PG3D-ViT: A Prompt-Guided 3D Vision Transformer for Medical Image Classification

***Abstract*—3D medical image classification is challenging due to small, subtle lesions and substantial irrelevant context, which often mislead deep models. Inspired by the top-down diagnostic process of clinicians—first identifying anatomical context, then locating anomalies—we propose Prompt-Guided 3D Vision Transformer (PG3D-ViT), a framework that simulates clinical reasoning through prompt-driven attention. To address limited 3D training data, PG3D-ViT leverages 2D masked autoencoder (MAE) pretraining to learn transferable image features. Through the prompt generation module, consistency difference analysis is performed between normal and abnormal samples to extract anatomical structure and global spatial prompt information related to the lesion context. These prompts are injected as query into a cross-attention mechanism, guiding the model to focus on lesion-relevant regions across the 3D volume. Evaluated on 7 public datasets spanning multiple modalities and pathologies, PG3D-ViT achieves a 1.88% average AUC improvement over state-of-the-art methods. The attention map visualizations demonstrate that the model can accurately localize lesion regions, validating the effectiveness of the clinical prompting mechanism in enhancing both the performance and interpretability of 3D medical image classification.The code is available at the provided link[1]**

***Index Terms*—3D Medical Image Classification, Vision Transformer, Prompting Mechanism, Masked Autoencoder Pretraining, Cross-Attention Mechanism.**

I. Introduction

In the field of medical image analysis, deep learning-based image classification methods have garnered extensive attention and achieved remarkable progress. Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) have been widely applied to various clinical tasks, such as lesion detection, organ classification, and disease prediction, attaining state-of-the-art performance across multiple benchmark datasets. However, despite their powerful capabilities in modeling both local textures and global dependencies, these methods generally adopt a uniform processing strategy for image regions or patches—treating all parts of the image equally without distinction. This approach overlooks a key characteristic of medical images: clinically significant structures are often spatially sparse (e.g., small lesions occupy only a tiny fraction of the image) and exhibit coarse-grained spatial determinism (e.g., anatomical structures such as the cruciate ligament or pancreas typically appear within specific spatial regions).

In recent years, self-supervised pretraining methods for 3D medical images have received increasing attention. Representative approaches such as masked autoencoding (e.g., Masked Autoencoder, MAE) and contrastive learning aim to mine internal correlations within unlabeled data, thereby enhancing the quality of learned feature representations. These methods have shown strong transferability, particularly in scenarios involving limited data or complex distributions. Nonetheless, their training objectives primarily focus on pixel- or patch-level reconstruction and similarity learning, without explicitly incorporating clinical prior knowledge related to spatial anatomy—for instance, the structural characteristics or spatial positioning of lesions. As a result, when confronted with 3D medical images exhibiting complex anatomical structures and heterogeneous distributions, the discriminative capacity of these methods remains constrained. Fig. 1 illustrates the typical appearance of meniscal injury in knee MRI scans.In (a), the green circle indicates a lesion in the anterior horn of the lateral meniscus, and in (b), the green circle marks a lesion in the posterior horn of the medial meniscus. Among the 27 slices in this MRI case, only these two slices exhibit clear lesion features, indicating that such pathological changes are highly localized in spatial distribution and characterized by sparse and ambiguous expression.

(a) (b)

Fig. 1. **The typical appearance of meniscal injury in knee MRI scans:** (a) The red circle indicates a lesion in the anterior horn of the lateral meniscus. (b) The red circle marks a lesion in the posterior horn of the medial meniscus.

To this end, we propose a clinically inspired prompt-guided 3D Transformer classification model, Prompt-Guided 3D Vision Transformer (PG3D-ViT). The core idea of this approach is to simulate the diagnostic process of clinicians by constructing lesion-context prompt vectors that encode anatomical structure information and spatial priors related to the lesion. These prompts emulate the cognitive patterns of radiologists during image interpretation and serve as auxiliary inputs to guide the Transformer to focus on lesion-relevant regions within the feature space.

To get a deeper understanding of the role of the prompt mechanism within the model, we conducted a visualization analysis of attention distributions under three prompting

[1] https://github.com/UMED-P/PG3D-ViT

conditions. As shown in Fig. 2, the horizontal axis represents different slices from the same MRI volume, while the vertical axis corresponds to attention maps under the global spatial location prompt, anatomical structure prompt, and no-prompt condition (with the no-prompt result derived from a pretrained ViT model). The results show that, under the no-prompt setting, attention is generally dispersed and often concentrated around the image edges. In contrast, with the introduction of global spatial and anatomical prompts, the model's attention becomes significantly more focused on the lesion regions. Notably, under the anatomical prompt, as the slices progress deeper into the volume and the true lesion context gradually shifts to the right, the model's attention also shifts accordingly—consistently tracking the evolving lesion area, demonstrating strong adaptability to structural information and sustained attention across slices.

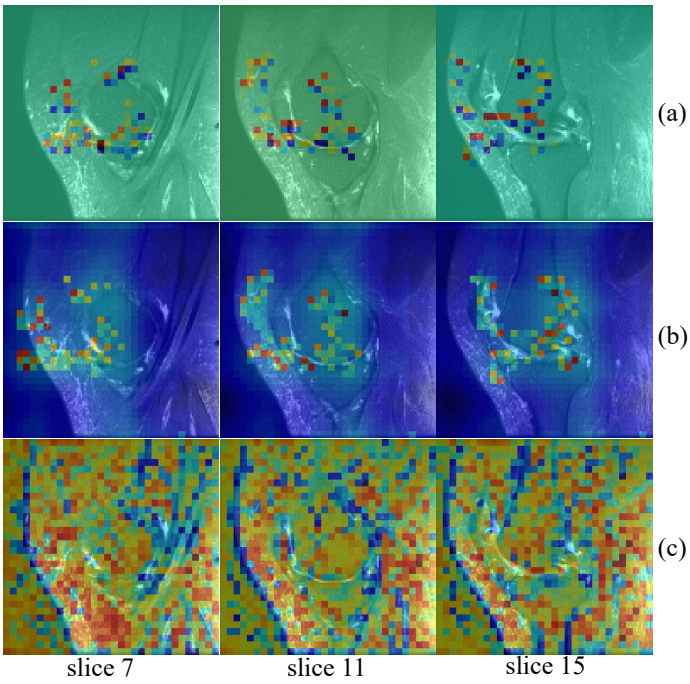

Fig. 2. **Attention Distribution:**(a) Attention distribution under global spatial location prompt, (b) under anatomical structure prompt, and (c) without any prompt. Attention under both global spatial and anatomical prompts is highly focused and concentrated around the lesion region, whereas in the no-prompt condition, attention is dispersed and predominantly distributed along the image edges. The global spatial attention remains largely invariant across slices, while the attention guided by anatomical prompts dynamically shifts in accordance with changes in the lesion context—consistently tracking the lesion across the volume. This demonstrates that the constructed prompt information effectively guides the model to focus on clinically relevant lesion regions.

