MEDGMAE: GAUSSIAN MASKED AUTOENCODERS FOR MEDICAL VOLUMETRIC REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised pre-training has emerged as a critical paradigm for learning transferable representations from unlabeled medical volumetric data. Masked autoencoder based methods have garnered significant attention, yet their application to volumetric medical image faces fundamental limitations from the discrete voxel-level reconstruction objective, which neglects comprehensive anatomical structure continuity. To address this challenge, We propose MedGMAE, a novel framework that replaces traditional voxel reconstruction with 3D Gaussian primitives reconstruction as new perspectives on representation learning. Our approach learns to predict complete sets of 3D Gaussian parameters as semantic abstractions to represent the entire 3D volume, from sparse visible image patches. MedGMAE demonstrates dual utility across medical imaging applications. For representation learning, sparse Gaussian prediction produces superior encoder representations that outperform traditional MAE baselines on downstream segmentation, classification, and registration tasks. For volumetric reconstruction, the Gaussian decoder leverages pretrained anatomical priors to accelerate 3D CT volume reconstruction convergence. Extensive experiments across multiple medical imaging datasets demonstrate that our approach achieves superior performance, establishing a new framework for medical image pre-training. Code will be released soon.

1 INTRODUCTION

Volumetric medical imaging modalities, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), have become indispensable cornerstones of modern clinical practice, providing three-dimensional anatomical information crucial for diagnosis, treatment planning, and prognostic assessment. The advent of deep learning has heralded a new era in the automated analysis of these data, demonstrating unprecedented performance across a spectrum of tasks (Zhou et al., 2023c; Litjens et al., 2017). However, the full potential of these data-hungry models is severely constrained by a fundamental bottleneck: the scarcity of large-scale, expertly annotated datasets (Willemink et al., 2020; Ravì et al., 2016).

This challenge sparks an increasing interest in self-supervised pre-training methods that can harness unlabeled 3D data to improve performance in downstream tasks, such as segmentation, registration, and diagnosis. Due to the high anatomical similarity across different medical volumes, Masked Image Modeling (MIM) has emerged as a powerful 3D pre-training approach for learning local representations by reconstructing masked regions from visible context. Despite its promising results, we identify three fundamental yet underexplored challenges that limit the effectiveness of directly reconstructing masked regions via voxel-level regression: **(i) Discrete reconstruction conflicts with anatomical continuity:** conventional MIM methods typically regress discrete intensity voxels of masked regions (He et al., 2022; Chen et al., 2023). While this teaches the model to "fill in blanks" based on immediate spatial context works well for photorealistic data, it is ill-suited for capturing the underlying semantic continuity and geometric abstraction of anatomical structures in volumetric space. Discrete voxel regression often fails to model shape-consistent features, which are crucial for understanding medical images and transferring knowledge to downstream tasks. **(ii) Non-transferable decoder representations:** A common yet overlooked issue in voxel-based MIM

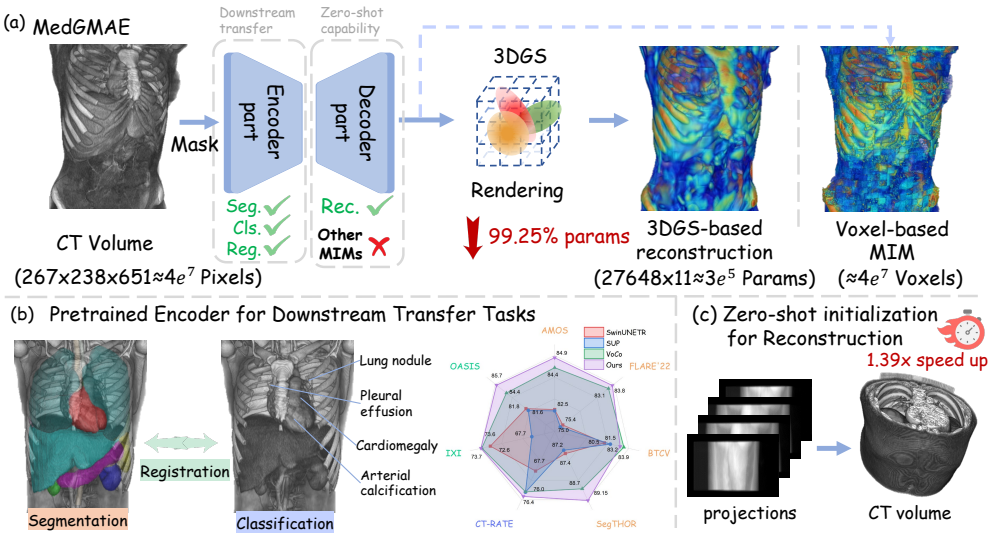

Figure 1: MedGMAE overview. (a) our MedGMAE pre-training with 3D Gaussian Splatting reconstruction leverages CT volume sparsity (anatomical organs occupy only 11.8% of space) to achieve 99.25% parameter reduction and superior coherence compared to voxel-based MIM methods. (b) Pre-trained encoder fine-tuning for downstream tasks: our MedGMAE could learn a strong encoder representation for downstream segmentation, registration, and classification tasks across multiple medical datasets. (c) our MedGMAE could bring a zero-shot capability for 3DGR-based CT reconstruction with 1.39× speed-up.

is that the decoder is designed purely for reconstructing low-level pixel intensities (Xie et al., 2022b; Tang et al., 2024; Tian et al., 2023). The pre-trained decoder is typically discarded, and the features it learns are rarely leveraged for downstream tasks, while its zero-shot capability is inherently constrained by the reliance on pixel-level reconstruction. **(iii) Sparse anatomical distribution leads to parameter inefficiency:** Unlike natural 2D images that contain dense textural information throughout, 3D medical images are inherently sparse in both semantic and intensity distributions. Redundant voxel-based representation fails to achieve optimal reconstruction efficiency.

To address these limitations, we introduce Medical Gaussian Masked Autoencoder (MedGMAE), a novel self-supervised framework tailored for 3D medical image pretraining grounded in a key insight: ***learning sparse 3D Gaussian representations instead of reconstructing dense voxel intensities***. As shown in Fig.1(a), our approach leverages 3D Gaussian primitives as an intermediate representation that naturally addresses the aforementioned challenges through three key advantages: **(i) Continuous geometric modeling for anatomical coherence:** 3D Gaussian primitives provide continuous, differentiable representations that inherently capture geometric abstractions and shape consistency across anatomical structures. Each Gaussian primitive encodes spatial position, orientation, and scale information, enabling the model to learn semantically meaningful geometric features that align with the continuous nature of anatomical boundaries (as shown in Fig.1(b)). **(ii) Transferable decoder:** Our Gaussian-based decoder remains useful after pre-training, directly serving as sophisticated initialization for Gaussian representation 3D medical reconstruction (as shown in Fig.1(c), faster 1.39× for coverage). **(iii) Parameter-efficient representation:**Our Gaussian-based approach naturally aligns with the sparse anatomical distribution in medical volumes, achieving superior parameter efficiency (99% reduction in parameters).

