Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A modified vision transformer architecture with scratch learning capabilities for plutonic rock classification

Sittiporn Kongsukho
Chulalongkorn University
Warunee Maneerat
Department of Mineral Resources
Narihiro Owada
Akita University
Tsuyoshi Adachi
Akita University
Peerapon Vateekul
Chulalongkorn University
Chakkaphan Sutthirat
chakkaphan.s@chula.ac.th
Chulalongkorn University

Research Article

Keywords: Deep learning, Vision Transformers, Self-attention mechanism, Plutonic rock classification, Photomicrography

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11255107/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A modified vision transformer architecture with scratch learning capabilities for plutonic rock classification

Sittiporn Kongsukho[a], Warunee Maneerat[b], Narihiro Owada[c], Tsuyoshi Adachi[c], Peerapon Vateekul[d**], Chakkaphan Sutthirat[a*]

[a] Department of Geology, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand

[b] Department of Mineral Resources, Rama VI Road, Rhayathai, Ratchathewi, Bangkok 10400, Thailand

[c] Faculty of International Resource Sciences, Akita University, 1-1 Tegata Gakuen-machi, Akita 010-8502, Japan

[d] Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand

[*]Corresponding Author: chakkaphan.s@chula.ac.th; c.sutthirat@gmail.com

**Co-Corresponding Author: peerapon.v@chula.ac.th

Abstract

The classification of plutonic rocks is essential for geological research, mineral exploration, and petrological analysis. Traditional methods rely on expert interpretation of thin-section images, making the process time-consuming and necessitates expertise. Vision Transformers (ViTs) provide promising tools to address these challenges by enabling automated classification. However, previous studies have focused primarily

on model accuracy without considering generalization and often lack a large plutonic rock dataset curated by petrologists. Thus, this study introduces the Petrographic Vision Transformer (PViT), a family of ViT models designed to automatically classify plutonic rock types in Thailand with high accuracy and generalization. The proposed models incorporate a new Gaussian Error Linear Unit (new GELU) activation in the position-wise feed-forward network (FFN) and an optimized multi-head attention mechanism with merged Query (Q), Key (K), and Value (V) projections. Training using a dataset of 2,886 thin-section images representing 15 plutonic rock types, PViT-Base achieved 98.34% accuracy, exceeding the standard ViT-Base by 0.62 percentage points when trained from scratch. Moreover, PViT achieved slightly higher accuracy than the Dual Attention Vision Transformer (DaViT) and EfficientNetV2, both state-of-the-art (SOTA) models originally benchmarked on ImageNet-1K. Grad-CAM visualizations generated by our model highlighted influential regions for classification, corresponding to key petrographic features, which closely resemble the interpretive approach used by petrologists. These findings demonstrate the potential of PViT to advance automated classification and offer strong potential for future geoscience applications.

Keywords: Deep learning, Vision Transformers, Self-attention mechanism, Plutonic rock classification, Photomicrography

1. Introduction

Rock classification is a fundamental task in geoscience, involving the observation of thin-sections to identify rock types based on framework minerals, textures, and internal structures (Streckeisen, 1976). These features, expressed through

shape, color, and microstructural characteristics, provide the basis for classification. Traditionally, rock samples are sliced into 30-µm-thick thin-sections mounted on glass slides and examined under polarized light microscopes. Thin-sections are then photographed as three-channel digital images (RGB photomicrographs), allowing geoscientists to study mineral composition and internal structures (e.g., Kongsukho and Kanjanapayont, 2025, Kongsukkho and Kanjanapayont, 2022). Photomicrograph analysis has consequently become a standard workflow in geological mapping, mineral exploration, and petroleum systems studies (e.g., Aligholi et al., 2019, Eldosouky et al., 2023, Hussain et al., 2023). Accurate interpretation of these images is essential for successful mineral exploration and geological assessments (Patel et al., 2017, Perez et al., 2011). However, the internal structures of rocks are heterogeneous and exhibit directional variability (Lepistö et al., 2005, Zhu et al., 2019).

Recent advances in artificial intelligence (AI) and machine learning (ML) have transformed computer vision into a field dominated by deep learning (DL) models (Krizhevsky et al., 2012, Ren et al., 2015, Shin et al., 2016). In geoscience, DL models have been increasingly adopted to automate thin-section image classification, particularly convolutional neural networks (CNNs) (e.g., Guojian and Peisong, 2021, Pires de Lima and Duarte, 2021, Polat et al., 2021, Qiqi et al., 2024, Seo et al., 2022, Su et al., 2020, Zheng et al., 2024b). These works demonstrate that CNNs can reduce network complexity and the number of training parameters, making them easier to train and optimize. As a result, CNN-based models provide an effective and efficient framework for automated classification of thin-section images. However, they tend to extract features from entire images without adequately attending to local detail, which can lead to misclassification (Seo et al., 2022, Zhang and Wang, 2023).

Recently, Vision Transformers (ViTs) have proven to be highly successful in computer vision tasks. Transformer architectures can achieve state-of-the-art performance in thin-section image classification (e.g., Wang et al., 2024, Zhang and Wang, 2023). The self-attention mechanism enables models to capture both global and local contextual information by allowing each patch to attend to all others (Vaswani et al., 2017). This mechanism is effective for modeling long-range dependencies beyond the local receptive fields of convolutional neural networks (Dosovitskiy et al., 2020). Consequently, self-attention within Transformer blocks facilitates the extraction of discriminative petrographic features and has been shown to outperform CNN-based models in thin-section image classification (e.g., Cai et al., 2024). However, ViTs typically rely on large-scale datasets for pretraining and require huge computational resources (Yang et al., 2021). Most previous studies have employed transfer learning strategies, which often prioritize high accuracy over generalization, both of which are essential for reliable applications in real-world geoscience (Lei et al., 2019, Yang et al., 2020). Additionally, the lack of training data from expert petrologists remains a major challenge. Large datasets curated by petrologists can significantly enhance model tuning and performance (Zheng et al., 2024a).

To address these challenges, this study proposes the Petrographic Vision Transformer (PViT), a family of Vision Transformer (ViT) models enhanced with two modifications: (1) a new Gaussian Error Linear Unit (GELU) activation in the position-wise feed-forward network (FFN) and (2) an optimized multi-head attention mechanism. The models are trained on thin-section images of plutonic rocks captured under both plane-polarized light (PPL) and cross-polarized light (XPL). The dataset comprises a total of 2,886 images representing 15 plutonic rock types collected from three granite belts across Thailand. The key features for classifying these rocks include

grain properties, internal structures, and basic categories, all of which are optical properties curated by expert petrologists. Model performance is evaluated using numerous metrics, and the Grad-CAM visualization highlights influential regions for classification by emphasizing distinguishing framework grains, textures, and microstructures. The results demonstrate that PViT not only achieves high accuracy but also extracts interpretable features related to real-world geoscience applications.

2. Related works

2.1 A short review of image processing using Vision Transformers

Vision Transformers (ViTs) have been adapted for image analysis by treating images as sequences of patch tokens rather than relying on local convolutions, as in traditional convolutional neural networks (CNNs). Each image is divided into fixed-size patches, which are linearly projected into embeddings and combined with positional encodings before being processed by Transformer encoder layers (Dosovitskiy et al., 2020, Vaswani et al., 2017). These encoder layers consist of multi-head attention (MHA) and multi-layer perceptron (MLP), integrated with residual connections and layer normalization. The MHA mechanism enables each patch token to attend globally, capturing long-range dependencies within a single layer (Vaswani et al., 2021, Zhang et al., 2022), unlike CNNs, which aggregate information locally through convolutional kernels (O'shea and Nash, 2015). After the MHA and MLP layers, a classification head is applied for downstream tasks.

Tokenization and positional encoding are crucial for preserving spatial structures. The self-attention mechanism assigns varying levels of importance to

different regions of an image. Within each self-attention block, patch tokens are linearly transformed into query, key, and value vectors. The attention mechanism computes similarity scores through query-key dot products and aggregates values accordingly. Multi-head attention performs this operation across multiple attention heads in parallel, allowing the model to capture complementary spatial relationships. This architecture enables ViTs to encode both global and local contextual information, frequently outperforming CNNs in image classification, object detection, and segmentation, particularly when pretrained on large-scale datasets with substantial computation resources (Yang et al., 2021).

Since their introduction, ViTs have been widely adopted for image classification (e.g., Chen et al., 2021). Subsequent variants such as Data-efficient image Transformers (DeiT) (Touvron et al., 2021), Token-to-Token ViT (T2T) (Yuan et al., 2021), Transformer in Transformer (TNT) (Han et al., 2021), and Swin Transformer (Liu et al., 2021) have improved efficiency by reducing the number of parameters, lowering computation requirements, increasing classification accuracy, and better handling high-resolution images, thereby addressing several limitations of the standard ViT.