We conducted a systematic evaluation of PG3D-ViT on seven publicly available medical imaging datasets. Experimental results demonstrate that the proposed method achieves significant performance improvements over existing state-of-the-art approaches in 3D medical image classification tasks, with an average AUC increase of 1.88%. Furthermore, attention heatmap visualizations further confirm the effectiveness of the prompt mechanism, showing that the model can accurately focus on target lesion regions, thereby exhibiting strong localization ability and interpretability.

Based on the above work, our contributions can be summarized as follows:

- We propose a top-down 3D Transformer framework inspired by clinical reasoning, which explicitly incorporates a prompt-guided mechanism to achieve global information fusion.
- We design a prompt generation module that constructs lesion-context prompts by aggregating consistency differences between abnormal and normal samples, encoding both anatomical structure and spatial prior information.
- We develop a 3D global prompt-guided aggregation module that injects the prompt as a Query into the Transformer's cross-attention mechanism, guiding the model to learn more discriminative spatial attention;
- Extensive experiments across various imaging modalities and classification tasks validate the effectiveness of the proposed method, highlighting the unique value of clinically inspired prompting in 3D medical image understanding.

II. Related Work

A. CNN-Based Models in Medical Image Classification

Convolutional Neural Networks (CNNs) [12], [25], [28], [30], [34], [35] have been widely applied in medical image classification tasks, owing to their stable architecture and strong capability for local feature modeling. A representative example is MRNet from Stanford, which segments MRI sequences into three orthogonal views and uses multiple 2D CNNs for fusion, achieving high-precision classification of knee abnormalities and structural damage [1].Residual architectures such as ResNet-3D [12], [13] have been widely adopted as robust baselines in 3D medical imaging tasks, including tumor detection and lung nodule classification. However, CNN has a fixed receptive field constrained by hierarchical stacking, making it insufficient for accurately modeling small-volume lesions in medical images. Additionally, it inherently struggles to capture long-range dependencies between distant structures, limiting its effectiveness in recognizing complex spatial configurations.

B. ViT-Based Models in Medical Imaging

The introduction of Transformer architectures into vision tasks has endowed ViT models with unique advantages in global modeling. The original Vision Transformer (ViT) proposed by Dosovitskiy et al. achieved CNN-level performance on ImageNet without using any convolutional layers, heralding a paradigm shift in vision modeling [4]. In the medical domain, Wang et al. incorporated a feature pyramid structure into ViT to improve multi-task classification accuracy on the MedMNIST series [5], while Manzari et al. proposed MedViT [6], which integrates convolutional perception with global attention, showing improved generalization and robustness compared to traditional CNNs. Despite ViT's innate advantage in capturing long-range dependencies, its lack of spatial inductive bias makes the recognition of small lesion regions highly dependent on effective guidance from the attention

mechanism. When lesions are small in size or exhibit low contrast, the model may struggle to autonomously focus on critical regions, thereby compromising the overall diagnostic performance.

C. Self-Supervised Pretraining for Medical Image Classification

To address the challenge of limited labeled data, self-supervised learning (SSL) [7], [14]–[18] has emerged as a powerful pretraining strategy for both ViTs and CNNs. The MAE model [18], which reconstructs masked regions using raw pixels as targets, has demonstrated strong reconstruction ability even when up to 75% of the input is masked. Zhou et al. proposed the UNIMISS framework, which bridges 2D and 3D medical image training, achieving significant gains in transfer learning [8]. Additionally, large-scale multi-modal models such as SAM [9] and CLIP [10] offer theoretical and practical support for prompt-based mechanisms and semantic fusion. Lai et al. proposed a self-supervised learning method based on residuals from language models [29] , which significantly enhances the feature representation ability and downstream task performance of biomedical imaging models without requiring additional annotations.

However, these methods primarily rely on bottom-up learning of low-level visual patterns—for example, reconstructing masked regions or constructing local representations through contrastive objectives—while lacking explicit clinical semantic prompts or spatial prior guidance. As a result, when dealing with sparsely distributed and structurally complex 3D medical images, the model may overlook subtle yet critical abnormal regions, making it difficult to capture the fine-grained cues essential for accurate diagnosis.

D. Prompting Mechanisms and Their Introduction

In the field of natural language processing (NLP), prompting techniques have been widely adopted to guide large models in completing specific tasks. In computer vision, general-purpose segmentation models such as Segment Anything (SAM) [9] introduced explicit prompts—such as points and boxes—into image space for the first time, which has attracted significant attention in the medical imaging community. Wu et al. proposed the Medical SAM Adapter (Med-SA) [21], which retains the original frozen SAM model while introducing a lightweight prompt adaptation module that incorporates medical knowledge into the network, thereby enhancing segmentation performance in medical imaging tasks. Liu et al. proposed the Segment Any Tissue (SAT) framework [22], which generates prompt points from a single annotated reference image to enable automatic segmentation of arbitrary tissue structures, without the need for retraining—offering strong guidance and adaptability.

Despite recent advances in prompt-based methods for medical image segmentation and classification, existing approaches remain predominantly focused on 2D image segmentation tasks and struggle to capture cross-slice structures and spatial consistency in 3D volumes, limiting their effectiveness in guiding 3D classification. Additionally, the use of low-level prompts such as points or boxes lacks the capacity to encode lesion context or global semantics. Moreover, the loose integration between prompts and model representation learning makes it difficult to achieve effective semantic focus and joint modeling.

To the best of our knowledge, this study is the first to introduce heuristic prompt vectors that simulate the cognitive process of clinicians into a ViT. The proposed lesion-context anatomical prompt functions similarly to an anatomical atlas, providing prior knowledge of lesion morphology, while the spatial location prompt conveys the lesion's position within the 3D volume. Within the model, these prompts explicitly guide attention to focus on lesion-relevant regions. This approach is distinct from existing prompting strategies and, to date, has no direct precedent in the current literature.

III. Methodology

In this study, we propose a clinically inspired 3D medical image classification framework—Prompt-Guided 3D Vision Transformer (PG3D-ViT)—designed to simulate the cognitive process of physicians who integrate structural knowledge and spatial cues during diagnosis. As illustrated in Fig. 3, the PG3D-ViT framework consists of three key components: (1) a self-supervised pretraining stage based on a Masked Autoencoder (MAE), which learns general-purpose image representations; (2) a lesion-context prompt generation module that constructs prompts including anatomical priors and spatial awareness; and (3) a full-space context aggregation module, which incorporates the prompt information through a cross-attention mechanism to guide the model's focus toward lesion-relevant regions in the feature space.