The main contributions of this work can be summarized as follows:

- First, we introduce MedGMAE, the first framework to successfully adapt and extend Gaussian-based masked autoencoding for self-supervised pre-training on 3D volumetric medical data. Our approach learns parameter-efficient representations that better captures continuous anatomical boundaries, enabling models to develop more structured and anatomically-aware representations.

- Second, we demonstrate a novel application for the pre-trained decoder by using it as a zero-shot, geometry-aware initializer for downstream 3D CT reconstruction tasks. The learned anatomical priors from pre-training significantly accelerate 3D Gaussian Representation-based CT reconstruction convergence, thus bridging self-supervised pre-training with practical medical image reconstruction applications.
- Third, extensive experiments across downstream tasks including segmentation, classification, and registration validate the superiority of our proposed approach compared to voxel-based masked representation methods. Additionally, experiments on low-dose CT reconstruction tasks demonstrate the zero-shot initialization capability of our proposed framework, showing significant acceleration in convergence while maintaining reconstruction quality.

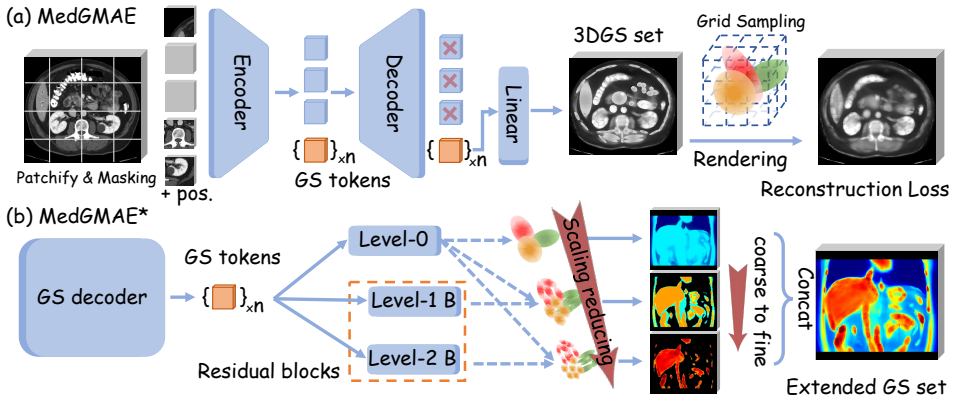

Figure 2: MedGMAE architecture. (a) MedGMAE pre-training framework that processes patchified and masked input through an encoder-decoder architecture to predict 3D Gaussian parameters, which are then rendered and optimized via reconstruction loss. (b) Extended MedGMAE* with multi-level residual blocks for progressive Gaussian parameters refinement.

2 Related Work

2.1 Patch-based masked image modeling

Masked autoencoders learn representations by reconstructing masked input regions. MAE (He et al., 2022) uses a ViT encoder (Dosovitskiy et al., 2020) processing only 25% visible patches and a lightweight decoder for reconstruction, enabling efficient pretraining. In medical imaging, existing self-supervised methods focus on different architectural designs and masking strategies to improve representation learning (Zhou et al., 2023b; Xie et al., 2022b; Tang et al., 2024; Tian et al., 2023; Tang et al., 2022a; Goncharov et al., 2023). Despite their architectural variations and different masking mechanisms, all these approaches remain fundamentally constrained by the voxel-level reconstruction objective, which encourages local interpolation rather than global structural understanding of anatomical features. We propose a fundamentally different approach using 3D Gaussian parameter prediction instead of voxel-level reconstruction. Unlike discrete intensity prediction, our method reconstructs anatomy through continuous geometric primitives, enabling structured representation learning aligned with anatomical continuity.

2.2 3D Gaussian splatting for medical imaging

3D Gaussian Splatting (3DGS) was developed for rendering 3D natural scenes (Kerbl et al., 2023). This approach has since been applied across diverse medical reconstruction scenarios, including 3D CT reconstruction (Li et al., 2025; Cai et al., 2024; Zha et al., 2024), coronary artery reconstruction (Fu et al., 2024), and 4D CT reconstruction (Fu et al., 2025; Yu et al., 2025).

Limitations and motivation. While Gaussian Masked Autoencoders ((Rajasegaran et al., 2025)) pioneered this approach for 2D images—using the z-axis of 3D Gaussians to infer abstract 2.5D

layers for spatial understanding—our motivation is fundamentally different. Instead of inferring abstract structure, we leverage the continuous and parameter-efficient nature of 3D Gaussians to holistically represent true 3D anatomical volumes, directly addressing the limitations of discrete voxel models for capturing continuous anatomy. This objective is better suited for downstream 3D tasks like segmentation and registration, and also unlocks a novel application in accelerating CT reconstruction. Our key insight is: leverage 3D Gaussian primitives as intermediate representations for masked autoencoder pre-training, learning geometric structures rather than discrete voxels. This shifts the objective from local reconstruction to geometric reasoning, encouraging spatial reasoning and anatomical priors while addressing initialization challenges in existing 3DGS medical methods.

3 METHOD

3.1 PRELIMINARIES

3D Gaussian primitives and volume rendering. In medical imaging domain, each 3D Gaussian primitives is parameterized by a center position $\mu \in \mathbf{R}^3$ and a covariance matrix $\Sigma \in \mathbf{R}^{3\times3}$, which jointly define the spatial distribution and morphology of the Gaussian primitive. In addition, each Gaussian carries an intensity value I that represents the intensity at the Gaussian center. In our implementation, we follow the standard practice of decomposing the covariance matrix $\Sigma = RSS^TR^T$ into a scaling matrix $S = \text{diag}(s) \in \mathbf{R}^{3\times3}$ represented by a scale vector $s \in \mathbf{R}^3$, and a rotation matrix $R \in \mathbf{R}^{3\times3}$ parameterized by a rotation quaternion $\phi \in \mathbf{R}^4$. Consequently, each Gaussian is represented by an 11-dimensional parameter vector $g = \{\mu, s, \phi, I\} \in \mathbf{R}^{11}$. For volumetric rendering, we reconstruct the complete 3D volume by evaluating the Gaussian field at discrete grid positions corresponding to the target volumetric dimensions. Each Gaussian contribution to any spatial position X is mathematically described by:

$$G_i(X|g_i) = I_i \cdot e^{-\frac{1}{2}(X-\mu_i)^T\Sigma_i^{-1}(X-\mu_i)}, \tag{1}$$

where $X \in \mathbf{R}^3$ denotes a position in the 3D space, and $g_i = \{\mu_i, s_i, \phi_i, I_i\}$ represents the parameters of the i-th Gaussian. The exponential term defines the spatial decay of the Gaussian influence based on the Mahalanobis distance from its center, naturally encoding the ellipsoidal shape through the covariance structure. The final volumetric intensity is computed as a spatially-localized aggregation of contributions from nearby Gaussians:

$$V(X|g_i) = \sum_{i:||X-\mu_i||\leq d_i} G_i(X|g_i), \tag{2}$$

where d_i defines the effective radius of influence for each Gaussian, typically set based on the eigenvalues of the covariance matrix to ensure computational efficiency while maintaining rendering quality. This localized aggregation strategy enables efficient rendering by avoiding computations for Gaussians with negligible contributions, making the differentiable rendering process tractable for large-scale medical volumes.