2.2 Applications of Vision Transformers for Petrographic Image Classification

Although recent studies have demonstrated the strong potential of Vision Transformers (ViTs) in various image classification tasks, their application to thin-section image classification remains limited. Zhang and Wang (2023) developed a hybrid Transformer model that combines Region-to-Region and Token-to-Token attention, achieving fine-grained mineral classification in sandstone and limestone with a top-1 accuracy of 88.04%. Cai et al. (2024) applied the Swin Transformer with

transfer learning to classify mineral phases (e.g., arsenopyrite, pyrite, and gold), achieving 92% accuracy and outperforming CNN-based models such as MobileNetV2 and ResNet50. Wang et al. (2024) introduced Rock-ViT for sandstone classification and achieved 91.7% accuracy by incorporating supervised contrastive loss. More recently, Zheng et al. (2024a) proposed Trans-SedNet, a dual-modal ViT trained on XPL and PPL images, built upon the SegFormer backbone with enhanced attention mechanisms. This model outperformed CNN-based architectures such as U-Net and DeepLabV3+ in semantic segmentation by accurately identifying individual mineral grains. These studies employed simplified classification to distinguish mineral grains in thin-section images. Although Vision Transformer models have achieved state-of-the-art performance in petrographic image classification, their application to plutonic rock classification remains unexplored, highlighting a clear research gap.

3. Data

3.1 Plutonic rock sampling and thin-section preparation

Plutonic rocks for this study were collected from three granite belts across Thailand (Fig. 1), categorized by geochemistry, tectonic settings, and geochronology (e.g., Charusiri et al., 1993, Cobbing, 2011). These belts correlate with granites in neighboring countries such as Malaysia, Myanmar, Indonesia, and Lao PDR (Mitchell, 1977). Sampling sites included: (i) the Western Granite Belts in Kanchanaburi Province (Fig. 1A); (ii) the Eastern and Central Belts in Chachoengsao, Chon Buri, Chanthaburi, and Rayong Provinces (Fig. 1B); and (iii) the Central and Western Belts in Phetchaburi and Prachuap Khiri Khan Provinces (Fig. 1C).

The Western Granite Belt, extending along the Thai-Myanmar border, consists of granodiorite to syenogranite with K-feldspar phenocrysts, quartz, and biotite. Most samples are S-type peraluminous granites; I-type equigranular varieties also occur. Ages range from the Cretaceous to the Paleocene (Beckinsale et al., 1979, Charusiri et al., 1993). The Eastern Granite Belt consists of medium- to coarse-grained syenogranite and gabbro with accessory hornblende and minor mafic intrusions. Although S-type and A-type granites are present, most rocks are I-type granites of the magnetite series, associated with Cu-Au and Mo mineralization (Khin et al., 2014). Ages range from the Late Carboniferous to the Late Triassic (Gardiner et al., 2018). The Central Granite Belt comprises large batholiths of syenogranite to monzogranite, often displaying megacrystic K-feldspar and perthitic textures. Some rocks exhibit gneissic fabric due to metamorphism (Kongsukkho and Kanjanapayont, 2022), while younger alkaline plutons intrude the main bodies. These granites are largely S-types (Ilmenite series), dated to the Late Triassic-Early Jurassic (Nantasin et al., 2012).

All samples were prepared as standard thin-sections (30 μm thickness) and examined under petrographic microscopes. Rock classification followed the International Union of Geological Sciences (IUGS) schemes: the QAPF diagram for granitoids (Fig. 2A), the gabbroic rocks diagram (Fig. 2B), and the ultramafic rock diagram (Fig. 2C). These diagrams were applied based on modal mineral proportions following the IUGS classification system (Streckeisen, 1976).

3.2 Image collection

Thin-section images were obtained using a polarized light microscope (Nikon ECLIPSE LV100 POL) at the Department of Geology, Chulalongkorn University.

Magnifications of 1×, 5×, and 10× were employed: lower magnifications (1×, 5×) provided broader views, whereas 10× captured detailed mineral frameworks. Each sample was photographed under both plane-polarized light (PPL) and cross-polarized light (XPL). The original thin-section images had a resolution of 4908 × 3264 pixels and were acquired in three color channels (RGB).

Polarized light microscopy enables mineral identification by exploiting optical properties. Under PPL, mineral grains can be distinguished based on petrographic features such as crystal shape, cleavage, color, and pleochroism. Pleochroism is an optical property in which a mineral displays different colors depending on the orientation of incident light. In contrast, under XPL, additional features, including interference colors, twinning, zoning, and cleavage, become visible. Interference colors arise when light splits into two rays as it passes through a mineral between crossed polars; minerals are generally light-colored and exhibit vivid colors that vary with birefringence. Twinning, often clearly observable under XPL, is particularly useful for distinguishing feldspars such as alkali feldspar and potassium feldspar. Simple twins consist of two crystal segments that extinguish at different angles during stage rotation. Contact twins share a smooth twin plane, whereas penetration twins intersect at irregular junctions. Polysynthetic twinning, characterized by numerous twin segments joined along parallel twin planes, is commonly found in plagioclase. Zoning, frequently observed in plagioclase and potassium feldspar, results from compositional changes during crystal growth and appears as color bands under XPL. These optical properties provide essential criteria for mineral identification and are critical for the classification of plutonic rocks in thin-section images.

3.3 Image dataset and category distribution – Petrological features

We collected 1,443 original thin-section images from 15 plutonic rock types, classified using the IUGS system (Fig. 2). Each rock type was imaged under both plane-polarized (PPL) and cross-polarized light (XPL), with details illustrated in Figure 3.

Anorthosite is dominated by calcic plagioclase feldspar (3-4 mm grains) with serrated, irregular boundaries, often partially altered to sericite, a fine-grained white mica (Fig. 3A). Diorite, a fine- to medium-grained mafic rock, contains andesine, oligoclase, pyroxene, and biotite, together with accessory hornblende, zircon, apatite, and Fe-Ti oxides; it is characterized by albite twinning in plagioclase and pleochroism in biotite (Fig. 3B). Dunite, dominated by olivine, displays high birefringence, distinct fracturing, chemical zoning, and alteration products, with no cleavage (Fig. 3C). Foidolite, a rare feldspathoid-rich rock, is predominantly composed of leucite and zoned clinopyroxene with iron variations (Fig. 3D). Gabbro primarily consists of labradorite and bytownite, with amphibole, clinopyroxene, and olivine; plagioclase shows pericline twinning, while olivine is frequently altered to hematite along cracks and to chlorite at grain margins (Fig. 3E). Granite, a felsic rock, contains quartz, plagioclase, alkali feldspar, biotite, and muscovite; feldspars commonly display exsolution and intergrowth textures, such as perthite and myrmekite (Fig. 3F). Granodiorite consists primarily of plagioclase and alkali feldspar, resembling granite but with more plagioclase and less alkali feldspar. It also contains minor minerals such as biotite, amphibole, and possibly apatite, with biotite grains often showing pleochroic halos around zircon (Fig. 3G). Greisen, an altered granitic or pegmatitic rock, is typically characterized by coarse-grained quartz and fine-grained topaz, fluorite, and tourmaline (Fig. 3H). Harzburgite, an ultramafic rock, is composed mainly of olivine, enstatite, and bronzite; olivine grains exhibit irregular cracks forming a mesh texture

and high birefringence, while orthopyroxene shows schiller structures and first-order yellow to low birefringence (Fig. 3I). Hornblendite is primarily composed of hornblende, appearing as slender prismatic or bladed crystals with a diamond-shaped cross section; amphibole cleavage is visible at 56° and 124°, with common concentric or patchy color zoning (Fig. 3J). Lherzolite is dominated by olivine and orthopyroxene, with minor clinopyroxene, spinel, and garnet, also contains plagioclase (Fig. 3K). Monzonite is an intermediate rock between syenite and diorite, composed of altered plagioclase and orthoclase with numerous inclusions, together with biotite and amphibole, both of which are partly altered to chlorite (Fig. 3L). Norite resembles gabbro but contains more orthopyroxene, showing augite exsolution lamellae that appear either tabular or pinched and swollen, forming rows of blebs (Fig. 3M). Orthopyroxenite is composed entirely of orthopyroxene, characterized by low birefringence, parallel extinction, pale pink to green pleochroism, two cleavages intersecting at ~88°, and fine exsolution lamellae (Fig. 3N). Syenite, a felsic rock, is primarily composed of orthoclase, aegirine, and augite, with minor plagioclase and accessory minerals such as titanite, actinolite, allanite-(Ce), ilmenite, and zircon. Orthoclase often shows perthitic texture with exsolution blebs as well as albite lamellae, while Carlsbad twinning, dividing crystals into two irregular segments, is common (Fig. 3O).

In total, 2,886 images were produced (PPL and XPL). The dataset was divided into training (70%), validation (15%), and test (15%) sets (Fig. 4). All original images were split into overlapping patches of 512 × 512 pixels to effectively model local structural information and enhance feature richness (Yuan et al., 2021). Additionally, pixel values within each patch were normalized using min-max scaling between 0 and

1 to ensure consistent feature magnitudes and stabilize the learning process (Patro and Sahu, 2015).

4 Methods

4.1 Model implementation

4.1.1 Overview of the Vision Transformer backbone

In this study, we modified the standard Vision Transformer (ViT) backbone from scratch, based on the model developed by Dosovitskiy et al. (2020), to address the task of plutonic rock classification. The standard ViT model consists of three main components: (i) a patch embedding layer, (ii) stacked Transformer encoder blocks, and (iii) an MLP classification head (Fig. 5A). The patch embedding layer divides each thin-section image into non-overlapping patches of size [*P, P*]. Each patch is flattened into a vector of dimension [*P* x *P* x *C*] and transformed into a one-dimensional sequence of embeddings. A learnable classification token [CLS] is prepended to the sequence to capture global semantics, and the resulting sequence is then fed into the encoder blocks (Fig. 5B).