A. Image Representation Pretraining Module

To obtain robust feature representations prior to downstream classification tasks, we first perform self-supervised pretraining on a large collection of unlabeled medical images. We adopt the Masked Autoencoder (MAE) strategy, aiming to learn general-purpose visual representations under unsupervised conditions,and subsequently transfer the pretrained weights to downstream 3D medical image classification tasks.

1) ***Data Preprocessing****:* The input training medical volume is denoted as $T \in \mathbb{R}^{h\times w\times l\times c}$, where h, w, l, and c represent the height, width, number of slices, and number of channels, respectively. Due to the high memory and computational cost of directly processing 3D volumes, we divide the 3D data along the slice axis into l 2D images during the pretraining phase, denoted as $X \in \mathbb{R}^{h\times w\times c}$.

2) ***MAE****:* Each 2D image is further partitioned into n non-overlapping patches of size $p \times p \times c$,and the set of patches is denoted as $\{x_i\}_{i=1}^{n}$.

After the encoder extracts features, its output is combined with a set of mask tokens corresponding to the masked patches, and the associated positional encodings are added. The combined sequence is then fed into a lightweight decoder to reconstruct the original pixels. The reconstruction loss, computed only over the masked regions, is defined in Eq. (1) as follows:

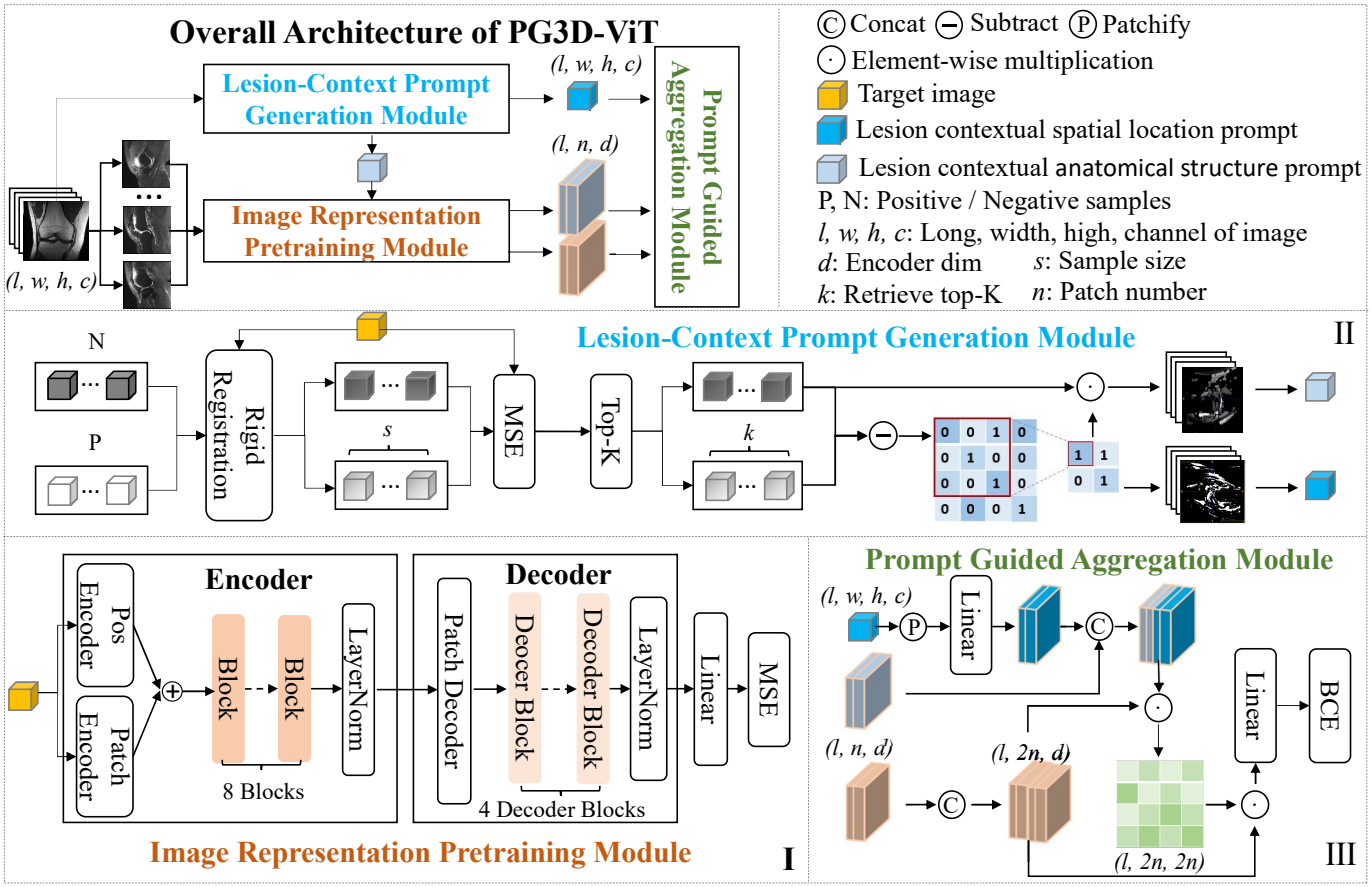

Fig. 3. **Overall Architecture of PG3D-ViT:** The framework consists of three key modules: the image representation pretraining module, the lesion-context prompt generation module, and the prompt-guided global context aggregation module. Ⅰ. The image representation pretraining module leverages a masked autoencoder (MAE) to pretrain the ViT encoder on large-scale 2D medical images. Ⅱ. The lesion-context prompt generation module extracts anatomical structure and spatial prior information related to the lesion to construct prompts. Ⅲ. The prompt-guided aggregation module injects the prompt vectors as Query tokens into the Transformer's cross-attention mechanism, guiding the model to focus on lesion-relevant regions across the entire image volume.For more technical details, please refer to the Methods section.

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\hat{x}_i - x_i\|_2^2 \tag{1}$$

where $\mathcal{M}$ denotes the set of masked patch indices, $\hat{x}_i$ is the reconstructed patch, and x_i is the ground truth.

Considering that most medical images are grayscale, we use single-channel input and adopt a relatively small embedding dimension along with a shallow Transformer architecture to control model complexity. After pretraining, only the encoder parameters are retained as the initialization for the feature extractor in downstream 3D medical image classification tasks.

B. Lesion-Context Prompt Generation Module

To explicitly incorporate lesion-context information, we design a lesion-context prompt module to generate anatomical structure and spatial location cues related to lesions within the 3D volume. This prompt is constructed by aggregating consistency differences between positive samples (containing lesions) and negative samples (healthy sample), thereby capturing contextual priors that reflect lesion-specific characteristics. A dedicated prompt module is constructed for each type of lesion. The detailed generation process is as follows:

1) ***Sample Selection****:* Randomly select s positive volumes $\{V_i^+\}_{i=1}^s$ and s negative volumes $\{V_j^-\}_{j=1}^s$ from the training dataset.

