98**	80.25	88.11	82.98	66.49	69.16	60.88	50.45	76.03
MoCo v3	91.96	92.85	92.42	68.25	72.77	94.91	78.82	88.21	81.59	71.15	75.76	66.48	58.81	79.54
Jigsaw	94.62	93.41	93.55	75.63	73.21	95.71	80.80	89.41	84.78	71.02	79.57	65.68	60.22	81.35
PositionLabel	94.35	93.15	93.21	75.39	73.24	95.76	80.69	88.80	84.04	71.18	79.02	65.11	60.07	81.09
With Medical SSL														
ModelGenesis	91.99	93.52	91.81	65.11	76.14	95.98	**86.88**	89.29	83.59	71.79	81.62	67.97	63.18	81.45
Rotation	91.75	93.13	91.62	65.09	76.55	94.21	86.16	89.74	83.08	71.13	81.55	67.90	63.72	81.20
Vicregl	90.32	94.15	91.30	65.12	75.41	94.76	86.00	89.13	82.54	71.26	81.01	67.66	63.08	80.89
Rubik++	**96.21**	90.41	89.33	75.22	72.64	97.44	79.25	89.65	83.76	74.74	78.35	67.14	61.97	81.38
PCRLv1	95.73	89.66	88.53	75.41	72.33	96.20	78.99	89.11	83.06	74.47	77.88	67.02	61.85	80.78
PCRLv2	95.50	91.43	89.52	76.15	73.54	97.28	79.64	90.16	84.17	75.20	78.71	68.74	62.93	81.74
Swin-UNETR	95.25	93.16	92.97	63.62	73.96	96.21	79.32	89.98	83.19	76.11	**82.25**	68.99	65.11	81.54
SwinMM	94.33	94.18	94.16	72.97	74.75	96.37	83.23	89.56	82.91	70.65	75.52	69.17	62.90	81.81
GL-MAE	94.54	94.39	94.37	73.19	74.93	96.51	83.49	89.74	83.11	70.80	75.71	69.39	63.12	82.01
GVSL	95.27	91.22	92.25	72.69	73.56	96.44	82.40	88.90	84.22	70.84	76.42	67.48	63.25	81.87
VoCo	95.73	**96.53**	**94.48**	76.02	75.60	97.41	78.43	91.21	**86.12**	**78.19**	80.88	**71.47**	67.88	**83.85**
Our method	95.91	94.42	94.29	**78.96**	**76.74**	96.80	81.73	91.15	85.70	74.72	78.92	68.30	**68.76**	83.57

Table 4: Experimental results on BTCV.

Method	MSD Spleen	MM-WHS
From Scratch		
3D Unet	93.71	83.09
UNETR	94.20	85.85
Swin-UNETR	94.63	86.11
With General SSL		
MAE3D	95.20	86.03
MoCo v3	94.32	84.16
Jiasaw	94.29	85.88
PositionLabel	94.16	85.52
With Medical SSL		
ModelGenesis	94.40	86.36
VicRegl	94.12	84.72
UniMiss	95.09	84.68
PCRLv1	94.32	86.58
PCRLv2	94.94	86.82
Rubik++	95.11	87.02
Swin-UNETR	95.02	87.06
SwinMM	95.34	86.98
JointSSL	94.92	84.89
GVSL	95.47	88.27
VoCo	**96.34**	90.54
Our method	96.11	**90.67**

Table 5: Experimental results on MSD Spleen and MM-WHS.

Method	Epoch time	Peak mem	FLOPS
VoCo	723.19	53.11	1499.03
Our method	**489.85**	**10.41**	**104.59**

Table 6: Pre-training efficiency comparison under identical hardware and training settings.

Loss objective	MM-WHS	AMOS22
$\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{rot}} + \mathcal{L}_{\text{mask}}$	**90.67**	**89.78**
$\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{rot}}$	89.82	85.98
$\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{mask}}$	89.37	85.90
$\mathcal{L}_{\text{rec}}$ only	89.65	86.31

Table 7: Ablation on loss function components (Dice %).

Perturbation	BTCV	WORD
[n=2, p=0.3, K=9]	83.19	79.60
[n=2, p=0.7, K=9]	82.79	82.17
[n=4, p=0.3, K=9]	**83.67**	84.20
[n=4, p=0.7, K=9]	83.11	**84.55**
[n=8, p=0.9, K=9]	82.93	83.26

Table 8: Ablation of perturbation strength for the proposed method (Dice %). Here, n denotes the patch grid size, p the masking rate, and K the number of rotations used during pre-training.