In the encoder, input sequences are projected into Query (Q), Key (K), and Value (V) vectors. The attention scores are calculated using scaled dot-product attention as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where d_k is the dimensionality of the key vectors (Vaswani et al., 2017). The SoftMax function is applied to these scores to obtain weights for the values. Rather than performing a single attention function with the same dimensionality for keys, values, and queries, the standard ViT employs multi-head attention (MHA). In MHA, the Q, K, and V matrices are each linearly projected h times using different learned projection matrices, resulting in h parallel attention computations. Each head produces an output of dimension d_v, which is subsequently concatenated and projected back to the original hidden dimension. After the multi-head attention step, a position-wise feed-forward network (FFN) is applied independently to each token. The FFN consists of two linear transformations with a Gaussian Error Linear Unit (GELU) activation function applied between them. Each encoder block is wrapped with residual connections and layer normalization (Fig. 5C).

Finally, the MLP classification head processes the [CLS] token output and maps it to the predefined class space through a fully connected layer, producing a probability distribution over the target classes (Fig. 5D).

4.1.2 Petrographic Vision Transformer (PViT)

After implementing the standard Vision Transformer (ViT) from scratch, we explored an efficient backbone to increase accuracy and generalization in real-world geosciences, thereby facilitating plutonic rock classification. Each encoder layer contains a position-wise feed-forward network (FFN) and multi-head attention (MHA) (Dosovitskiy et al., 2020), both of which play crucial roles in the Transformer neural network by learning abstract features through the encoder blocks and enabling the model to capture a variety of spatial relationships in thin-section images. However, the

FFN becomes saturated for very high or very low input values due to the GELU activation function, leading to the vanishing-gradient problem (e.g., Lee, 2023). Moreover, self-attention in MHA typically models dependencies for a single position per layer, requiring substantial computation resources and long training times. To overcome these limitations associated with the inefficient encoder backbone of standard ViT, we propose two modifications: (i) a new Gaussian Error Linear Unit (new GELU) within the FFN, and (ii) an optimized multi-head attention mechanism. These modifications result in the development of a novel Vision Transformer architecture (Fig. 6A), hereafter referred to as the Petrographic Vision Transformer (PViT).

4.1.2.1 New Gaussian Error Linear Unit (GELU) Activation

Activation functions (AFs) are nonlinear layers in neural networks. Among them, the Gaussian Error Linear Unit (GELU) (Hendrycks and Gimpel, 2016) has gained popularity due to its higher accuracy, training stability, and smoother gradient behavior compared with classical activation functions such as ReLU (Nair and Hinton, 2010) and ELU (Clevert et al., 2015). Mathematically, the GELU activation function is based on the Gaussian cumulative distribution function (CDF). The GELU activation function is formally defined using the Gaussian CDF as:

$$GELU(x) = xP(X \leq x) = x\Phi(x) = x \cdot \frac{1}{2}\left[1 + \mathrm{erf}(x/\sqrt{2})\right] \quad (2)$$

where Φ(x) represents the probability that a standard normal distribution takes on a value less than or equal to *x*, and erf is the Gaussian error function.

A widely used approximation of GELU, proposed by Hendrycks and Gimpel (2016), is given as:

$$GELU(x) = 0.5x\left(1 + tanh\left[\sqrt{2/\pi}(x + 0.044715x^3)\right]\right) \quad (3)$$

This tanh-based approximation is widely used in image classification. However, thin-section images contain diverse and complex features such as shapes, textures, and microstructures, which differ significantly from those in general datasets such as ImageNet-21k. To improve the learning of these petrographic features, we implemented a new GELU variant in the FFN block of each Transformer encoder by adopting a sigmoid-based approximation.

The Sigmoid-based GELU is defined as:

$$\mathrm{GELU}(x) = x \cdot \sigma(1.702x) \quad (4)$$

where σ(x) is the sigmoid function. This approach exhibits a more linear behavior in the intermediate activation range, which reduces the extreme compression of mid-level features and yields smoother gradients compared with the tanh-based approximation. Such behavior is advantageous for capturing petrographic features and leads to improved classification accuracy.

4.1.2.2 Optimized multi-head attention

Self-attention (SA) is a core component of Transformer-based models, enabling the modeling of long-range dependencies by computing weighted averages over token representations (Bahdanau et al., 2014, Graves et al., 2014). Traditional self-attention typically models dependencies for a single position per layer, whereas multi-head attention (MHA) improves this by attending to multiple positions simultaneously (Vaswani et al., 2017), which has been central to the success of Transformer architectures (Chen et al., 2018). However, in standard MHA, attention heads within

the same layer operate independently, which limits interaction and computational efficiency. To address this limitation, we optimize the mechanism by fusing the Query (Q), Key (K), and Value (V) projections into a single joint linear transformation (Fig. 6B).

Given an input tensor $X \in \mathbb{R}^{B \times L \times D}$, where B is the batch size, L is the sequence length, and D is the hidden dimension, the Q, K, and V matrices are computed using a single linear projection (Fig. 7):

$$[Q, K, V] = XW_{qkv}, \qquad W_{qkv} \in \mathbb{R}^{Dx3D} \tag{5}$$

The resulting tensor is then split into three components and reshaped to support multi-head attention with H heads, where each head has dimension $N = D/H$:

$$[Q, K, V] \in \mathbb{R}^{B \times H \times L \times N} \tag{6}$$

Scaled dot-product attention is applied to compute the attention scores and outputs:

$$A = QK^T / \sqrt{N}, \qquad A_c = \mathrm{SoftMax}(A), \qquad Z = A_c V \tag{7}$$

Here, O denotes the final output of the multi-head attention layer after concatenating all attention heads and applying a linear projection back to the original hidden dimension:

$$O = ZW_o, \qquad W_o \in \mathbb{R}^{DxD} \tag{8}$$

This fused QKV design increases parallelism, enhances GPU memory locality, and reduces computational overhead while maintaining or improving performance.

4.2 Attention visualization

To interpret the internal workings of our models, we apply Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017, Zhou et al., 2016). This approach combines classification output gradients with self-attention scores to produce

class-specific heatmaps (Fig. 8). Gradients, backpropagated from the output to each encoder block, identify the contributions of encoded features, while self-attention feature maps, normalized by a sigmoid function, suppress sharp activations and retain meaningful patch correlations (Leem and Seo, 2024).

In the PViT architecture, the [CLS] serves as the learnable classification token; therefore, only the first row of each attention matrix, representing the connections between the [CLS] token and all patch tokens, is considered. Through residual connections, the [CLS]-based representation aggregates information from all encoder blocks.

Let $A^{k}_{h,1} \in \mathbb{R}^{N}$ denote the first row of the self-attention matrix from the h^{th} head in the k^{th} encoder block. This row represents the attention between the [CLS] token and all patch tokens. The feature map F^{k}_{h} is then defined as:

$$F_h^k = G(A_{h,1}^k) \tag{9}$$

when G (·) denotes the element-wise sigmoid function, applied to suppress sharp activations and enhance localization quality.

To generate the final CAMs, we compute the gradient $\alpha_h^{k,c}$, which represents the influence of $A^{k}_{h,1}$ on the classification output for class c. We combine this gradient with the corresponding feature map as follows:

$$L^c = \sum_{k=1}^{K} \sum_{h=1}^{H} F_h^k \odot \mathrm{ReLU}(\alpha_h^{k,c}) \tag{10}$$

where ⊙ denotes the Hadamard product, and the ReLU function ensures that only positive contributions are retained. The resulting vector $L^c \in \mathbb{R}^N$ is reshaped into a 2D spatial heatmap and upsampled to the original image resolution, enabling interpretable visualization of petrographic features relevant to classification.

5. Experimental setup

We evaluated the classification performance of Petrographic Vision Transformer (PViT) by comparing it with the standard Vision Transformer (Dosovitskiy et al., 2020) and state-of-the-art (SOTA) models, including DaViT (Ding et al., 2022), and EfficientNetV2 (Tan and Le, 2021), trained from scratch on our plutonic rock dataset. In addition, an ablation study was conducted to examine the effects of the new GELU activation and the optimized multi-head attention. To ensure a fair comparison, all experiments were trained and tested under the same scheme using the same dataset. The dataset comprises input images of size [512 x 512 x 3]. Hyperparameters were adopted from the standard ViT settings (Dosovitskiy et al., 2020), as summarized in Table 1. For instance, PViT-L/16 denotes the "Large" model variant with a 16 × 16 input patch size. Model performance was evaluated using accuracy, precision, recall, and F1-score. In addition, attention heatmaps were generated to interpret key petrographic features.

All experiments were conducted on a Windows workstation equipped with an Intel Xeon W5-2445 CPU and an NVIDIA RTX A4500 GPU (CUDA 12.8). All models were implemented in Python 3.12.7 using the PyTorch framework. AdamW, with a weight decay of 0.001, was employed to improve model generalization (Loshchilov and Hutter, 2017).

6. Results

6.1 Experimental results of plutonic rock classification

The performance of the Petrographic Vision Transformer (PViT), the standard Vision Transformer (ViT), and other state-of-the-art (SOTA) models across different configurations is summarized in Table 2. The results show that the PViT models achieved higher classification accuracy than these models when trained from scratch on the plutonic rock dataset. PViT-Base exhibited the highest classification accuracy of 98.34%, outperforming ViT-Base, DaViT-Base, and EfficientNetV2-M, which have similar model sizes (86M parameters). The classification accuracies for ViT-Base, DaViT-Base, and EfficientNetV2-M were 97.72%, 98.14%, and 98.14%, respectively. Comparing PViT-Large with ViT-Large, the PViT-Large model also achieved higher accuracy. PViT-Large (307M parameters) yielded an accuracy of 97.52%, while ViT-Large reached only 96.38%. Finally, we also adopted a larger model size of 638M parameters and obtained an accuracy of 96.07% for both the PViT-Huge and ViT-Huge models.