2) ***3D Rigid Registration****:* Each candidate volume is rigidly registered to the target volume T by optimizing the mutual information loss, with the rigid transformation defined in Equation Eq. (2).

$$\mathcal{L}_{\text{MI}} = -I(T, V_i^{+/-}) \tag{2}$$

The voxel-level Mean Squared Error (MSE) is computed as the similarity metric, as defined in Eq. (3):

$$\text{MSE}(T, V_i^{+/-}) = \frac{1}{|\Omega|} \sum_{x \in \Omega} \|T(x) - V_i^{+/-}(x)\|^2 \tag{3}$$

The top k positive and negative samples with the lowest MSE are selected. Their average volumes are computed as shown in Eq. (4) :

$$\bar{V}^+ = \frac{1}{k}\sum_{i=1}^{k} V_i^+, \quad \bar{V}^- = \frac{1}{k}\sum_{j=1}^{k} V_j^- \tag{4}$$

*3) **Difference Map and Prompt Generation**:* A voxel-wise difference map is computed as defined in Eq. (5):

$$D(x) = \frac{|\bar{V}^+(x) - \bar{V}^-(x)|}{\bar{V}^+(x) + \bar{V}^-(x) + \epsilon} \tag{5}$$

where ϵ is a small constant to prevent division by zero. The top-activated voxels are selected to form a difference mask $p(x)$.

*4) **Context Enhancement via Logical Convolution Kernel**:* To enhance the model's perception of lesion-context regions, we introduce a logical convolution kernel for spatial enhancement. Specifically, a sliding window of size $w \times w \times w$ is defined, and a logical operation is applied to each voxel, as formulated in Eq. (6) : if the sum of activation values within the window exceeds a predefined threshold θ, the output at that position is set to 1, otherwise, it is set to 0.

Unlike traditional weighted convolution, logical convolution uses binary activation as its core criterion, effectively expanding the model's receptive field. This mechanism allows the response at the lesion point to propagate into the surrounding space, enabling the model to better extract semantic information from the lesion and its context during subsequent guidance. It is particularly beneficial for detecting small lesions or those with unclear boundaries. Moreover, logical operations are computationally lightweight, offering high efficiency and structural simplicity.

$$p'(x) = \begin{cases} 1, & \sum_{y \in \mathcal{W}_x} p(y) \geq \theta \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $p(y)$ denotes the initial spatial prompt map, θ is a predefined activation threshold, and $p'(x)$ is the spatial prompt after logical convolution enhancement.

Based on the enhanced prompt map $p'(x)$, we extract the lesion-context anatomical structure Eq. (7) from a negtive registration sample $\bar{V}^-(x)$ via spatial masking:

$$S'(x) = \bar{V}^-(x) \cdot p'(x) \tag{7}$$

where $\cdot$ denotes element-wise multiplication. This operation retains only the regions activated by the spatial prompt within the positive sample, effectively capturing anatomical structure information relevant to the lesion. The resulting structural prompt $s'(x)$ is then used as input to the subsequent prompt-guided module, enabling the model to better focus on clinically meaningful regions.

C. Full-Space Prompt-Guided Aggregation Module

The full-space prompt-guided aggregation module is designed to achieve deep fusion between prompt information and image features, guiding the model to focus on lesion-relevant regions across the entire image volume. This mechanism enables the model to explicitly capture the response relationships between the prompt signals and visual features, thereby enhancing its ability to perceive small lesions and complex contextual patterns, and improving overall discriminative performance. The detailed steps are as follows:

*1) **Image, Anatomical, and Spatial Position Representations**:* To fully capture the key semantic information in 3D medical images, we construct multi-dimensional representations from three perspectives: image content, anatomical structure, and spatial position. Specifically, each original image slice is fed into a ViT encoder initialized with MAE pretraining to extract visual feature representations Z_i. Simultaneously, a lesion-context anatomical prompt map is introduced as a structural prior and passed through the same encoder to obtain anatomical structure representations S_i as defined in Eq. (8). In addition, the spatial location map p' is partitioned into patches consistent with the image and linearly projected to match the feature dimension, forming the spatial position representation P_i. These three types of features are subsequently integrated in a cross-attention module to provide rich contextual information for lesion recognition and classification:

$$\begin{aligned} Z_i &= \mathrm{ViT}_{\mathrm{enc}}(T_i) \in \mathbb{R}^{l \times d}, \\ S_i &= \mathrm{ViT}_{\mathrm{enc}}(S_i') \in \mathbb{R}^{l \times d} \end{aligned} \tag{8}$$

*2) **Cross-Attention**:* The anatomical representation S_i and spatial position representation p_i are concatenated along the channel dimension to form the Query, while the image representation Z_i is duplicated and concatenated to serve as both the Key and the Value, as defined in Eq. (9):

$$\begin{aligned} Q_i &= [S_i \parallel P_i] \in \mathbb{R}^{2*l \times d}, \\ K_i = V_i &= [Z_i \parallel Z_i] \in \mathbb{R}^{2*l \times d} \end{aligned} \tag{9}$$

Through the attention mechanism defined in Eq. (10), the full-space prompts are effectively integrated with the image features, facilitating the extraction of critical contextual representations.

$$A_i = \mathrm{Attention}(Q_i, K_i, V_i) = \mathrm{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{D}}\right) V_i \tag{10}$$

The fused features are subsequently aggregated and forwarded to a classification head, where binary cross-entropy, as defined in Eq. (11), is used as the loss function:

$$\mathcal{L}_{\mathrm{cls}} = -\left[y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right] \tag{11}$$

This module injects the prompt information into the network in the form of attention, allowing structural priors to be incorporated in the early stages of feature aggregation. It achieves two levels of integration: on one hand, the semantic fusion between the prompt and the image representation; on the other, the aggregation of global volumetric context across the 3D space, thereby enhancing the model's joint perception of anatomical structures and spatial localization. The cross-attention mechanism enables spatially guided focus based on external prompts, improving the model's ability to capture long-range dependencies between small lesions and complex anatomical structures, thus enhancing both reasoning accuracy

and interpretability.In the multi-class task, each lesion category is guided by its own structural prior, and the final objective is obtained by averaging the Binary Cross-Entropy (BCE) losses computed for each category.

IV. Experiments

A. Experimental Datasets

We conducted a systematic evaluation of the proposed model on medical imaging data from four different modalities, including MRI, CT, MRA, and electron microscope images, covering a wide range of clinical application scenarios. The experimental tasks encompass both 2D and 3D medical image classification, involving various anatomical structures and tissue types such as the brain, abdomen, and vasculature. All datasets used are publicly available and span diverse task types, including disease screening, organ abnormality detection, and small lesion recognition, demonstrating strong research representativeness and practical challenge. An overview of the datasets is presented in Table I.