3.2 PROPOSED APPROACH

We propose MedGMAE, a framework that replaces voxel-level reconstruction with 3D Gaussian parameters prediction for medical volumetric representation learning in Fig.2(a).

MedGMAE representation learning: The model consists of a Vision Transformer (ViT)-based encoder, a lightweight Transformer decoder, and a differentiable Gaussian renderer specifically designed for volumetric medical data reconstruction. For a given 3D medical image patch with dimensions $96 \times 96 \times 96$, we first patchify it into N non-overlapping patches of size $12 \times 12 \times 12$, resulting in $N = 512$ patches. We then randomly mask these patches with a masking ratio r, typically set to 0.75, yielding n visible patches where $n = N \times (1 - r)$. The ViT encoder processes only the visible patches and encodes them from raw patch representations to latent embeddings $x_i \in \mathbf{R}^{d_{enc}}$, where $i \in \{1, 2, 3, \ldots, n\}$. The decoder employs k learnable query tokens $q_j \in \mathbf{R}^{d_{dec}}$, $j \in \{0, 1, 2, \ldots, k-1\}$, where k represents the number of 3D Gaussians to be predicted. Importantly, k can be set to any value independent of the number of masked tokens, providing flexibility in controlling the reconstruction granularity. We project the encoder latent embeddings to the decoder

dimension space as $\hat{x}_i \in \mathbf{R}^{d_{dec}}$ and construct the decoder input by concatenating three components: the encoder class token, the learnable Gaussian query tokens, and the remaining encoder tokens:

$$X_{dec} = \{\hat{x}_1\} \cup \{q_j\}_{j=1}^{k} \cup \{\hat{x}_i\}_{i=2}^{n} \tag{3}$$

The decoder processes the X_{dec} tokens through multiple Transformer blocks with multi-head self-attention mechanisms. This allows the query tokens to attend to the visible patch embeddings and aggregate spatial-semantic information necessary for accurate 3D Gaussian parameter prediction. The decoder outputs k sets of Gaussian parameters, with each query token predicting one 3D Gaussian primitive through dedicated parameter heads. Each predicted 3D Gaussian is an 11-dimensional vector comprising position coordinates $\mu \in \mathbf{R}^3$, anisotropic scaling factors $s \in \mathbf{R}^3$, rotation quaternion $\phi \in \mathbf{R}^4$, and intensity $I \in \mathbf{R}^1$. The conversion from decoder features to Gaussian parameters is accomplished through four specialized linear prediction heads: a Gaussian center head, a scale head, a rotation head, and an intensity head. Each head applies appropriate activation functions to ensure parameter validity: sigmoid activation for positions and densities to constrain values within [0,1], and L2 normalization for rotation quaternions to maintain unit length. To ensure stable training and balanced parameter distributions across the three spatial dimensions, we employ custom initialization strategies for the prediction heads. All heads utilize Xavier uniform initialization for weights, while biases are initialized specifically for each parameter type: position and rotation heads use zero initialization, the scale head employs a constant bias of -1.386 (resulting in approximately 0.2 after sigmoid activation), and the density head uses a bias of -0.405 (yielding approximately 0.5 after sigmoid activation). This initialization scheme promotes consistent scale distributions across x, y, and z dimensions while providing reasonable starting values for Gaussian intensity.

Differentiable volumetric rendering and training: Once we obtain k predicted 3D Gaussians, we employ a differentiable volumetric renderer to reconstruct the 3D medical image. The renderer accumulates the contributions of all Gaussians within the volume space, with each Gaussian influence determined by its spatial extent and intensity. During training, we apply the reconstruction loss only to the originally masked regions, computed as the mean squared error between the rendered volume and the ground truth image. This masked reconstruction objective encourages the model to learn meaningful 3D representations while maintaining computational efficiency.

Extended MedGMAE for reconstruction: For enhanced reconstruction performance, we further present MedGMAE* with multi-level residual blocks (a hierarchically extended MedGMAE structure Hyun & Heo (2024)), utilizing more Gaussians to capture fine-grained volumetric details in Fig.2(b). We define a hierarchical structure with levels $l \in \{0, 1, 2\}$, from coarse to fine granularity, where each level contains a set of Gaussian parameters. Specifically, we establish dependencies between Gaussian parameters of adjacent levels, where Level 0 contains N_0 base Gaussians, Level 1 expands to $m_1 \times N_0$ Gaussians, and Level 2 expands to $m_2 \times N_0$ Gaussians. We model the 3D representation in a coarse-to-fine manner by assigning coarser- and finer-level Gaussians for coarser and finer details. For scale parameters, we enforce hierarchical reduction as:

$$s^l = s^0 + \hat{s}^l \cdot \sigma_{scale} - \Delta s^l \tag{4}$$

where $\sigma_{scale} = 0.1$ controls the residual magnitudes, and $\Delta s^l > 0$ to ensure monotonic scale reduction. $\Delta s^1 = 0.02$ for Level 1 and $\Delta s^2 = 0.05$ for Level 2 are adopted. For other parameters, we compute new positions as: $\mu^l = \mu^0 + \hat{\mu}^l \cdot \sigma_\mu$,where $\hat{\mu}^l$ are the predicted residual position parameters. We define residual transformations as: $I^l = I^0 + \hat{I}^l \cdot \sigma_I, \quad \phi^l = \text{normalize}(\phi^0 + \hat{\phi}^l \cdot \sigma_{rot})$ where $\sigma_\mu, \sigma_I, \sigma_{rot}$ control the residual magnitudes. Note that all residual prediction modules adopt tanh activation functions to ensure bounded residual outputs and stable training dynamics. This hierarchical densification enables coarse-to-fine reconstruction while maintaining spatial coherence through base Gaussian constraints, significantly enhancing the model's ability to capture fine-grained details in CT reconstruction.

4 Experiments

4.1 Datasets

Pre-training datasets. For self-supervised pre-training, we utilize the AbdomenAtlas1.0Mini dataset (Li et al., 2024a), which contains 5,195 CT scans. All scans are first resampled spacing

Table 1: Comparison of different methods with different proportions on AMOS (Ji et al., 2022), FLARE'22 (Ma et al., 2024), BTCV (Landman et al., 2015) and SegTHOR (Lambert et al., 2020). The DSC (%) is reported. **val** (bold) / val (underline) : top method / second method. † denotes we utilize official pre-training weights. ‡ denotes the results are copied from VoCo (Wu et al., 2024b).