The loss and accuracy curves, together with the confusion matrices for test set, support these results, as shown in Figures 9. The PViT models converged rapidly in both training and validation curves at around 12 epochs. Similarly, ViT, DaViT, and EfficientNetV2 converged at approximately 20 training epochs. The confusion matrix for the test dataset demonstrates that the PViT model made very few mistakes; hornblendite was misclassified as dunite.

6.2 Visualization of petrographic features related to plutonic rock classification

Grad-CAM visualization generated from the Petrographic Vision Transformer (PViT) effectively highlight influence regions. For anorthosite, the models highlight

albite twinning in plagioclase (Fig. 10A-a, b) and thin, diffuse lamellae intersecting the twinning at acute angles (Fig. 10A2-c–f). They also identify fine alteration of sericite and associated fine-grained white mica within plagioclase. In diorite, attention is focused on exsolution lamellae in augite (Fig. 10B2-a, c), distinct cleavage and pleochroism in biotite (Fig. 10B2-b), and irregular boundaries of plagioclase (Fig. 10B2-d). In dunite, the models highlight distinctive fracturing (Fig. 10C2-a–c) and compositional zoning of olivine. In foidolite, the models focus on low-birefringence colors in leucite (Fig. 10D2-a), complex polysynthetic twinning (Fig. 10D2-b, c), two intersecting cleavages in clinopyroxene (Fig. 10D2-d), and compositional zoning in clinopyroxene (Fig. 10D2-e, f). For granodiorite, the heatmaps emphasize pleochroism in biotite (Fig. 10E2-a–c), albite twinning in plagioclase (Fig. 10E2-d), and quartz exhibiting undulose extinction (Fig. 10E2-e, f). For harzburgite, the key petrographic features include undulatory zoning with high birefringence in olivine (Fig. 10F2-a, b) and irregular cracking with lower birefringence in orthopyroxene (Fig. 10F2-c–e). The models accurately identify bronzite based on its schiller structure and two perpendicular cleavages (Fig. 10F2-f), successfully distinguishing hornblende from bronzite due to their differing cleavage systems. In hornblendite, attention is focused on slender prismatic and bladed amphibole crystals (Fig. 10G2-a) that exhibit diamond-shaped cross-sections intersecting at approximately 56^{o} and 124^{o} (Fig. 10G2-b). For lherzolite, the PViT models highlight high birefringence (Fig. 10H2-a), undulose extinction in olivine (Fig. 10H2-b–d), and exsolution lamellae in clinopyroxene (Fig. 10H2-e). In gabbro, our models emphasize exsolution lamellae in clinopyroxene (Fig. 10I2-a–c) and albite twinning in sodium-rich plagioclase (Fig. 10I2-d, e). They also identified pericline twinning in calcic plagioclase (Fig. 10I2-f, g). For granite, key features include grid twinning in microcline (Fig. 10J2-a, b), pleochroism in biotite (Fig. 10J2-c), and

493 undulose extinction in quartz (Fig. 10J2-d). Some detection of perthite composed of
494 albite in microcline and/or orthoclase is also noted (Fig. 10J2-e). In syenite, attention is
495 focused on myrmekite, consisting of irregular plagioclase and/or quartz grains within
496 orthoclase (Fig. 10K2-a, b), and perthite, which exhibits exsolution lamellae of albite
497 in orthoclase (Fig. 10K2-c, d). For monzonite, the models concentrate on large quartz
498 grains (Fig. 10L2-a), exsolution blebs and lamellae of albite in microcline (Fig. 10L2-
499 b), and euhedral crystals of hornblende (Fig. 10L2-c, d). In greisen, the models highlight
500 cloudy quartz associated with their inclusions (Fig. 10M2-a–d), and patchy brown and
501 blue tourmaline (Fig. 10M2-e, f). For norite, the models captured orthopyroxene
502 exhibiting characteristic undulose extinction (Fig. 10N2-a–c) and interstitial
503 plagioclase (Fig. 10N2-d, e). Finally, in orthopyroxenite, the models emphasize
504 prismatic cleavages intersecting at nearly 90^{o} and tabular exsolution lamellae of augite
505 (Fig. 10O2-a–c).
506
507 **6.3 Ablation Study**
508
509 The Petrographic Vision Transformer (PViT) is a modified model based on the
510 Vision Transformer (ViT) architecture, composed of two modifications: a new GELU
511 and an optimized multi-head attention mechanism, as described in Section 4.1.2. To
512 optimize the model's performance for plutonic rock classification, classical activation
513 functions and attention mechanisms were incorporated into our model architecture for
514 comparison. All models were implemented under the same training scheme using the
515 plutonic rock dataset.
516 From the intensive experiments on different activation functions shown in Table
517 3, it is apparent that both the new GELU and the optimized multi-head attention

mechanism outperform the standard ViT, which consists of the standard GELU and standard multi-head attention mechanism, yielding a 0.62% improvement over the baseline model. When comparing the experimental results of the standard multi-head attention with different activation functions, the accuracy decreases by 1.03% with SiLU, 0.41% with Mish, 2.27% with Laplace. Similarly, when SiLU, Mish, and Laplace adopt the optimized multi-head attention, the accuracy decreases by 1.45%, 1.24%, and 3.72%, respectively. However, the standard GELU combined with the optimized multi-head attention mechanism improves model performance by 0.21% (Table 3; Fig. 11).

Furthermore, intensive experiments on different attention mechanisms were trained under the same training scheme to further identify the effects of the optimized multi-head attention mechanism, as summarized in Table 4 and Figure 11. The optimized multi-head attention mechanisms combined with the new GELU achieve the highest accuracy, surpassing the standard multi-head attention. The results show that classical attention mechanisms, including Multi-Query Attention, Talking-Head Attention, and Flash Attention, are not sufficient to match the performance of GELU in both Tanh-based and Sigmoid-based approximations. Multi-Query Attention and Flash Attention can effectively improve throughput, which reduces computational resources and improves training speed, but their models exhibit lower accuracy compared with the standard ViT (Table 4; Fig. 11). Talking-Head Attention performs poorly in terms of accuracy (1.24% decrease), throughput (2.8 image/second decrease), and GPU memory usage (6.7 GB increase), making it inefficient for our dataset.

7. Discussion

7.1 Performance of the Petrographic Vision Transformer (PViT) versus state-of-the-art (SOTA) models for plutonic rock classification

Based on the numerical metrics and Grad-CAM visualizations, the Petrographic Vision Transformer (PViT) demonstrates excellent classification performance on the plutonic rock dataset from Thailand. PViT consistently achieves higher classification accuracy than the standard ViT (Dosovitskiy et al., 2020), DaViT (Ding et al., 2022), and EfficientNetV2 (Tan and Le, 2021) when trained from scratch. Specifically, our model provides improvements of 0.62% over the standard ViT and 0.2% over both DaViT and EfficientNetV2. In addition to its accuracy, Grad-CAM visualizations from PViT reveal influence regions that align with petrologists' knowledge (Fig. 10). The model focuses on petrographic features, such as cleavage, twinning, zoning, and extinction to classify the plutonic rocks. This interpretability is crucial, as it demonstrates that PViT does not rely on random feature extractions but instead learns petrographic features consistent with established geological knowledge. Integrating these interpretable regions and the complementary spatial relationships enabled by self-attention probably provides a strong foundation for classifying thin-section images under both plane-polarized (PPL) and cross-polarized light (XPL), overcoming limitations reported in previous studies (e.g., Su et al., 2020).

Although PViT outperforms these SOTA models when all are trained from scratch on our dataset (Fig. 12A; Table 2), its performance remains slightly lower than that of SOTA models, which were pretrained on large datasets such as ImageNet-1K and trained with significantly higher computational resources (Fig. 12B; Table 5). Among pretrained models, PViT surpasses only the fine-tuned CoCa-Base (Yu et al., 2022). Nevertheless, PViT exhibits strong potential for thin-section image classification

because it requires considerably fewer computational resources and fewer training samples than the pretrained models. The performance gap between PViT and the strongest pretrained model is relatively small (less than a 2.0% accuracy gap for base-scale models and less than 4.0% for huge-scale models), highlighting the model's ability to learn petrographic features despite the limited dataset sizes, and the gap might be closed with additional fine-tuning. Lastly, experienced petrologists can easily classify plutonic rock types based on distinctive grains, textures, and microstructures. PViT offers an alternative way that reduces the time required for petrographic analysis. By embedding petrographic knowledge into a single model, PViT has strong potential to support and accelerate plutonic rock classification in workflows, which further validates the superiority of our proposed PViT.