B. Benchmarking

*1) **Design of Baseline Models**:* We select four representative categories of models as comparative baselines:

*a) **CNN-based models**:* including AlexNet [1], VGG [25], ResNet [12], DenseNet [34], EfficientNet [35], and their 2.5D, 3D, and ACS variants [31], as well as AutoML-based approaches such as Auto-sklearn [32] and Auto-Keras [33]. These models exhibit strong local feature modeling capabilities and are well-suited for tasks with well-defined anatomical structures.

*b) **ViT-based models**:* such as ViT/B-16 [4], MedViT [6], and FPVT [5],ViT 3D [4] leverage global attention mechanisms to enhance lesion perception, making them effective in recognizing complex structures.

*c) **Self-supervised models**:* including MAE [18], DINO [7], EVA [37], and UniMiSS [8], R-LLM [29],employ unsupervised pretraining to improve performance on small-sample and fine-grained recognition tasks.

*d) **Prompt-based models**:* represented by SAM ViT/B-16 [9], enable spatially guided target localization via point or bounding-box prompts.

To ensure reliability, the SOTA performance is mostly based on the original data in the original paper. we implement the model with PyTorch and train it on NVIDIA A100 GPU.

*2) **Quantitative Performance Evaluation**:* We evaluate the classification accuracy and Area Under the Receiver Operating Characteristic Curve (AUC) across multiple publicly available benchmark datasets.

Table II reports the experimental results of our model on the MedMNIST v2 dataset. The AUC scores on AdrenalMNIST3D, NoduleMNIST3D, VesselMNIST3D, and SynapseMNIST3D are improved by 3.50%, 1.80%, 1.20%, and 3.30%, respectively. These results demonstrate the effectiveness of our proposed framework in handling small-sample and high-complexity medical image classification tasks.

On the MRNet dataset, the evaluation metrics of the comparative experiments are shown in Table III. Compared to state-of-the-art (SOTA) models, our method achieves an average AUC improvement of 2.27%.

Our model is compatible with 2D medical imaging data.From Table IV, it can be seen that PG3D-ViT achieved the second-best performance on the OrgansMNIST dataset, while on the BreastMNIST dataset, the AUC improved by 0.4%.The model demonstrates superior performance on challenging datasets.

We conducted a comparative analysis of two mainstream model architectures: the prompt-based SAM and the visual backbone ViT. On the OrgansMNIST and BreastMNIST datasets, the proposed PG3D-ViT achieved AUC improvements of 1.2% and 10.1% respectively over the existing prompt learning model SAM ViT/B, demonstrating its architectural advantages. PG3D-ViT leverages lesion-context prompts constructed from consistency differences between positive (diseased) and negative (healthy) samples, introducing clinically interpretable coarse-grained priors that effectively indicate potential lesion regions. In contrast, SAM ViT/B relies on the self-learned attention without explicit supervision, which lead to dispersed attention or misalignment with diagnostically relevant regions. Moreover, by integrating both anatomical cues and spatial cues (approximate lesion locations), PG3D-ViT effectively guides the Transformer to focus on pathology-relevant regions during both modeling and attention aggregation, resulting in more accurate and robust classification performance.

Compared to ViT-B, PG3D-ViT achieves an AUC improvement of 8.82% on these datasets. ViT-B models all image patches equally using self-attention, lacking structural priors, which limits its ability to detect small or subtle lesions.

Meanwhile, comparisons across different datasets indicate that 3D images significantly outperform 2D images in terms of performance. This is mainly attributed to the fact that training samples for 3D images are more scarce, and the problem of redundant and complex anatomical structures is more pronounced. In this context, prompt-guided modeling effectively helps the model focus on lesion-relevant regions, which aligns well with our expectations.

*3) **Model Complexity and Efficiency Analysis**:* In practical applications, the number of parameters and the overall model size directly impact computational resource consumption, inference speed, and the feasibility of deployment on edge devices. Therefore, in addition to core performance metrics, we conducted a comparative analysis of each model's parameter scale and storage footprint. All evaluation metrics were calculated based on an input size of 20×256×256.

From the perspective of model complexity in Table V, our proposed PG3D-ViT exhibits significantly lower computational burden compared to large-scale models such as VGG and ViT-Base, with a parameter count comparable to that of ResNet. Across multiple medical imaging datasets, models with fewer parameters,such as MRNet, ResNet-18/50+3D, FPVT, MedViT-S, and PG3D-ViT—consistently outperform overparameterized architectures, with PG3D-ViT achieving the best performance in particularly challenging settings involving

TABLE I
SUMMARY STATISTICS OF BENCHMARK DATASETS FOR MEDICAL IMAGE CLASSIFICATION

Statistic	Organ	Modalities	Size	Labels	Train	Test
MRNet [1]	Knee	MRI	~20*256*256	3	1000	120
OrgansMNIST [19]	Liver	Abdominal CT	28*28	11	13,932	8,827
BreastMNIST [19]	Breast	CT	28*28	2	546	156
AdrenalMNIST3D [20]	Ribs	CT	28*28*28	3	1,027	240
NoduleMNIST3D [20]	Lungs	CT	28*28*28	2	1,158	310
VesselMNIST3D [20]	Brain	MRA	28*28*28	2	1,335	382
SynapseMNIST3D [20]	Synapse	Electron Microscope	28*28*28	2	1,230	352

TABLE II
QUANTITATIVE ANYLYSIS ON ADRENALMNIST3D, NODULEMNIST3D, VESSELMNIST3D AND SYNAPSEMNIST3D

Model	AdrenalMNIST3D		NoduleMNIST3D		VesselMNIST3D		SynapseMNIST3D	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
ResNet-18 + 2.5D [20]	0.718	0.772	0.838	0.835	0.748	0.846	0.634	0.696
ResNet-18 + 3D [20]	0.827	0.721	0.863	0.844	0.874	0.877	0.820	0.745
ResNet-18 + ACS [20]	**0.839**	0.754	0.873	0.847	**0.930**	**0.928**	0.705	0.722
ResNet-50 + 2.5D [20]	0.732	0.763	0.835	0.848	0.751	0.877	0.669	0.735
ResNet-50 + 3D [20]	0.828	0.745	0.875	0.847	0.907	0.918	**0.851**	**0.795**
ResNet-50 + ACS [20]	0.828	0.758	0.886	0.841	0.912	0.858	0.719	0.709
auto-sklearn [20]	0.828	0.802	0.914	0.874	0.910	0.915	0.631	0.730
AutoKeras [20]	0.804	0.705	0.844	0.834	0.773	0.894	0.538	0.724
FPVT	0.801	0.704	0.814	0.822	0.770	0.888	0.530	0.712
ViT-3D [29]	0.820	0.819	0.915	0.868	0.826	0.910	/	/
R-LLM [29]	**0.839**	**0.829**	**0.924**	**0.897**	0.837	0.910	/	/
PG3D-ViT [20]	**0.874**	**0.832**	**0.942**	**0.952**	**0.942**	**0.938**	**0.884**	**0.813**

TABLE III
QUANTITATIVE ANYLYSIS ON MRNET: SINCE THE MRNET DATASET WAS ORIGINALLY DESIGNED FOR A COMPETITION AND ITS OFFICIAL TEST SET IS NOT PUBLICLY AVAILABLE, THIS STUDY USES THE ORIGINAL VALIDATION SET AS THE TEST SET. MEANWHILE, THE ORIGINAL TRAINING SET IS RE-SPLIT INTO A NEW TRAINING SET AND VALIDATION SET FOR MODEL TRAINING AND HYPERPARAMETER TUNING.