Pretrain Method	AMOS			FLARE'22			BTCV			SegTHOR		
	1%	10%	100%	1%	10%	100%	1%	10%	100%‡	1%	10%	100%
Training from scratch												
UNETR	23.67	60.06	77.02	22.47	56.46	70.81	28.05	42.85	79.82	42.31	72.72	85.82
SwinUNETR	28.94	63.45	82.51	35.89	63.38	75.38	27.71	51.33	80.53	44.82	73.93	87.35
General self-supervised methods												
SparK	36.14	71.68	84.07	36.48	71.74	80.67	30.69	51.26	-	44.76	80.36	88.08
MAE	54.67	72.94	83.61	62.35	77.01	82.56	62.04	75.01	-	66.72	83.60	88.52
Medical self-supervised methods												
MG†	25.72	46.94	62.99	27.30	48.18	57.33	29.27	38.04	81.45	36.96	60.16	83.79
TransVW†	18.72	66.91	82.58	4.81	62.07	75.78	5.63	8.42	-	8.91	31.30	87.46
UniMiSS†	29.49	66.34	79.92	24.92	60.99	74.71	32.95	47.08	-	42.92	76.59	84.34
SUP†	25.60	64.95	82.45	33.72	60.35	74.96	28.75	49.67	81.54	41.74	73.46	87.22
PCRLv2†	21.07	39.07	54.14	27.71	42.97	54.29	24.01	30.48	81.74	40.22	74.71	85.77
GVSL†	24.25	63.45	81.38	26.33	59.54	73.27	24.86	41.79	81.87	42.56	77.40	86.98
vox2vec†	32.76	62.30	74.78	34.11	61.99	70.33	35.29	51.77	-	47.21	73.98	86.77
HySparK†	34.50	64.32	**85.58**	37.54	73.60	82.35	35.81	51.54	-	58.81	**83.95**	88.74
VoCo†	55.81	73.34	84.44	57.66	**78.84**	83.12	**73.20**	**77.85**	**83.85**	67.12	83.87	88.70
MedGMAE	**58.79**	**75.65**	84.90	**62.72**	78.72	**83.77**	66.19	77.11	83.22	**70.92**	83.91	**89.15**

of 1.5mm×1.5mm× 1.5mm using trilinear interpolation. The Hounsfield Unit (HU) values are then clipped to the range [-175, 250]. Finally, the intensity values are normalized to the range [0, 1].

Downstream datasets. For segmentation tasks, we conduct experiments on four public datasets: AMOS (Ji et al., 2022), FLARE'22 (Ma et al., 2024), BTCV (Landman et al., 2015), and SegTHOR (Lambert et al., 2020), with official training-validation split with 1%, 10% and 100% proportions. Medical image classification tasks are conducted on the CT-RATE dataset (Hamamci et al., 2024) with official data partition. For registration tasks we perform experiments on IXI (Kim et al., 2021) and OASIS (Marcus et al., 2007) with same data split as (Wu et al., 2024a). Also, CT reconstruction experiments are conducted on the low-dose Chest and Abdomen CT: AAPM-Mayo dataset (Moen et al., 2021).

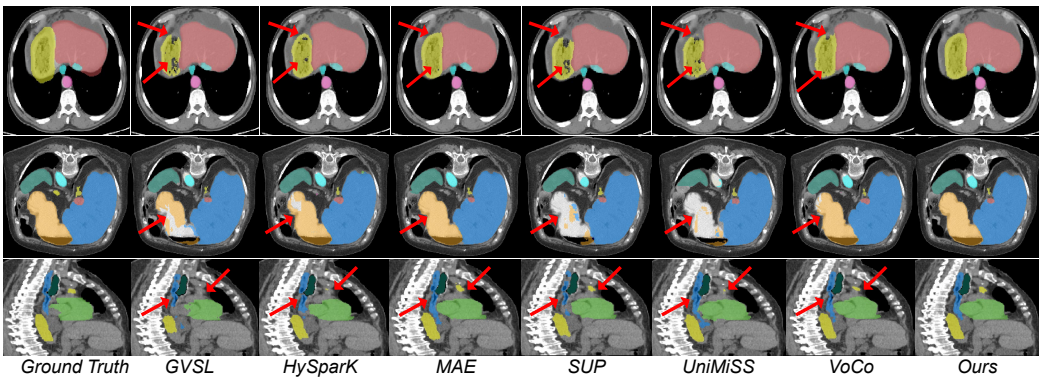

Figure 3: Visualization of one-shot segmentation results for AMOS (row 1), FLARE'22 (row 2) and SegTHOR (row 3).

4.2 Implementation Details

For pre-training, we sample the pre-training volumes into $96 \times 96 \times 96$ patches by ratios of positive and negative as 3:1 in 8 sub-crops. Augmentation probabilities for random flip, rotation, intensities scaling, and shifting are set to 0.5, 0.3, 0.1, 0.1, respectively. We use the AdamW optimizer, an initial learning rate of 1e-4, and a cosine-annealing scheduler for all experiments. The pre-training use a batch size of 192 and train the model for 400K steps. All experiments use a fixed random seed of 41 to ensure reproducibility. We evaluate our method using task-specific metrics: Dice Similarity Coefficient (DSC) for segmentation, Area Under the Curve (AUC) for classification, DSC for registration, and Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) for reconstruction tasks.

Table 2: Performance comparison on CT-RATE dataset for classification task. The AUC (%) is shown. **val** (bold) / <u>val</u> (underline) : top method / second method. † denotes official pre-training weights.

Method		AUC
Scratch	UNETR	71.43
	SwinUNETR	74.29
Fine-tuning	VoCo-10K†	72.11
	VoCo-160K†	76.02
	SUP†	<u>76.04</u>
	MedGMAE	**76.40**

Table 3: The DSC(%) of registration on IXI and OASIS datasets. ‡ denotes the results are copied from VoCo Wu et al. (2024a). The best results are in **bold**.

Method	IXI	OASIS
Training From Scratch		
VoxelMorph†	71.5	78.6
TransMorph†	74.5	81.6
SwinUNETR†	72.6	81.8
Fine-tuning		
SUP†	67.7	81.5
SuPreM†	72.9	81.2
VoCo†	73.6	84.4
MedGMAE	**73.7**	**85.7**

For downstream transfer tasks, we adopt UNETR (Hatamizadeh et al., 2022) as the baseline network followed as (Chen et al., 2023; Tang et al., 2024). For segmentation task, all the pre-processing strategies are the same as (Tang et al., 2022b). For classification task, We resample the volume to $1.5 \times 1.5 \times 3.0$ mm, clip the HU range to $[-1000, 1000]$ and rescale it to [0, 1]. The volume size is set to be $192 \times 192 \times 96$. The model is trained for 100 epochs with a batch size of 96, using AdamW as the optimizer with a learning rate of 3e-2 and weight decay of 0.05. For registration task, our registration algorithm based on TransMorph (Chen et al., 2022) and all the registration pre-processing and training strategies are the same as (Wu et al., 2024a). All experiments are conducted on NVIDIA H20 GPUs.