7.2 The impact of a new activation function and optimized multi-head attention: A case study on a plutonic rock dataset

This study presents an image classification using the Petrographic Vision Transformer (PViT) to automatically classify plutonic rock types in Thailand with high accuracy and strong generalization. The ablation study shows that the synergistic potential functions between the new GELU activation function in the position-wise feed-forward network (FFN) and the optimized multi-head attention mechanism improve classification accuracy while reducing redundancy by shortening training time and lowering computational requirements compared with standard Vision Transformer architectures (Dosovitskiy et al., 2020). This evidence indicates that the combination of the enhanced activation function in the FFN and the optimized multi-head attention effectively models structural information in the images and enriches feature

representation, ultimately enabling more accurate classification for plutonic rock types. The optimized multi-head attention mechanism likely has a greater impact on performance than the new activation function because self-attention is the core operation of the Transformer encoder (Vaswani et al., 2017). The modified multi-head attention with optimization makes it possible for the model to learn rich associations and local information from different sub-representation spaces at different positions. In contrast, the new activation function mainly enhances nonlinear mapping within the FFN. While both components contribute to improved performance, the optimized multi-head attention appears to be the primary drive behind the observed accuracy values.

7.3 Future directions to enhance plutonic rock classification through improving AI technologies and advancing petrographic techniques.

Our study demonstrates the potential of Transformer-based models for automating plutonic rock classification, supported by a curated dataset of Thai samples. However, misclassifications persist due to complex internal structures and overlapping mineralogical features, underscoring the need for more advanced AI methods and refined petrographic techniques.

Generative AI, particularly GANs, offers a promising direction for augmenting limited datasets. PetroGAN (Ferreira et al., 2022), for example, has shown success in generating realistic petrographic images, while GAN-based approaches have been applied to reconstruct images of hydrothermal rocks (Amiri et al., 2023) and to develop super-resolution method that enhance the resolution of two-dimensional and three-dimensional volcanic rock images obtained via X-ray microtomography (Buono et al.,

2023). These methods can expand training data and improve model robustness. Parallel efforts with Transformer-based super-resolution methods (Koeshidayatullah et al., 2023) confirm that attention mechanisms improve both image quality and interpretability. Beyond classification, segmentation models such as Trans-SedNet (Zheng et al., 2024a) demonstrate the value of dual models using XPL and PPL for mineral grain identification, a strategy that could be adapted to plutonic rock studies.

Advances in imaging also offer opportunities to improve model performance. Multi-channel optical imaging (Launeau et al., 1994) and polarized light at multiple orientations (Barraud, 2006) have already demonstrated improved mineral recognition. Combining cross-polarized and plane-polarized light has become standard (Zheng et al., 2024b), but challenges such as interference color overlap and difficulties in distinguishing clinopyroxene from hornblende remain. To address this, future studies could incorporate gypsum plates for additional contrast, multi-angle stage rotations, and varied thin-section orientations, thereby enhancing both mineral discrimination and model interpretability. Finally, broadening datasets to include more diverse plutonic rock types and geological settings will improve the generalization of deep learning models and real-world geoscience applications.

8. Conclusions

This study proposes the Petrographic Vision Transformer (PViT) for automatically classifying plutonic rock types in thin-section images from Thailand. The model was trained on a dataset composed of XPL and PPL images with 15 categories. PViT outperforms the standard ViTs with higher classification accuracy and achieves comparable performance to DaViT and EfficientNetV2 when trained from scratch on

the plutonic rock dataset under the same scheme using the same dataset. An ablation analysis was conducted to evaluate the performance of the new GELU with a sigmoid-based approximation in the position-wise feed-forward network (FFN) and the optimized multi-head attention mechanism in PViT. The combination of these components effectively improves classification accuracy and reduces redundancy by shortening training time and lowering computational resources compared with standard architectures. Additionally, the application of PViT through Grad-CAM visualization is useful for producing interpretable heatmaps in which characteristic framework grains, textures, and microstructures are given weight in the classification process. This study paves the way for further development of Transformer-based models in computer vision for classifying plutonic rock types in thin-section images.

Code availability statement

The code is made open on GitHub repository at: https://github.com/Sittiporn-GT/PViT-Petrographic-Vision-Transformer.git

CRediT authorship contribution statement

Sittiporn Kongsukho: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Warunee Maneerat**: Data curation. **Narihiro Owada**: Writing – review & editing, Software. **Adachi Tsuyoshi**: Resources, Funding acquisition. **Peerapon Vateekul**: Writing – review & editing, Formal analysis, Investigation, Conceptualization, Supervision. **Chakkaphan**

Sutthirat: Writing – review & editing, Data curation, Conceptualization, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

S.K. sincerely expresses his gratitude to the Development and Promotion of Science and Technology Talented Project (DPST) and the Institute for the Promotion of Teaching Science and Technology (IPST) in Thailand for their generous scholarship and research grant support during his doctoral studies. We also thank the Master of Science Program in Energy Geosciences at the Department of Geology, Chulalongkorn University and Mineral Economics and Mining Informatics Laboratory, Graduate School of International Resource Sciences, Akita University, for providing access to high-performance computing resources for training and evaluating models. Additionally, we are deeply grateful to the Geological Survey Division of the Department of Mineral Resources (DMR) in Thailand for their invaluable assistance

with geological information. This research was supported by the 90th Anniversary of Chulalongkorn University Scholarship under the Ratchadapisek Somphot Endowment Fund.

References

Ainslie, J., Lee-Thorp, J., De Jong, M., Zemlyanskiy, Y., Lebrón, F. and Sanghai, S., 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. arXiv preprint arXiv:2305.13245.

Aligholi, S., Lashkaripour, G.R. and Ghafoori, M., 2019. Estimating engineering properties of igneous rocks using semi-automatic petrographic analysis. Bulletin of Engineering Geology and the Environment, 78(4): 2299-2314.

Amiri, H., Vasconcelos, I., Jiao, Y., Chen, P.-E. and Plümper, O., 2023. Quantifying microstructures of earth materials using higher-order spatial correlations and deep generative adversarial networks. Scientific Reports, 13(1): 1805.

Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Barraud, J., 2006. The use of watershed segmentation and GIS software for textural analysis of thin sections. Journal of Volcanology and Geothermal Research, 154(1-2): 17-33.

Beckinsale, R.D., Suensilpong, S., Nakapadungrat, S. and Walsh, J.N., 1979. Geochronology and geochemistry of granite magmatism in Thailand in relation to a plate tectonic model. Journal of the Geological Society, 136(5): 529-537.

Buono, G., Caliro, S., Macedonio, G., Allocca, V., Gamba, F. and Pappalardo, L., 2023. Exploring microstructure and petrophysical properties of microporous volcanic

716 rocks through 3D multiscale and super-resolution imaging. Scientific Reports,
717 13(1): 6651.
718 Cai, Y.-W., Qiu, K.-F., Petrelli, M., Hou, Z.-L., Santosh, M., Yu, H.-C., Armstrong,
719 R.T. and Deng, J., 2024. The application of “transfer learning” in optical
720 microscopy: The petrographic classification of opaque minerals. American
721 Mineralogist, 109(12): 2060-2072.
722 Charusiri, P., Clark, A.H., Farrar, E., Archibald, D. and Charusiri, B., 1993. Granite
723 belts in Thailand: evidence from the 40Ar/39Ar geochronological and
724 geological syntheses. Journal of Southeast Asian Earth Sciences, 8(1): 127-136.
725 Chen, J., Yang, T., Zhang, D., Huang, H. and Tian, Y., 2021. Deep learning based
726 classification of rock structure of tunnel face. Geoscience Frontiers, 12(1): 395-
727 404.
728 Chen, M.X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L.,
729 Parmar, N., Schuster, M. and Chen, Z., 2018. The best of both worlds:
730 Combining recent advances in neural machine translation. arXiv preprint
731 arXiv:1804.09849.
732 Clevert, D.-A., Unterthiner, T. and Hochreiter, S., 2015. Fast and accurate deep network
733 learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289.
734 Cobbing, E.J., 2011. Granitic rocks.
735 Dao, T., Fu, D., Ermon, S., Rudra, A. and Ré, C., 2022. Flashattention: Fast and
736 memory-efficient exact attention with io-awareness. Advances in neural
737 information processing systems, 35: 16344-16359.
738 Ding, M., Xiao, B., Codella, N., Luo, P., Wang, J. and Yuan, L., 2022. Davit: Dual
739 attention vision transformers, European conference on computer vision.
740 Springer, pp. 74-92.

741 Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T.,
742 Dehghani, M., Minderer, M., Heigold, G. and Gelly, S., 2020. An image is
743 worth 16x16 words: Transformers for image recognition at scale. arXiv preprint
744 arXiv:2010.11929.
745 Eldosouky, A.M., El-Desoky, H.M., Henaish, A., Abdel-Rahman, A.M., Fahmy, W.,
746 El-Awny, H. and Pour, A.B., 2023. 6 - Geological data for mineral exploration.
747 In: A.B. Pour, M. Parsa and A.M. Eldosouky (Editors), Geospatial Analysis
748 Applied to Mineral Exploration. Elsevier, pp. 233-277.
749 Elfwing, S., Uchibe, E. and Doya, K., 2018. Sigmoid-weighted linear units for neural
750 network function approximation in reinforcement learning. Neural networks,
751 107: 3-11.
752 Ferreira, I., Ochoa, L. and Koeshidayatullah, A., 2022. On the generation of realistic
753 synthetic petrographic datasets using a style-based GAN. Scientific Reports,
754 12(1): 12845.
755 Gardiner, N.J., Searle, M.P., Morley, C.K., Robb, L.J., Whitehouse, M.J., Roberts,
756 N.M.W., Kirkland, C.L. and Spencer, C.J., 2018. The crustal architecture of
757 Myanmar imaged through zircon U-Pb, Lu-Hf and O isotopes: Tectonic and
758 metallogenic implications. Gondwana Research, 62: 27-60.
759 Graves, A., Wayne, G. and Danihelka, I., 2014. Neural turing machines. arXiv preprint
760 arXiv:1410.5401.
761 Guojian, C. and Peisong, L., 2021. Rock Thin-Section Image Classification based on
762 Residual Neural Network, 2021 6th International Conference on Intelligent
763 Computing and Signal Processing (ICSP), pp. 521-524.