Model	Abnormal		ACL		Meniscus	
	AUC	ACC	AUC	ACC	AUC	ACC
MRNET	**0.931**	**0.896**	**0.889**	**0.825**	0.762	**0.731**
ResNet50 [27]	0.900	/	0.870	/	**0.820**	/
VGG19	0.915	0.869	0.875	0.803	0.756	0.655
ViT/B-16	0.622	0.532	0.676	0.561	0.638	0.554
MAE ViT/B-16	0.711	0.657	0.428	0.304	0.644	0.527
UniMiss	0.775	0.705	0.632	0.587	0.697	0.562
PG3D-ViT	**0.948**	**0.912**	**0.892**	**0.829**	**0.810**	**0.783**

TABLE IV
QUANTITATIVE ANYLYSIS ON ORGANSMNIST AND BREASTMNIST

Model	OrgansMNIST		BreastMNIST	
	AUC	ACC	AUC	ACC
AlexNet [3]	0.973	0.786	0.891	0.865
VGG16 [19]	0.976	0.789	0.875	0.853
ResNet-18 [19]	0.972	0.782	0.901	0.863
ResNet-50 [19]	0.972	0.770	0.857	0.812
DenseNet-121 [19]	0.972	0.777	0.862	0.833
EfficientNet-B4 [19]	0.951	0.680	0.765	0.765
auto-sklearn [19]	0.945	0.672	0.836	0.803
AutoKeras [19]	0.974	**0.813**	0.871	0.831
Google AutoML Vision [19]	0.964	0.749	0.919	0.861
ViT/B-16	0.971	0.654	0.848	0.821
CLIP ViT/B-16	0.965	0.654	0.752	0.776
EVA-ViT/B-16	0.958	0.719	0.749	0.748
FPVT	0.976	0.785	**0.938**	0.891
MedViT-T	0.972	0.789	0.934	0.896
MedViT-S	**0.987**	0.805	**0.938**	**0.897**
MedViT-L	0.973	0.806	0.929	0.883
DINO ViT/B-16	0.971	0.743	0.884	0.848
R-LLM [29]	/	/	0.882	0.782
SAM ViT/B-16 [36]	0.970	0.824	0.841	0.868
PG3D-ViT	**0.982**	**0.816**	**0.942**	**0.914**

high noise and subtle lesions. This phenomenon suggests that in domains like medical imaging, where training data are often limited, complex models are more susceptible to overfitting and generalize poorly. In contrast, compact models, due to their reduced parameter space, tend to be more stable during training and exhibit stronger robustness under noisy, low-sample conditions. Our model achieves strong performance with relatively low parameter count and computational cost, due to its well-designed architectural framework. In our performance evaluation, we further analyzed the computational overhead of the prompt generation stage. The rigid registration operation takes an average of approximately 3 seconds per image pair in a GPU environment. During training, all prompts can be pre-generated through preprocessing, resulting in a negligible impact on training efficiency. During inference, clinical applications typically tolerate delays on the order of a few seconds without strict real-time requirements. Therefore, the proposed prompt generation strategy is practically feasible and does not pose a substantial barrier to the deployment or

TABLE V
THE MODEL COMPLEXITY ANALYSIS OF BASELINE AND OUR PROPOSED MODEL

CNNS				ViTs			
Model	**Backbone**	**#P(M)**	**#F(G)**	**Model**	**Backbone**	**#P(M)**	**#F(G)**
MRNET	2D AlexNet	60.0	36.0	MAE ViT/B-16	2D ViT-Base	87.0	458
VGG16	2D VGG-16	138.0	405.0	UniMiss	2D Transformer	87.0	458
VGG19	2D VGG-19	144.0	515.0	ViT/B-16	2D ViT-Base	87.0	458
AlexNet	2D CNN	60.0	36.0	CLIP ViT/B-16	2D ViT-Base	87.0	458
ResNet-18	2D CNN	11.7	47.6	EVA-ViT/B-16	2D ViT-Base	87.0	458
ResNet-50	2D CNN	25.6	108	FPVT	CNN + Transformer	20.0	50.0
DenseNet-121	2D CNN	8.1	78.0	MedViT-T	2D ViT-Tiny	10.8	34.0
EfficientNet-B4	2D CNN	19.3	38.0	MedViT-S	2D ViT-Small	23.6	128
auto-sklearn	2D ResNet-18	12.0	48.0	MedViT-L	2D ViT-Large	45.8	350
AutoKeras	2D NAS	20.0	40.0	ViT 3D	2D ViT-Base	87.0	458
Google AutoML Vision	EfficientNet-like	66.0	740	R-LMM	2D ViT-Base	87.0	458
ResNet-18 + 2.5D	2.5D CNN	12.0	48.0	DINO ViT/B-16	2D ViT-Base	87.0	458
ResNet-18 + 3D	3D CNN	34.0	7.1	SAM ViT/B-16	2D ViT-Base/16	87.0	458
ResNet-18 + ACS	3D(ACS)	11.7	2.4	PG3D-ViT	2D ViT	51.4	23.4
ResNet-50 + 2.5D	2.5D CNN	25.6	108				
ResNet-50 + 3D	3D CNN	76.8	16.1				
ResNet-50 + ACS	3D CNN (ACS)	25.6	5.4				

operation of the overall system.

C. Visualization Analysis

To qualitatively assess the model's attention mechanism, we visualize the attention heatmaps generated by PG3D-ViT. Fig. 4 illustrates an example from a knee MRI slice.

We conducted a qualitative analysis of the prompts generated by the model. We present a knee MRI slice in which (a1) and (a2) are the input images to be analyzed, with red circles indicating the lesion areas of the anterior cruciate ligament (ACL) and the meniscus, respectively. (b1) and (b2) show the anatomical prompts generated by the model for the two distinct lesion contexts. Despite the close spatial proximity of the ACL and meniscus, the corresponding anatomical prompts exhibit clear differences. The prompt in (b1) distinctly highlights the characteristics of an ACL tear, with a significantly increased signal in the sagittal view, consistent with clinical diagnostic features. (c1) and (c2) display the spatial location prompts, which exhibit high activation around the ACL and meniscus regions, respectively. This suggests that the model is able to reasonably capture the spatial positions of the lesions, demonstrating a high degree of focused attention.