For CT reconstruction task, we follow the experimental setup of (Li et al., 2025). The key difference is that FBP reconstruction results are cropped using non-overlapping sliding windows and fed into MedGMAE as input. The output Gaussian parameters undergo volume rendering and are concatenated back to original size. To reduce FBP artifact interference, we conduct experiments with 80, 120, and 160 projections. All experiments are trained for 15,000 iterations on Nvidia 3090 GPUs. We evaluate the original 3DGR, 3DGR with MedGMAE initialization, and 3DGR with MedGMAE* initialization using training time, iterations required to reach PSNR=35 and SSIM=90%, and final PSNR and SSIM after complete training.

Comparison with state-of-the-art methods. We select both general and medical self-supervised learning methods for comprehensive comparison. Following (Wu et al., 2024b), we select UNETR (Hatamizadeh et al., 2022), SwinUNETR (Tang et al., 2022a) as compared baseline model. For segmentation tasks, we compare against prominent masked image modeling (MIM) methods, including MAE (He et al., 2022; Chen et al., 2023) and SparK Tian et al. (2023), under identical experimental settings. Additionally, we select nine recent and well-known self-supervised methods: Models Genesis (MG) (Zhou et al., 2021), TransVW (Haghighi et al., 2021), UniMiSS (Xie et al., 2022a), Swin UNETR Pretrained method (SUP) (Tang et al., 2022a), PCRLv2 (Zhou et al., 2023a), GVSL (He et al., 2023), vox2vec (Goncharov et al., 2023), HySparK (Tang et al., 2024), and VoCo (Wu et al., 2024b). For registration tasks, we compare against methods trained from scratch, including VoxelMorph (Balakrishnan et al., 2019), TransMorph (Chen et al., 2022), and SwinUNETR (Hatamizadeh et al., 2021), as well as methods with pre-training such as SUP (Tang et al., 2022b), SuPreM (Li et al., 2024b), and VoCo (Wu et al., 2024b). To ensure fair comparison, we utilize official implementations and loaded official pre-trained weights for all medical SSL methods before fine-tuning.

5 Results

5.1 Promising downstream transfer results

Medical image segmentation. Following previous work, we fine-tuned pre-trained models using 1%, 10%, and 100% of the training data on AMOS, FLARE'22, BTCV, and SegTHOR datasets, respectively. The segmentation results are presented in Table 1. MedGMAE achieves the best or

Table 4: Comprehensive reconstruction comparison across different projection views. 3DGR refers to the initialization method employed in the original paper Li et al. (2025), whereas MedGMAE and MedGMAE* indicate initialization using Gaussian points estimated through zero-shot inference by our proposed model. Values are reported as mean ± standard deviation. The best results are in **bold**.

Method	Time(min)	iter(P=35)	iter(S=90%)	PSNR(full)	SSIM(full)
80 projections					
3DGR	$397_{\pm 39.5}$	$1670_{\pm 371.6}$	$1140_{\pm 145.7}$	$\mathbf{44.6}_{\pm 1.19}$	$\mathbf{98.4}_{\pm 0.32}$
MedGMAE	$303_{\pm 30.9}$	$1135_{\pm 95.0}$	$1100_{\pm 102.5}$	$43.9_{\pm 1.01}$	$97.5_{\pm 0.51}$
MedGMAE*	$\mathbf{251}_{\pm 19.8}$	$\mathbf{990}_{\pm 137.5}$	$\mathbf{820}_{\pm 60.0}$	$44.1_{\pm 0.97}$	$98.0_{\pm 0.38}$
120 projections					
3DGR	$507_{\pm 47.8}$	$1660_{\pm 382.6}$	$1150_{\pm 206.5}$	$45.2_{\pm 1.49}$	$98.7_{\pm 0.32}$
MedGMAE	$357_{\pm 22.0}$	$1040_{\pm 91.7}$	$980_{\pm 60.0}$	$\mathbf{46.2}_{\pm 1.17}$	$98.5_{\pm 0.29}$
MedGMAE*	$\mathbf{335}_{\pm 20.4}$	$\mathbf{920}_{\pm 74.8}$	$\mathbf{780}_{\pm 32.0}$	$45.8_{\pm 1.15}$	$\mathbf{98.7}_{\pm 0.27}$
160 projections					
3DGR	$594_{\pm 140.5}$	$1711_{\pm 449.8}$	$1137_{\pm 211.8}$	$45.1_{\pm 1.53}$	$98.7_{\pm 0.34}$
MedGMAE	$\mathbf{373}_{\pm 20.5}$	$1055_{\pm 85.7}$	$967_{\pm 69.9}$	$\mathbf{46.8}_{\pm 1.36}$	$98.7_{\pm 0.25}$
MedGMAE*	$388_{\pm 36.6}$	$\mathbf{960}_{\pm 96.8}$	$\mathbf{780}_{\pm 33.1}$	$45.8_{\pm 1.39}$	$\mathbf{98.7}_{\pm 0.21}$

second-best DSC scores among all compared methods across different data regimes. Notably, our method demonstrates particularly strong performance in low-data scenarios, outperforming the previous best method VoCo by 2.98% and 5.06% on AMOS and FLARE'22, respectively, with 1% data. Compared to training from scratch baselines Hatamizadeh et al. (2022; 2021), our pre-trained MedGMAE demonstrates substantial improvements, with gains of 20-35% in 1% data scenarios across all datasets. Even with full training data, pre-training consistently provides meaningful improvements of 2-8% over the corresponding from-scratch methods. These results demonstrate that our method could learn a strong anatomical representation by using Gaussian representation. Fig. 3 shows the visualization results.

Medical image classification. Table 2 presents the performance comparison on the CT-RATE dataset. Compared to training from scratch, MedGMAE shows substantial improvements over the best scratch-trained baseline Swin-Bv2 by 2.11%. Among pre-trained methods, MedGMAE surpasses the previous best performers VoCo-160K and SUP by 0.38% and 0.36% respectively, demonstrating the effectiveness of our pre-training approach.

Medical image registration. Table 3 presents the DSC performance comparison on IXI and OASIS datasets for medical image registration tasks. MedGMAE achieves the best performance on OASIS and competitive results on IXI. Compared to the best scratch-trained baselines, our method provides substantial improvements of 1.1% on IXI and 3.9% on OASIS. Among pre-trained methods, MedGMAE outperforms the previous state-of-the-art VoCo by 1.3% on OASIS, confirming the effectiveness of our pre-training approach for medical image registration tasks. It worth noting that both IXI and OASIS are from *unseen* MRI modality, which demonstrates the generalization ability of MedGMAE.

5.2 Geometry-aware zero-shot initialization for 3DGS-based medical image reconstruction

As shown in Table 4 and Fig. 4, MedGMAE demonstrates significant acceleration in training convergence across all projection settings. For training efficiency, MedGMAE reduces training time by 31.0%, 35.0%, and 37.2% compared to 3DGR baseline with 80, 120, and 160 projections respectively. More importantly, MedGMAE substantially accelerates convergence speed, requiring 39.4% and 28.1% fewer iterations to reach PSNR=35 and SSIM=90% benchmarks on average. The residual-extended MedGMAE* further improves convergence performance, achieving even faster iteration counts for quality thresholds while maintaining comparable final reconstruction quality. These results demonstrate that our pre-training approach significantly enhances training efficiency for 3D Gaussian representation-based CT reconstruction without compromising final image quality. Statistical analysis using t-tests revealed that our proposed MedGMAE initialization methods significantly outperformed 3DGR ($p < 0.001$) in traiing efficiency.