764 Han, K., Xiao, A., Wu, E., Guo, J., Xu, C. and Wang, Y., 2021. Transformer in
765 transformer. Advances in neural information processing systems, 34: 15908-
766 15919.
767 Hendrycks, D. and Gimpel, K., 2016. Gaussian error linear units (gelus). arXiv preprint
768 arXiv:1606.08415.
769 Hussain, J., Zhang, J., Iqbal, S.M., Hussain, J., Fitria, F., Lina, X., Ali, N., Hussain, S.,
770 Akram, W. and Ali, M., 2023. Exploring the potential of late permian aggregate
771 resources for utilization in engineering structures through geotechnical,
772 geochemical and petrographic analyses. Scientific Reports, 13(1): 5088.
773 Khin, Z., Meffre, S., Lai, C.-K., Burrett, C., Santosh, M., Graham, I., Manaka, T.,
774 Salam, A., Kamvong, T. and Cromie, P., 2014. Tectonics and metallogeny of
775 mainland Southeast Asia — A review and contribution. Gondwana Research,
776 26(1): 5-30.
777 Koeshidayatullah, A., Ferreira-Chacua, I. and Li, W., 2023. Is attention all geosciences
778 need? Advancing quantitative petrography with attention-based deep learning.
779 Computers & Geosciences, 181: 105466.
780 Kongsukho, S. and Kanjanapayont, P., 2025. Kinematic flow and temperature variation
781 within the Three Pagodas shear Zone, western Thailand. Journal of Asian Earth
782 Sciences, 289: 106572.
783 Kongsukkho, S. and Kanjanapayont, P., 2022. Petrography and deformation history of
784 the three Pagodas fault, western Thailand. Bulletin of Earth Sciences of
785 Thailand, 14(1): 50-61.
786 Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep
787 convolutional neural networks. Advances in neural information processing
788 systems, 25.

789 Launeau, P., Cruden, A.R. and Bouchez, J.-L., 1994. Mineral recognition in digital
790 images of rocks; a new approach using multichannel classification. The
791 Canadian Mineralogist, 32(4): 919-933.
792 Le Maitre, R.W., Streckeisen, A., Zanettin, B., Le Bas, M., Bonin, B. and Bateman, P.,
793 2005. Igneous rocks: a classification and glossary of terms: recommendations
794 of the International Union of Geological Sciences Subcommission on the
795 Systematics of Igneous Rocks. Cambridge University Press.
796 Lee, M., 2023. Gelu activation function in deep learning: a comprehensive
797 mathematical analysis and performance. arXiv preprint arXiv:2305.12073.
798 Leem, S. and Seo, H., 2024. Attention guided CAM: Visual explanations of vision
799 transformer guided by self-attention, Proceedings of the AAAI Conference on
800 Artificial Intelligence, pp. 2956-2964.
801 Lei, S., Zhang, H., Wang, K. and Su, Z., 2019. How training data affect the accuracy
802 and robustness of neural networks for image classification.
803 Lepistö, L., Kunttu, I. and Visa, A., 2005. Color-based classification of natural rock
804 images using classifier combinations, Image Analysis: 14th Scandinavian
805 Conference, SCIA 2005, Joensuu, Finland, June 19-22, 2005. Proceedings 14.
806 Springer, pp. 901-909.
807 Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin
808 transformer: Hierarchical vision transformer using shifted windows,
809 Proceedings of the IEEE/CVF international conference on computer vision, pp.
810 10012-10022.
811 Loshchilov, I. and Hutter, F., 2017. Decoupled weight decay regularization. arXiv
812 preprint arXiv:1711.05101.

Ma, X., Zhou, C., Kong, X., He, J., Gui, L., Neubig, G., May, J. and Zettlemoyer, L., 2022. Mega: Moving average equipped gated attention. arXiv preprint arXiv:2209.10655.

Misra, D., 2019. Mish: A self regularized non-monotonic activation function. arXiv preprint arXiv:1908.08681.

Mitchell, A., 1977. Tectonic settings for emplacement of Southeast Asian tin granites.

Nair, V. and Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines, Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807-814.

Nantasin, P., Hauzenberger, C., Liu, X., Krenn, K., Dong, Y., Thöni, M. and Wathanakul, P., 2012. Occurrence of the high grade Thabsila metamorphic complex within the low grade Three Pagodas shear zone, Kanchanaburi Province, western Thailand: Petrology and geochronology. Journal of Asian Earth Sciences, 60: 68-87.

O'shea, K. and Nash, R., 2015. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

Patel, A.K., Chatterjee, S. and Gorai, A.K., 2017. Development of machine vision-based ore classification model using support vector machine (SVM) algorithm. Arabian Journal of Geosciences, 10: 1-16.

Patro, S. and Sahu, K.K., 2015. Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462.

Perez, C.A., Estévez, P.A., Vera, P.A., Castillo, L.E., Aravena, C.M., Schulz, D.A. and Medina, L.E., 2011. Ore grade estimation by feature selection and voting using boundary detection in digital image analysis. International Journal of Mineral Processing, 101(1-4): 28-36.

838 Pires de Lima, R. and Duarte, D., 2021. Pretraining Convolutional Neural Networks for
839 Mudstone Petrographic Thin-Section Image Classification. Geosciences, 11(8):
840 336.
841 Polat, Ö., Polat, A. and Ekici, T., 2021. Automatic classification of volcanic rocks from
842 thin section images using transfer learning networks. Neural Computing and
843 Applications, 33(18): 11531-11540.
844 Qiqi, Z., Sai, W., Shun, T., Liangbin, Y., Kunlun, Q. and Qingfeng, G., 2024.
845 RockS2Net: Rock image classification via a spatial localization siamese
846 network. Computers & Geosciences, 185: 105560.
847 Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object
848 detection with region proposal networks. Advances in neural information
849 processing systems, 28.
850 Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017.
851 Grad-cam: Visual explanations from deep networks via gradient-based
852 localization, Proceedings of the IEEE international conference on computer
853 vision, pp. 618-626.
854 Seo, W., Kim, Y., Sim, H., Song, Y. and Yun, T.S., 2022. Classification of igneous
855 rocks from petrographic thin section images using convolutional neural
856 network. Earth Science Informatics, 15(2): 1297-1307.
857 Shin, H.-C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. and
858 Summers, R.M., 2016. Deep convolutional neural networks for computer-aided
859 detection: CNN architectures, dataset characteristics and transfer learning. IEEE
860 transactions on medical imaging, 35(5): 1285-1298.
861 Streckeisen, A., 1976. To each plutonic rock its proper name. Earth-Science Reviews,
862 12(1): 1-33.

Su, C., Xu, S.-j., Zhu, K.-y. and Zhang, X.-c., 2020. Rock classification in petrographic thin section images based on concatenated convolutional neural networks. Earth Science Informatics, 13(4): 1477-1484.

Tan, M. and Le, Q., 2021. Efficientnetv2: Smaller models and faster training, International conference on machine learning. PMLR, pp. 10096-10106.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jégou, H., 2021. Training data-efficient image transformers & distillation through attention, International conference on machine learning. PMLR, pp. 10347-10357.

Vaswani, A., Ramachandran, P., Srinivas, A., Parmar, N., Hechtman, B. and Shlens, J., 2021. Scaling local self-attention for parameter efficient visual backbones, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12894-12904.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Wang, C., Li, P., Long, Q., Chen, H., Wang, P., Meng, Z., Wang, X. and Zhou, Y., 2024. Deep Learning for Refined Lithology Identification of Sandstone Microscopic Images. Minerals, 14(3): 275.

Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L. and Gao, J., 2021. Focal self-attention for local-global interactions in vision transformers. arXiv preprint arXiv:2107.00641.

Yang, Y.-Y., Rashtchian, C., Zhang, H., Salakhutdinov, R.R. and Chaudhuri, K., 2020. A closer look at accuracy vs. robustness. Advances in neural information processing systems, 33: 8588-8601.

887 Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M. and Wu, Y., 2022.
888 Coca: Contrastive captioners are image-text foundation models. arXiv preprint
889 arXiv:2205.01917.
890 Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.-H., Tay, F.E., Feng, J. and
891 Yan, S., 2021. Tokens-to-token vit: Training vision transformers from scratch
892 on imagenet, Proceedings of the IEEE/CVF international conference on
893 computer vision, pp. 558-567.
894 Zhang, H. and Wang, S., 2023. An Optimized Hybrid Transformer for Enhanced Ultra-
895 Fine-Grained Thin Sections Categorization via Integrated Region-to-Region
896 and Token-to-Token Approaches. Applied Sciences, 13(13): 7853.
897 Zhang, X., Zeng, H., Guo, S. and Zhang, L., 2022. Efficient long-range attention
898 network for image super-resolution, European conference on computer vision.
899 Springer, pp. 649-667.
900 Zheng, D., Hou, L., Hu, X., Hou, M., Dong, K., Hu, S., Teng, R. and Ma, C., 2024a.
901 Sediment grain segmentation in thin-section images using dual-modal Vision
902 Transformer. Computers & Geosciences, 191: 105664.
903 Zheng, D., Zhong, H., Camps-Valls, G., Cao, Z., Ma, X., Mills, B., Hu, X., Hou, M.
904 and Ma, C., 2024b. Explainable deep learning for automatic rock classification.
905 Computers & Geosciences, 184: 105511.
906 Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A., 2016. Learning deep
907 features for discriminative localization, Proceedings of the IEEE conference on
908 computer vision and pattern recognition, pp. 2921-2929.
909 Zhu, Y., Bai, L., Peng, W., Zhang, X. and Luo, X., 2019. Depthwise separable
910 convolution feature learning for ihomogeneous rock image classification,
911 Cognitive Systems and Signal Processing: 4th International Conference,

ICCSIP 2018, Beijing, China, November 29-December 1, 2018, Revised Selected Papers, Part I 4. Springer, pp. 165-176.