While occasional attention drift or anatomical misalignment in prompt generation may occur, the reliance on coarse-grained anatomical and spatial priors provides only a global constraint. This design enables the model to tolerate local deviations and still maintain robustness, ensuring that the diagnostic reasoning process remains coherent and reliable.

As shown in the attention distribution maps in Fig. 2, the attention guided by the prompts consistently focuses on the lesion-context regions. These observations confirm that PG3D-ViT, under the guidance of prompts, can effectively focus on the key lesion areas identified by the prompt mechanism. This aligns with our original design motivation: the prompt mechanism enables the model to explicitly concentrate on clinically relevant regions, implementing a top-down attention strategy that simulates the clinical reasoning process.

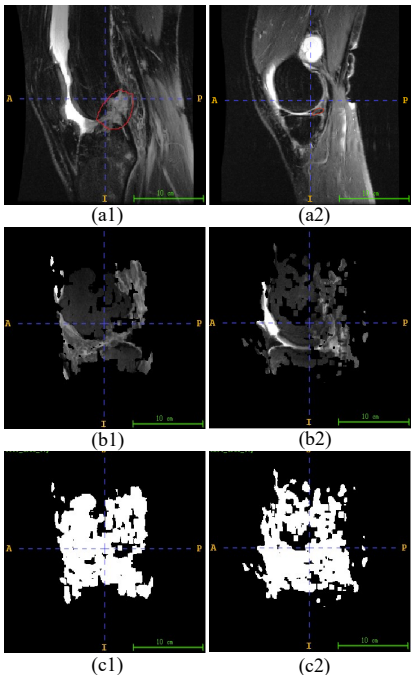

Fig. 4. **Lesion Context Prompt:** (a1) and (a2) are the original input slices containing lesion regions of the ACL and meniscal tear, with the red circles indicating the lesion areas; (b1) and (b2) show the anatomical prompts extracted from similar positive samples; (c1) and (c2) display the spatial location prompts, which are primarily activated around the anterior cruciate ligament and meniscus regions, respectively.

D. Ablation Study

1) ***Without MAE Pretraining (w/o MAE Pretraining):*** This experiment aims to evaluate the impact of MAE models trained on medical images in enhancing feature representation. Specifically, we replace the ViT encoder weights with the

original MAE weights pretrained on natural images, while keeping all other components unchanged, in order to observe how the model performs in medical image classification tasks under direct transfer settings.

*2) **Without Lesion-Context Prompt Module (w/o Prompting Module)):*** To assess the role of the prompt mechanism in guiding the model's attention to critical regions, we remove the lesion-context prompting module and directly feed raw image features into the MAE encoder, followed by a linear classification head. No anatomical or spatial prompt information derived from the consistency differences between normal and abnormal samples is introduced.

*3) **Organ Prompting Module**:* To further investigate how the granularity of prompts affects model guidance, we replace the fine-grained lesion-context prompts with a coarse-grained organ prompting module, in order to explore the influence of different levels of prompt granularity on attention focusing.

*4) **Without Prompt-Guided Aggregation Module (w/o Prompt-Guided Aggregation)**:* This experiment aims to assess the contribution of the cross-attention mechanism in fusing prompt information. We replace the cross-attention operation with a simple feature concatenation strategy, where the prompt vector is no longer used as the Query. Instead, the prompt and image features are concatenated and passed through a linear transformation for fusion, allowing us to evaluate performance under conditions without explicit attention guidance.

All ablation variants were evaluated under the same training configurations as the complete PG3D-ViT model, and the results are summarized in Table VI. Among them, removing the MAE module pretrained on medical images results in the most substantial performance drop, suggesting that the semantic gap between natural and medical images limits the transferability of original MAE weights. Likewise, eliminating the lesion-context prompting module also causes a noticeable decline, indicating that explicit structural prompts play a crucial role in guiding the model to key diagnostic regions.Replacing lesion-level prompts with organ-level prompts results in a moderate performance decline, further indicating that fine-grained prompts are more effective in guiding the model to focus on lesion regions, which typically occupy only a small portion of the organ. Finally, removing the prompt-guided aggregation module leads to only a slight performance drop, suggesting that the cross-attention mechanism offers additional benefits in fusing prompt information and modeling long-range spatial dependencies.

TABLE VI
ABLATION STUDY ON PROMPT AND FUSION MODULES

Model Configuration	Vessel MNIST3D	Nodule MNIST3D
Full PG3D-ViT	0.942	0.942
w/o MAE Pretraining	0.856	0.920
w/o Lesion-Context Prompt Module	0.918	0.922
Organ Prompt Module	0.926	0.930
w/o Prompt-Guided Aggregation Module	0.932	0.935

In summary, the three core modules in PG3D-ViT work synergistically to improve both the accuracy and generalizability of 3D medical image classification.

E. Generality and Limitation Study

Generality Analysis:The prompt mechanism introduced in PG3D-ViT demonstrates strong adaptability across a variety of medical image classification tasks. For tasks involving well-defined anatomical structures and relatively stable lesion distributions, the prompt serves as an effective structural prior, guiding the model to focus on diagnostically relevant regions and significantly improving classification performance. For example, although anterior cruciate ligament (ACL) tears and meniscal injuries present in various clinical forms, their anatomical locations are well-defined and associated with describable structural references. In most patients, such lesions exhibit consistent spatial distributions, making the anatomical context reliable. PG3D-ViT leverages prompt generation through alignment and contrast between positive and negative samples to extract clinically meaningful lesion-context cues, thereby enabling the model to focus efficiently on regions of the knee joint most prone to pathological changes, ultimately enhancing diagnostic accuracy and robustness.

Limitation Analysis:Despite its effectiveness, the prompt mechanism used in PG3D-ViT shows limitations in tasks where lesions are highly heterogeneous in appearance or lack consistent spatial patterns. In dermatological lesion classification tasks (e.g., DermatologyMNIST), lesions vary widely in morphology, boundaries, and location, and lack stable anatomical references. This makes it difficult for consistency-based prompts to provide meaningful guidance. Experimental results show that in such tasks, introducing prompt mechanisms may not only fail to improve performance but may also misguide the model's attention away from diagnostically relevant areas, leading to overall performance degradation. These observations indicate that in tasks lacking clear anatomical priors or stable lesion structures, the use of prompts should be approached with caution, as their generalization capacity may be fundamentally limited.

V. CONCLUSION

This paper proposes a novel Vision Transformer architecture for 3D medical image classification, named PG3D-ViT. The model is inspired by the diagnostic reasoning process of radiologists, which typically begins with global anatomical structures and gradually narrows focus to subtle local lesions in a top-down manner. By introducing contextual prompts that encode anatomical structures and spatial priors, PG3D-ViT implements a guided attention mechanism within the Transformer framework.