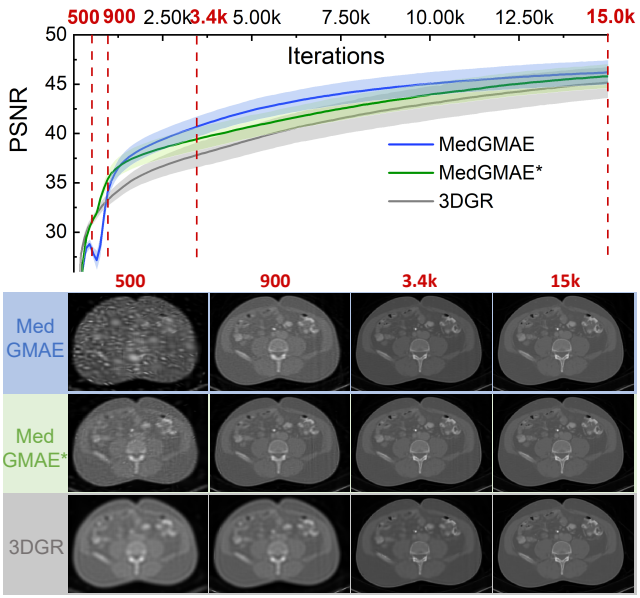

Figure 4: CT reconstruction convergence analysis on AAPM-Mayo dataset. Top: Average PSNR curves with standard error bands showing reconstruction quality improvement over training iterations for different methods. Bottom: Visual comparison of reconstructed CT slices at 500, 900, 3.4k, and 15k iterations for MedGMAE, MedGMAE*, and 3DGR methods, demonstrating the faster convergence and superior reconstruction quality of our approaches.

5.3 Ablation Study

Table 5 presents the ablation study results on MedGMAE components across three segmentation datasets. Adding voxel-based SSL provides substantial improvements of 6-12% over the baseline. Our proposed Gaussian-based SSL further enhances performance by 1-2% compared to voxel-based approaches, confirming the superiority of 3D Gaussian representation over voxel-based reconstruction.

Table 5: Transfer ablation on MedGMAE. The DSC (%) is reported.

Proxy		SSL	AMOS	FLARE'22	SegTHOR
Voxel	Gaussian				
			77.02	70.81	85.82
✓		✓	83.61	82.56	88.52
	✓	✓	84.90	83.77	89.15

6 Conclusion

In this paper, we present MedGMAE, a novel self-supervised pre-training framework that replaces voxel-level reconstruction with 3D Gaussian representation. Leveraging the more efficient and continuous 3D Gaussian primitives, MedGMAE achieves promising encoder transfer performance on diverse downstream tasks including segmentation, classification, and registration. Besies, the transferable decoder enables a 1.39× acceleration compared to original 3DGR-CT reconstruction methods. Extensive experimental results demonstrate the effectiveness of MedGMAE across multiple medical imaging applications. However, in CT reconstruction tasks, the result are affected by noise from FBP reconstruction, which could be improved by training a multi-view 3D Gaussian foundation model.

REFERENCES

Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

Yuanhao Cai, Yixun Liang, Jiahao Wang, Angtian Wang, Yulun Zhang, Xiaokang Yang, Zongwei Zhou, and Alan Yuille. Radiative gaussian splatting for efficient x-ray novel view synthesis. In *European Conference on Computer Vision*, pp. 283–299. Springer, 2024.

Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical image analysis*, 82:102615, 2022.

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*, pp. 1970–1980, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Xueming Fu, Yingtai Li, Fenghe Tang, Jun Li, Mingyue Zhao, Gao-Jun Teng, and S Kevin Zhou. 3dgr-car: Coronary artery reconstruction from ultra-sparse 2d x-ray views with a 3d gaussians representation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 14–24. Springer, 2024.

Xueming Fu, Pei Wu, Yingtai Li, Xin Luo, Zihang Jiang, Junhao Mei, Jian Lu, Gao-Jun Teng, and S. Kevin Zhou. Dyna3dgr: 4d cardiac motion tracking with dynamic 3d gaussian representation, 2025. URL https://arxiv.org/abs/2507.16608.

Mikhail Goncharov, Vera Soboleva, Anvar Kurmukov, Maxim Pisov, and Mikhail Belyaev. vox2vec: A framework for self-supervised contrastive learning of voxel-level representations in medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 605–614. Springer, 2023.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE transactions on medical imaging*, 40(10):2857–2868, 2021.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *CoRR*, 2024.

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*, pp. 272–284. Springer, 2021.

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*, pp. 574–584, 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*, pp. 16000–16009, 2022.

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*, pp. 9538–9547, 2023.

Sangeek Hyun and Jae-Pil Heo. Gsgan: Adversarial learning for hierarchical generation of 3d gaussian splats. *Advances in Neural Information Processing Systems*, 37:67987–68012, 2024.

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.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Boah Kim, Dong Hwan Kim, Seong Ho Park, Jieun Kim, June-Goo Lee, and Jong Chul Ye. Cyclemorph: cycle consistent unsupervised deformable image registration. *Medical image analysis*, 71:102036, 2021.

Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: Segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–6. Ieee, 2020.

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, pp. 12. Munich, Germany, 2015.

Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, 97:103285, 2024a.

Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation. In *The Twelfth International Conference on Learning Representations*, volume 1, 2024b.

Yingtai Li, Xueming Fu, Han Li, Shang Zhao, Ruiyang Jin, and S Kevin Zhou. 3dgr-ct: Sparse-view ct reconstruction with a 3d gaussian representation. *Medical Image Analysis*, pp. 103585, 2025.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.

Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Mae, Adamo Young, Cheng Zhu, Xin Yang, Kangkang Meng, Ziyan Huang, et al. Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the flare22 challenge. *The Lancet Digital Health*, 6(11):e815–e826, 2024.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

Taylor R Moen, Baiyu Chen, David R Holmes III, Xinhui Duan, Zhicong Yu, Lifeng Yu, Shuai Leng, Joel G Fletcher, and Cynthia H McCollough. Low-dose ct image and projection dataset. *Medical physics*, 48(2):902–911, 2021.

Jathushan Rajasegaran, Xinlei Chen, Rulilong Li, Christoph Feichtenhofer, Jitendra Malik, and Shiry Ginosar. Gaussian masked autoencoders, 2025. URL https://arxiv.org/abs/2501.03229.

Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21, 2016.

Fenghe Tang, Ronghao Xu, Qingsong Yao, Xueming Fu, Quan Quan, Heqin Zhu, Zaiyi Liu, and S Kevin Zhou. Hyspark: Hybrid sparse masking for large scale medical image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 330–340. Springer, 2024.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. 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, 2022a.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. 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, 2022b.

Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing bert for convolutional networks: Sparse and hierarchical masked modeling. *arXiv preprint arXiv:2301.03580*, 2023.

Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020.

Linshan Wu, Jiaxin Zhuang, and Hao Chen. Large-scale 3d medical image pre-training with geometric context priors. *arXiv preprint arXiv:2410.09890*, 2024a.

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*, pp. 22873–22882, 2024b.

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*, pp. 558–575. Springer, 2022a.

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*, pp. 9653–9663, 2022b.

Weihao Yu, Yuanhao Cai, Ruyi Zha, Zhiwen Fan, Chenxin Li, and Yixuan Yuan. X^{2}-gaussian: 4d radiative gaussian splatting for continuous-time tomographic reconstruction. *arXiv preprint arXiv:2503.21779*, 2025.

Ruyi Zha, Tao Jun Lin, Yuanhao Cai, Jiwen Cao, Yanhao Zhang, and Hongdong Li. R^2-gaussian: Rectifying radiative gaussian splatting for tomographic reconstruction. *arXiv preprint arXiv:2405.20693*, 2024.

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, 2023a.

Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. 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, 2023b.

S Kevin Zhou, Hayit Greenspan, and Dinggang Shen. *Deep learning for medical image analysis*. Academic Press, 2023c.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021.

7 APPENDIX

7.1 ETHICS STATEMENT

This work involves the analysis of human CT scan data exclusively sourced from publicly available datasets. We acknowledge that the representativeness and potential biases present in these public datasets may influence the fairness and generalizability of our proposed model. We encourage future work to validate our methods on more diverse and representative datasets to ensure equitable healthcare outcomes. All datasets used in this study were previously collected with appropriate ethical approvals and consent procedures as documented by the original data contributors. Beyond these considerations, we have identified no additional ethical conflicts or concerns related to this research.

7.2 REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we have made the following efforts: All architectural details, hyperparameters, and training procedures for our proposed method are comprehensively described in Section 3.2 and Section 4.2 of the main paper. For comparative baseline results, we have directly reported performance metrics from the original publications to avoid potential inconsistencies that may arise from reimplementation differences, with proper citations provided throughout. Upon acceptance of this paper, we commit to releasing the complete source code, including training scripts, model implementations, and evaluation, to facilitate full reproducibility of our results.

7.3 NETWORK ARCHITECTURE CONFIGURATION

Encoder Architecture (ViT Large)

Our MedGMAE employs a Vision Transformer (ViT) Large configuration as the encoder backbone. The detailed specifications are presented in Table 6.

Table 6: ViT Large Encoder Configuration Details

Component	Configuration
Embedding Dimension	1536
Number of Attention Heads	16
Number of Transformer Layers	12
MLP Ratio	4.0
Patch Size	$16 \times 16 \times 16$ or $12 \times 12 \times 12$
Input Image Size	$96 \times 96 \times 96$
Number of Patches	512 (for 12^3) or 216 (for 16^3)
Dropout Rate	0.0
Attention Dropout Rate	0.0
Drop Path Rate	0.1

The encoder processes 3D medical images through the following pipeline:

Decoder Architecture (Lightweight Design)

The decoder employs a lightweight Transformer architecture optimized for Gaussian parameter prediction, as detailed in Table 7.

Decoder Input Token Composition

The decoder processes a carefully constructed sequence of tokens:

$$\mathbf{X}_{dec} = \{\mathbf{x}_{cls}\} \cup \{\mathbf{q}_j\}_{j=1}^{512} \cup \{\mathbf{x}_i\}_{i=2}^{n} \tag{5}$$

where:

- $\mathbf{x}_{cls}$: Class token from encoder (1 token)
- $\{\mathbf{q}_j\}_{j=1}^{512}$: Gaussian query tokens (512 tokens)

Table 7: Gaussian Decoder Configuration Details

Component	Configuration
Embedding Dimension	528
Number of Attention Heads	16
Number of Transformer Layers	8
MLP Ratio	4.0
Number of Gaussian Query Tokens	512
Encoder-to-Decoder Projection	$1536 \rightarrow 528$ Linear Layer
Dropout Rate	0.0
Attention Dropout Rate	0.0
Drop Path Rate	0.1

- $\{\mathbf{x}_i\}_{i=2}^{n}$: Remaining visible patch tokens (127 tokens for 75% masking)

Total decoder input length: $1 + 512 + 127 = 640$ tokens.

7.4 Gaussian Parameter Prediction Heads

Four Specialized Prediction Heads

Each Gaussian is parameterized by an 11-dimensional vector comprising position, scale, rotation, and intensity. Four specialized linear heads predict these parameters:

Table 8: Gaussian Parameter Prediction Head Specifications

Parameter	Dimension	Activation	Range	Bias Init
Position ($\boldsymbol{\mu}$)	3	Sigmoid	$[0, 1]^3$	0.0
Scale ($\mathbf{s}$)	3	Sigmoid	$[0, 1]^3$	-1.386
Rotation ($\boldsymbol{\phi}$)	4	L2 Normalize	Unit Quaternion	0.0
Density (α)	1	Sigmoid	$[0, 1]$	-0.405

Custom Initialization Strategy

To ensure balanced parameter distributions across spatial dimensions, we employ specialized initialization:

$$\text{Position Head:} \quad \mathbf{W} \sim \mathcal{U}(-\sqrt{6/d}, \sqrt{6/d}), \quad \mathbf{b} = \mathbf{0} \tag{6}$$

$$\text{Scale Head:} \quad \mathbf{W} \sim \mathcal{U}(-\sqrt{6/d}, \sqrt{6/d}), \quad \mathbf{b} = -1.386 \tag{7}$$

$$\text{Rotation Head:} \quad \mathbf{W} \sim \mathcal{U}(-\sqrt{6/d}, \sqrt{6/d}), \quad \mathbf{b} = \mathbf{0} \tag{8}$$

$$\text{Density Head:} \quad \mathbf{W} \sim \mathcal{U}(-\sqrt{6/d}, \sqrt{6/d}), \quad \mathbf{b} = -0.405 \tag{9}$$

The bias initialization ensures reasonable starting distributions:

- Scale bias of -1.386 results in $\sigma(-1.386) \approx 0.2$ after sigmoid activation
- Density bias of -0.405 results in $\sigma(-0.405) \approx 0.5$ after sigmoid activation

7.5 Differentiable Gaussian Rendering Algorithm

CUDA Implementation Details

Our CUDA implementation employs several optimization strategies:

[h!] CUDA Gaussian Rendering Kernel [1] Gaussian parameters $\{\boldsymbol{\mu}_i, \mathbf{s}_i, \boldsymbol{\phi}_i, \alpha_i\}_{i=1}^{N}$ Grid points $\{\mathbf{x}_j\}_{j=1}^{M}$, Pixel mask $\mathbf{M}$ Rendered intensity grid $\mathbf{I}$ Initialize shared memory buffers for covariance

matrices and centers gaussian_idx $=$ atomicAdd(work_counter, 1) $< N$ Load Gaussian parameters into shared memory Compute bounding box using 2σ rule: $\text{expand}_d = 2.0 \times s_{i,d} \times \text{grid\_size}_d$ for $d \in \{x, y, z\}$ $\text{bounds}_d = [\mu_{i,d} - \text{expand}_d, \mu_{i,d} + \text{expand}_d]$ each voxel $\mathbf{x}_j$ in bounding box $\mathbf{M}[j] = 1$ (masked region) Compute $\Delta\mathbf{x} = \mathbf{x}_j - \boldsymbol{\mu}_i$ Compute power $= -0.5\Delta\mathbf{x}^T\boldsymbol{\Sigma}_i^{-1}\Delta\mathbf{x}$ intensity $= \alpha_i \exp(\text{power})$ atomicAdd($\mathbf{I}[j]$, intensity)

Sparse Rendering Optimization

For masked regions, we implement sparse rendering that only computes intensities for required pixels:

[h!] Sparse Gaussian Rendering [1] Sparse grid points $\{\mathbf{x}_j\}_{j=1}^{M}$ (only masked pixels) Gaussian parameters $\{\boldsymbol{\mu}_i, \mathbf{s}_i, \boldsymbol{\phi}_i, \alpha_i\}_{i=1}^{N}$ Sparse intensity values $\{I_j\}_{j=1}^{M}$ each sparse point j in parallel $I_j = 0$ each Gaussian i Check if point $\mathbf{x}_j$ within 2σ bounds of Gaussian i within bounds Compute intensity contribution and add to I_j

This sparse approach reduces computational complexity from $O(N \times H \times W \times D)$ to $O(N \times M)$ where M is the number of masked pixels (typically $0.75 \times H \times W \times D$).

7.6 Training Configuration

The training parameters are shown in Table 9.

Table 9: Training Hyperparameters

Parameter	**Value**
Batch Size	8
Learning Rate	1×10^{-4}
Weight Decay	0.05
Optimizer	AdamW
Learning Rate Schedule	Cosine Annealing
Warmup Steps	2000
Max Training Steps	100000
Gradient Clipping	1.0
Gaussian Parameters	
Number of Gaussians	512
Maximum Scale	0.5
Temperature (τ)	0.5

7.7 Additional Experimental Results

Downstream Classification Performance on CT-RATE Dataset

CT Reconstruction Performance Analysis

Figure 6 demonstrates the superior performance of MedGMAE in accelerating CT reconstruction convergence. Our method shows significant improvements across different projection views (80, 120, and 160 projections), with MedGMAE achieving faster convergence and better reconstruction quality compared to the baseline 3DGR method.

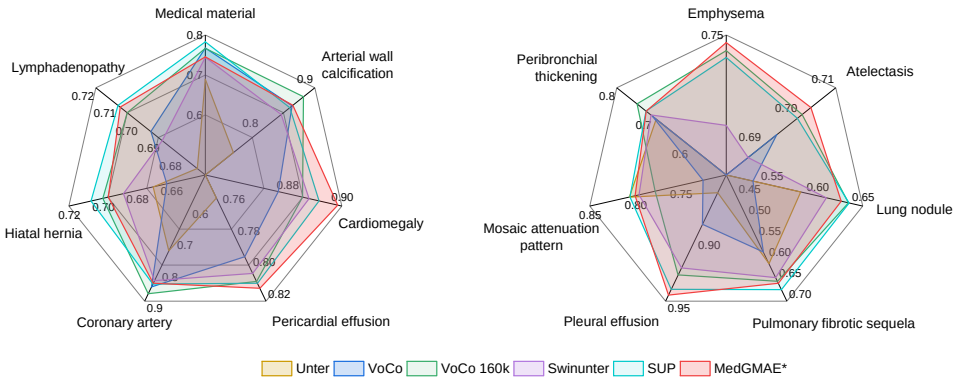

Figure 5: Classification performance comparison on CT-RATE dataset. The radar charts show the Area Under Curve (AUC) scores for different disease categories.

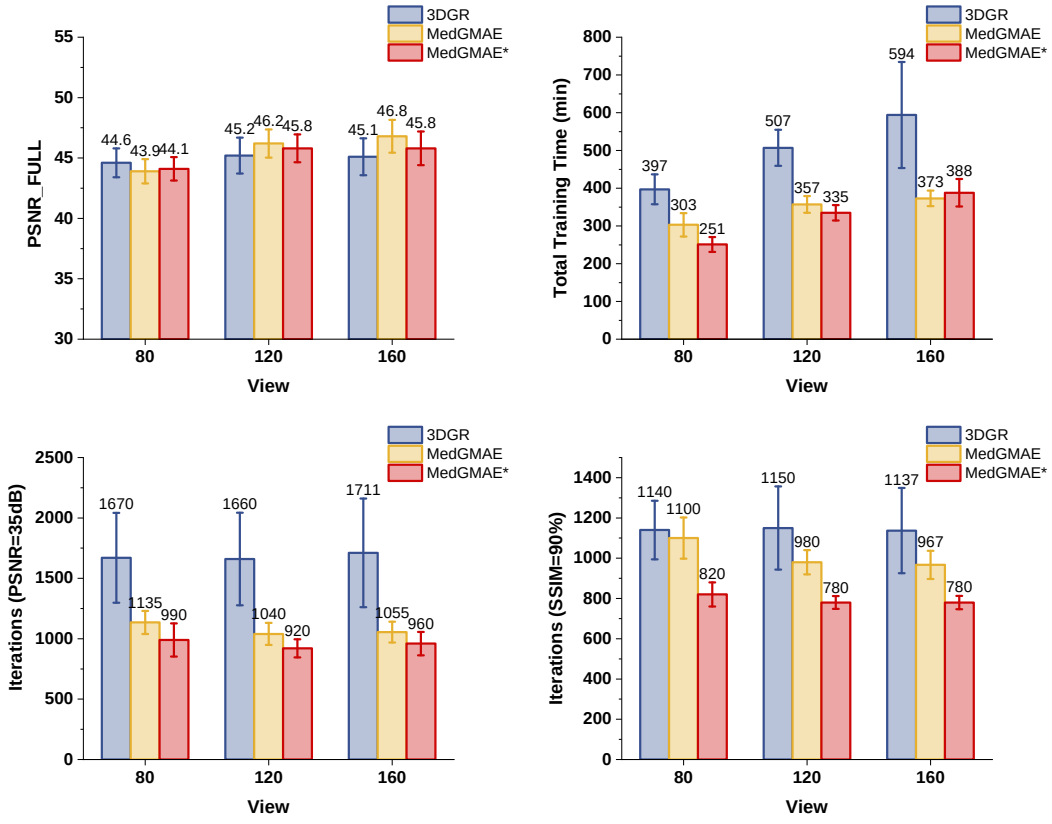

Figure 6: CT reconstruction performance comparison across different projection views. Top row shows PSNR convergence, training time, iterations to reach PSNR=35dB, and iterations to reach SSIM=90%. Bottom row presents the convergence curves and visual reconstruction quality at different iteration stages (500, 900, 3.4k, 15k iterations) for MedGMAE, MedGMAE*, and 3DGR methods.