Figure captions

Fig. 1. Geological map of Thailand showing granite distributions and sampling locations. Red boxes indicate the study areas: (A) Kanchanaburi Province, representing the Central Granite Belt; (B) Chachoengsao, Chon Buri, and Chanthaburi Provinces, representing the Eastern and Central Granite Belts; and (C) Phetchaburi and Prachuap Khiri Khan Provinces, representing the Western Granite Belt.

Fig. 2. Diagrams illustrating the International Union of Geological Sciences (IUGS) classification system for plutonic rocks (after Le Maitre et al. (2005): (A) Granitic and alkaline rocks, classified based on the modal proportions of quartz (Q), alkali feldspar (A), plagioclase (P), and feldspathoids (F); (B) Gabbroic rocks, based on the modal proportions of plagioclase, clinopyroxene (Cpx), and orthopyroxene (Opx); (C) Ultramafic rocks, based on the modal proportions of olivine (Ol), clinopyroxene (Cpx), and orthopyroxene (Opx).

Fig. 3. Fifteen studied plutonic rock types, each illustrated under plane-polarized light (PPL) and cross-polarized light (XPL). Rock types include: (A) Anorthosite, (B) Diorite, (C) Dunite, (D) Foidolite, (E) Gabbro, (F) Granite, (G) Granodiorite, (H) Greisen, (I) Harzburgite, (J) Hornblendite, (K) Lherzolite, (L) Monzonite, (M) Norite, (N) Orthopyroxenite, and (O) Syenite.

Fig. 4. Number of thin-section images in the training, validation, and test datasets. AN = Anorthosite, DI = Diorite, DU = Dunite, FO = Foidolite, GA = Gabbro, GR = Granite, GD = Granodiorite, GS = Greisen, HZ = Harzburgite, HB = Hornblendite, LH = Lherzolite, MO = Monzonite, NO = Norite, PY = Orthopyroxenite, and SY = Syenite.

Fig. 5. Overall architecture of Vision Transformer (ViT). (A) Simplified architecture of Vision Transformer based on the model developed by Dosovitskiy et al. (2020). (B) Details of the patch embedding layer. (C) Explanation of the self-attention mechanism within a Transformer encoder block. (D) Internal workflow of the MLP classification head.

Fig. 6. Overview of the Petrographic Vision Transformer (PViT) backbone. (A) Simplified architectures of PViT with two modifications, which are New GELU within the feed-forward network (FFN) and optimized multi-head attention mechanism. (B) The details of the optimized multi-head attention mechanism by fusing the Query (Q), Key (K), and Value (V) projections into a single joint linear transformation.

Fig. 7. Explanation of optimized multi-head attention.

Fig. 8. Workflow for generating visualization heatmaps, adapted from Leem and Seo (2024).

Fig. 9. (A) Loss curves (B) accuracy curves of Petrographic Vision Transformer (PViT), Vision Transformer (ViT), Dual Attention Vision Transformer (DaViT), and EfficientNetV2. The results of the training dataset are represented in solid lines, and the

results of the validation dataset are represented in dashed lines. (C) confusion matrices of PViT using the training dataset, and (D) confusion matrices of PViT using the test dataset. AN = Anorthosite, DI = Diorite, DU = Dunite, FO = Foidolite, GA = Gabbro, GR = Granite, GD = Granodiorite, GS = Greisen, HZ = Harzburgite, HB = Hornblendite, LH = Lherzolite, MO = Monzonite, NO = Norite, PY = Orthopyroxenite, SY = Syenite.

Fig. 10. Grad-CAM visualizations of Petrographic Vision Transformer (PViT), the heatmap highlights influential regions where models give more weight. (A) Anorthosite; (B) Diorite; (C) Dunite; (D) Foidolite; (E) Granodiorite; (F) Harzburgite; (G) Hornblendite; (H) Lherzolite; (I) Gabbro; (J) Granite; (K) Syenite; (L) Monzonite; (M) Greisen; (N) Norite; and (O) Orthopyroxenite.

Fig. 11. Performance of activation functions and attention mechanisms. (A) Accuracy across activation functions. (B) Accuracy versus throughput for different activation functions. (C) Accuracy across attention mechanisms with GELU and New GELU. (D) Accuracy versus throughput for different attention mechanisms.

Fig. 12. Comparison of model performance. (A) Accuracy versus number of parameters for PViT and stage-of-the-art (SOTA) models trained from scratch on plutonic rock datasets. (B) Accuracy versus number of parameters for PViT with state-of-the-art (SOTA) models pretrained on ImageNet-1K and fine-tuned with a linear head on the same dataset.

Table caption

Table 1. Configurations of the Petrographic Vision Transformer (PViT).

Model	Layers	Hidden size D	MLP size	Heads	Params
PViT-Base	12	768	3072	12	86 M
PViT-Large	24	1024	4096	16	307 M
PViT-Huge	32	1280	5120	16	632 M

Table 2. Performance comparison of Petrographic Vision Transformer (PViT) and state-of-the-art (SOTA) trained from scratch on the plutonic rock dataset.

Model	Model evaluation							
	Accuracy		Precision		Recall		F1-Score	
	Train	Test	Train	Test	Train	Test	Train	Test
Petrographic Vision Transformer: PViT (our study)								
PViT-Base	99.97	98.34	99.97	98.45	99.97	98.33	99.97	98.34
PViT-Large	99.96	97.52	99.97	97.61	99.97	97.52	99.97	97.51
PViT-Huge	98.05	96.07	97.66	95.56	97.83	95.11	97.72	95.17
Vision Transformer: ViT (Dosovitskiy et al., 2020)								
ViT-Base	99.63	97.72	99.64	97.89	99.63	97.72	99.63	97.66
ViT-Large	99.30	96.07	99.32	96.38	99.30	96.07	99.30	96.02
ViT-Huge	99.27	96.07	99.28	96.18	99.27	96.07	99.27	96.04
Dual Attention Vision Transformer: DaViT (Ding et al., 2022)								
DaViT-Base	100.00	96.27	100.00	95.71	100.00	93.54	100.00	94.30
DaViT-Large	100.00	98.34	100.00	98.44	100.00	98.34	100.00	98.34
EfficientNetV2 (Tan and Le, 2021)								
EfficientNetV2-M	100.00	98.14	100.00	98.21	100.00	98.17	100.00	98.12
EfficientNetV2-L	99.99	98.34	99.96	98.45	99.97	98.34	99.97	98.34

Table 3. Comparison of optimized multi-head attention and standard multi-head attention across different activation functions trained from scratch on the plutonic rock dataset. The PViT-B/16 serves as the baseline with the same training scheme applied for a fair comparison.

PViT- Base model 86M Parameters	**Standard multi-head attention** Vaswani et al. (2017)				**Optimized multi-head attention** ***(our study)***			
	Accuracy (Top1 %)	**Throughput (image/second)**	**Training Time (second/image)**	**Maximum GPU Memory (GB)**	**Accuracy (Top1 %)**	**Throughput (image/second)**	**Training Time (second/image)**	**Maximum GPU Memory (GB)**
Standard GELU Tanh-based approximation Hendrycks and Gimpel (2016)	97.72	14.50	0.069	19.70	97.93 (+0.21)	18.00 (+3.5)	0.056 (-0.013)	19.70 (0)
SiLU Sigmoid-Weighted Linear Unit Elfwing et al. (2018)	96.69 (-1.03)	14.80 (+0.3)	0.068 (-0.001)	16.00 (-3.7)	96.27 (-1.45)	19.00 (+4.5)	0.053 (-0.016)	16.20 (-3.5)
Mish Mish activation function Misra (2019)	97.31 (-0.41)	15.30 (+0.8)	0.065 (-0.004)	17.30 (-2.4)	96.48 (-1.24)	19.10 (+4.6)	0.052 (-0.017)	17.70 (-2.0)
Laplace Laplace activation function Ma et al. (2022)	95.45 (-2.27)	14.90 (+0.4)	0.067 (-0.002)	17.60 (-2.1)	94.00 (-3.72)	18.40 (+3.9)	0.054 (-0.015)	16.30 (-3.4)
New GELU Sigmoid-based approximation ***(our study)***	97.31 (-0.41)	15.00 (+0.5)	0.067 (-0.002)	17.50 (-2.2)	***98.34 (+0.62)***	***18.70 (+4.2)***	***0.053 (-0.016)***	***17.20 (-2.5)***

Table 4. Comparison of standard GELU and New GELU across different attention mechanisms trained from scratch on the plutonic rock dataset. The PViT-B/16 serves as the baseline with the same training scheme applied for a fair comparison.