Limitations. Results use a fixed training budget. Evaluation with broader hyperparameter sweeps and adaptive loss weighting [Kendall *et al.*, 2018] remains future work.

Acknowledgments

We thank collaborators and domain experts for their helpful feedback. Large language models (LLMs) were used only for minor text refinement and block diagram visualization. All experiments and analyses were conceived and verified by the authors.

Code and Data Availability

The datasets used in this study are publicly available. All preprocessing, training scripts, and JSON files are available in our https://github.com/kabir2505/Two-Fold repository to ensure full reproducibility.

Ethics Statement

This study used only publicly available, de-identified data. No human subjects or patient interaction were involved, and no institutional approval was required.

References

[Antonelli *et al.*, 2022] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[Baid *et al.*, 2021] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[Bardes *et al.*, 2022] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. *Advances in Neural Information Processing Systems*, 35:8799–8810, 2022.

[Barron, 2019] Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4331–4339, 2019.

[Caron *et al.*, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[Chen *et al.*, 2021a] Pengguang Chen, Shu Liu, and Jiaya Jia. Jigsaw clustering for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11526–11535, 2021.

[Chen *et al.*, 2021b] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.

[Chen *et al.*, 2023] Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Balan, and Kevin Brown. Masked image modeling advances 3d medical image analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1970–1980, 2023.

[Clark *et al.*, 2013] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.

[Hatamizadeh *et al.*, 2021] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.

[Hatamizadeh *et al.*, 2022] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[He *et al.*, 2023] Yuting He, Guanyu Yang, Rongjun Ge, Yang Chen, Jean-Louis Coatrieux, Boyu Wang, and Shuo Li. 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*, pages 9538–9547, 2023.

[Ji *et al.*, 2022] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, 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.

[Kendall *et al.*, 2018] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[Landman *et al.*, 2015] Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-*

atlas labeling beyond cranial vault—workshop challenge, volume 5, page 12. Munich, Germany, 2015.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[Luo *et al.*, 2022] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, 2022.

[Nguyen *et al.*, 2023] Duy MH Nguyen, Hoang Nguyen, Truong TN Mai, Tri Cao, Binh T Nguyen, Nhat Ho, Paul Swoboda, Shadi Albarqouni, Pengtao Xie, and Daniel Sonntag. Joint self-supervised image-volume representation learning with intra-inter contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14426–14435, 2023.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[Setio *et al.*, 2017] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

[Taleb *et al.*, 2020] Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. *Advances in neural information processing systems*, 33:18158–18172, 2020.

[Tao *et al.*, 2020] Xing Tao, Yuexiang Li, Wenhui Zhou, Kai Ma, and Yefeng Zheng. Revisiting rubik's cube: Self-supervised learning with volume-wise transformation for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 238–248. Springer, 2020.

[Wang *et al.*, 2023] Yiqing Wang, Zihan Li, Jieru Mei, Zihao Wei, Li Liu, Chen Wang, Shengtian Sang, Alan L Yuille, Cihang Xie, and Yuyin Zhou. Swinmm: masked multi-view with swin transformers for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 486–496. Springer, 2023.

[Wu *et al.*, 2024] Linshan Wu, Jiaxin Zhuang, and Hao Chen. 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*, pages 22873–22882, 2024.

[Xie *et al.*, 2022a] Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *European Conference on Computer Vision*, pages 558–575. Springer, 2022.

[Xie *et al.*, 2022b] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.

[Zhang and Gong, 2023] Zhemin Zhang and Xun Gong. Positional label for self-supervised vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3516–3524, 2023.

[Zhang *et al.*, 2020] Kang Zhang, Xiaohong Liu, Jun Shen, Zhihuan Li, Ye Sang, Xingwang Wu, Yunfei Zha, Wenhua Liang, Chengdi Wang, Ke Wang, et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*, 181(6):1423–1433, 2020.

[Zhou *et al.*, 2019] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International conference on medical image computing and computer-assisted intervention*, pages 384–393. Springer, 2019.

[Zhou *et al.*, 2023] Hong-Yu Zhou, Chixiang Lu, Chaoqi Chen, Sibei Yang, and Yizhou Yu. A unified visual information preservation framework for self-supervised pre-training in medical image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8020–8035, 2023.

[Zhuang *et al.*, 2025] Jiaxin Zhuang, Luyang Luo, Qiong Wang, Mingxiang Wu, Lin Luo, and Hao Chen. Advancing volumetric medical image segmentation via global-local masked autoencoders. *IEEE Transactions on Medical Imaging*, 2025.

[Zhuang, 2018] Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2933–2946, 2018.