PG3D-ViT offers a new paradigm for understanding 3D medical images. Unlike conventional approaches that treat all voxels equally, the proposed model emulates clinical diagnostic logic by leveraging prompt information to effectively guide attention toward critical regions, thereby enhancing both discriminative power and generalization capability. We believe

this method holds strong potential for broad application in medical imaging analysis.

Future work may explore more efficient and generalizable prompt generation strategies, such as incorporating learnable anatomical atlases or integrating large-scale prompt models. In addition, incorporating human-in-the-loop evaluation is of great significance, particularly by allowing human experts to interactively correct the prior attention. In this way, bidirectional collaboration can be achieved in prompt generation and attention guidance. This would help foster genuine human–machine interactive collaboration, further enhancing the clinical relevance and translational potential of the proposed approach.

REFERENCES

[1] N. Bien, P. Rajpurkar, R. L. Ball, et al., "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet," PLOS Medicine, vol. 15, no. 11, p. e1002699, 2018.

[2] S. Katabathula, Q. Wang, R. Xu, "Predict Alzheimer's disease using hippocampus MRI data: a lightweight 3D deep convolutional network model with visual and global shape representations," *Scientific Reports*, vol. 12, art. 17106, 2022.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, pp. 1097–1105, 2012.

[4] A. Dosovitskiy, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. ICLR*, 2021.

[5] H. Wang, et al., "Feature Pyramid Vision Transformer for Medical Image Classification," *Medical Image Analysis*, 2022.

[6] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, A. Ayatollahi, "MedViT: A Robust Vision Transformer for Generalized Medical Image Classification," *Computers in Biology and Medicine*, vol. 157, 106791, 2023.

[7] M. Caron, H. Touvron, I. Misra, et al., "Emerging properties in self-supervised Vision Transformers," *Nature*, vol. 604, no. 7904, pp. 695–700, 2022.

[8] Y. Xie, J. Zhang, Y. Xia, Q. Wu, "UniMiSS: Universal Medical Self-Supervised Learning via Breaking Dimensionality Barrier," arXiv:2112.09356, 2021.

[9] A. Kirillov, et al., "Segment Anything," arXiv:2304.02643, 2023.

[10] A. Radford, et al., "Learning Transferable Visual Models from Natural Language Supervision," in *Proc. ICML*, 2021.

[11] Z. Ji, Z. Chen, X. Ma, "Grouped multi-scale vision transformer for medical image segmentation," *Scientific Reports*, vol. 15, 11122, 2025.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[13] W. Wang, et al., "GC-WIR: 3D global coordinate attention wide inverted ResNet network for pulmonary nodules classification," *BMC Pulmonary Medicine*, vol. 24, Article 3272, 2024.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proc. ICML*, pp. 1597–1607, 2020.

[15] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," in *Proc. CVPR*, pp. 9729–9738, 2020.

[16] J.-B. Grill, F. Strub, F. Altché, et al., "Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning," in *Proc. NeurIPS*, vol. 33, pp. 21271–21284, 2020.

[17] M. Caron, H. Touvron, I. Misra, et al., "Emerging Properties in Self-Supervised Vision Transformers," in *Proc. ICCV*, pp. 9650–9660, 2021.

[18] K. He, X. Chen, S. Xie, et al., "Masked Autoencoders Are Scalable Vision Learners," in *Proc. CVPR*, pp. 16000–16009, 2022.

[19] J. Yang, R. Shi, B. Ni, et al., "MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis," in *Proc. IEEE ISBI*, 2021.

[20] J. Yang, R. Shi, D. Wei, et al., "MedMNIST v2: A Large-Scale Lightweight Benchmark for 2D and 3D Biomedical Image Classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[21] J. Wu, Z. Wang, M. Hong, et al., "Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation," *Medical Image Analysis*, vol. 102, p. 102882, 2025.

[22] X. Liu, G. Shi, R. Wang, et al., "Segment Any Tissue: One-shot Reference Guided Training-free Automatic Point Prompting for Medical Image Segmentation," *Medical Image Analysis*, vol. 102, p. 103550, 2025.

[23] Y. Huang, P. Cheng, R. Tam, and X. Tang, "Fine-grained Prompt Tuning: A Parameter and Memory Efficient Transfer Learning Method for High-resolution Medical Image Classification," arXiv preprint arXiv:2403.07576, 2024. [Online]. Available: https://arxiv.org/abs/2403.07576

[24] N. M. Khan, N. Abraham, M. Hon, *et al.*, "Transfer Learning with Intelligent Training Data Selection for Prediction of Alzheimer's Disease," *IEEE Access*, vol. 7, pp. 72726–72735, 2019.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. 3rd Int. Conf. Learning Representations (ICLR), 2015.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.

[27] S. Sharma, M. Umer, A. Bhagat, J. Bala, P. Rattan, and A. Rahmani, "A ResNet50-Based Approach to Detect Multiple Types of Knee Tears Using MRIs," *Mathematical Problems in Engineering*, vol. 2022, Article ID 7411081, 2022.

[28] WANG K, NIU X, DOU Y, *et al*. A siamese network with adaptive gated feature fusion for individual knee OA features grades prediction [J/OL]. *Scientific Reports*, 2021.

[29] Z. Lai, J. Wu, S. Chen, Y. Zhou, and N. Hovakimyan, "Residual-based Language Models are Free Boosters for Biomedical Imaging," arXiv preprint arXiv:2403.17343, 2024.

[30] P. Xie, K. Zuo, J. Liu, M. Chen, S. Zhao, W. Kang, and F. Li, *Interpretable Diagnosis for Whole-Slide Melanoma Histology Images Using Convolutional Neural Network*, Journal of Healthcare Engineering, vol. 2021, pp. 1–7, Nov. 2021.

[31] J. Yang, et al., "Reinventing 2D Convolutions for 3D Images," IEEE J. Biomed. Health Informatics, pp. 1–1, 2021.

[32] M. Feurer, A. Klein, K. Eggensperger, et al., "Auto-sklearn: Efficient and Robust Automated Machine Learning," in Automated Machine Learning, Springer, 2019, pp. 113–134.

[33] H. Jiang, Q. Song, X. Hu, "Auto-Keras: An Efficient Neural Architecture Search System," arXiv:1806.10282, 2018.

[34] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. CVPR, pp. 4700–4708, 2017.

[35] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, pp. 6105–6114, 2019.

[36] X. Yang, R. Shi, T. He, Y. Wang, L. Wang, and L. Wen, "Rethinking Foundation Models for Medical Image Classification through a Benchmark Study on MedMNIST," arXiv preprint arXiv:2312.06600, 2023.

[37] Y. Fang, J. Wang, Q. Zhang, Y. Wang, W. Chen, Z. Liu, Y. Wei, and H. Li, "EVA: Exploring the limits of masked visual representation learning at scale," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 19358–19369, 2023.