PViT- Base model 86M Parameters	**Standard GELU: Tanh-based approximation** Hendrycks and Gimpel (2016)				**NewGELU: Sigmoid-based approximation** ***(our study)***			
	Accuracy (Top1 %)	**Throughput (image/second)**	**Training Time (second/image)**	**Maximum GPU Memory (GB)**	**Accuracy (Top1 %)**	**Throughput (image/second)**	**Training Time (second/image)**	**Maximum GPU Memory (GB)**

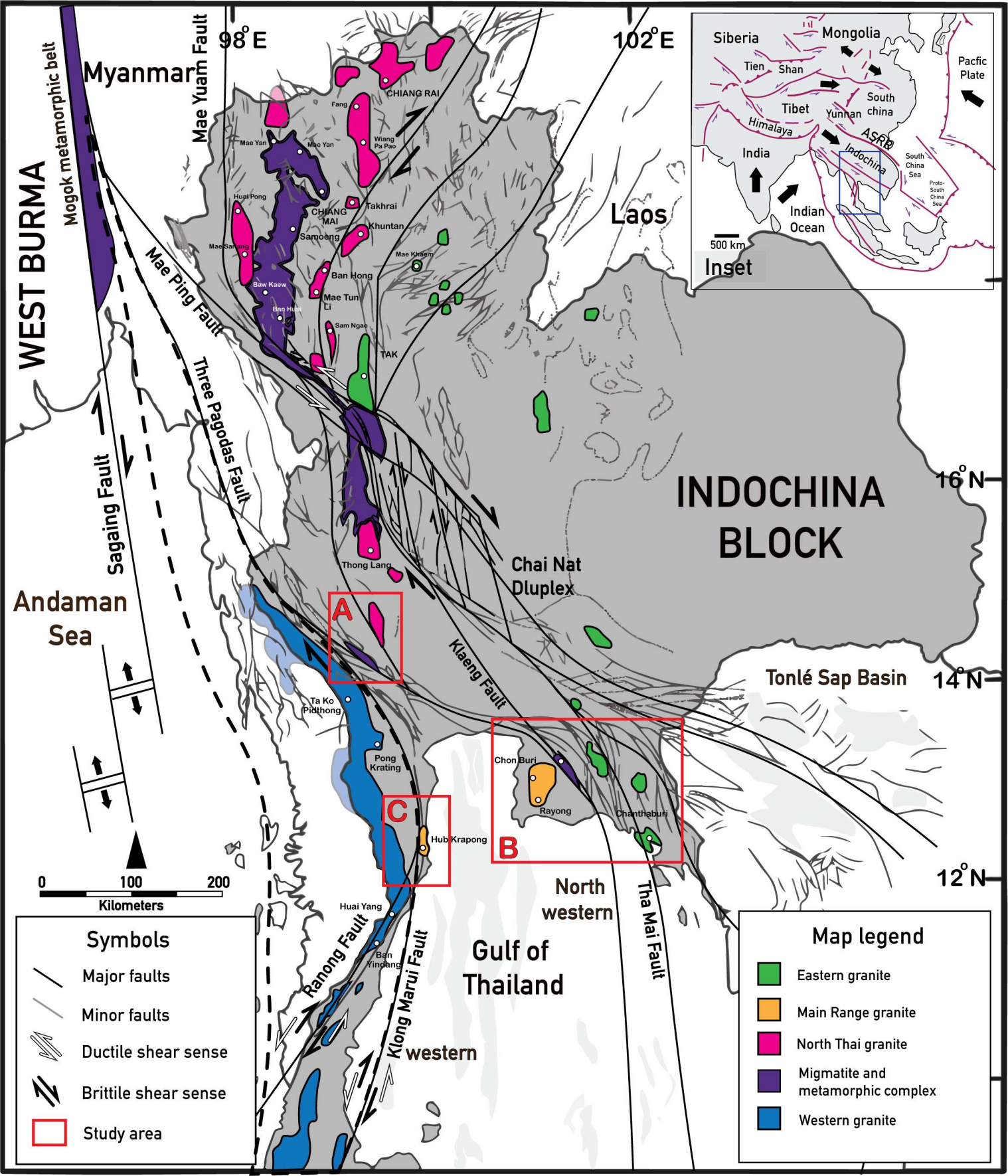
Standard Multi-Head Attention Vaswani et al. (2017)	97.72	14.50	0.069	19.70	97.31 (-0.41)	15.00 (+0.5)	0.067 (-0.002)	17.50 (-2.2)
Multi-Query attention Ainslie et al. (2023)	97.31 (-0.41)	19.40 (+4.9)	0.052 (-0.017)	19.20 (-0.5)	96.69 (-1.03)	20.20 (+5.7)	0.050 (-0.019)	17.00 (-2.7)
Talking Head attention Shazeer et al. (2020)	96.89 (-0.83)	11.30 (-3.2)	0.088 (+0.019)	28.50 (+8.8)	96.48 (-1.24)	11.67 (-2.8)	0.086 (+0.017)	26.41 (+6.7)
FlashAttention Dao et al. (2022)	93.79 (-3.93)	21.40 (+6.9)	0.047 (-0.022)	9.20 (-10.5)	97.31 (-0.41)	22.40 (+7.9)	0.045 (-0.024)	6.90 (-12.8)
Optimized Multi-Head Attention ***(our study)***	97.93 (+0.21)	18.00 (+3.5)	0.056 (-0.013)	19.70 (0)	***98.34 (+0.62%)***	***18.70 (+4.2)***	***0.053 (-0.016)***	***17.20 (-2.5)***

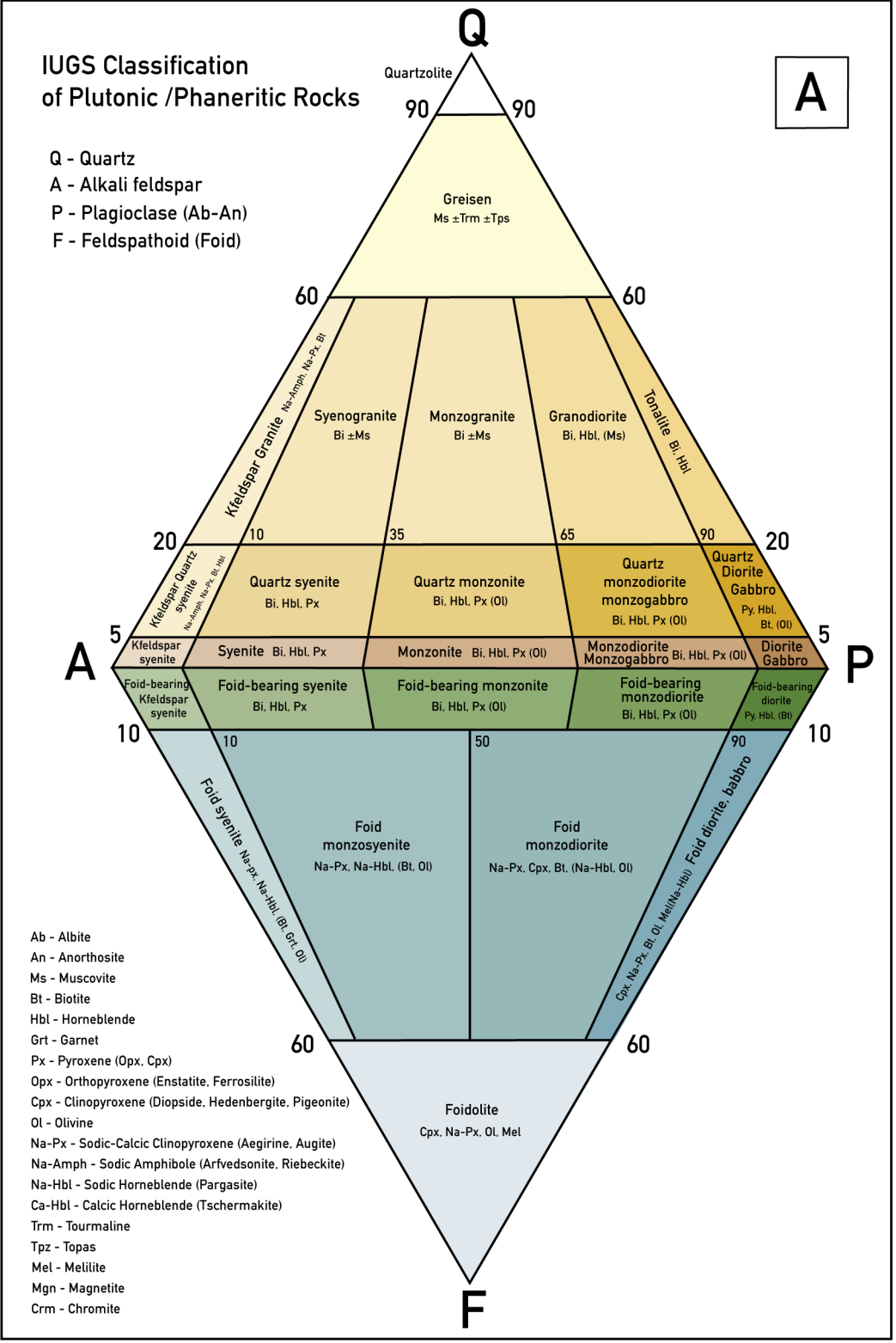
Table 5. Performance comparison between Petrographic Vision Transformer (PViT) and state-of-the-art (SOTA) models pretrained on ImageNet and fine-tuned with a linear head on the same dataset.

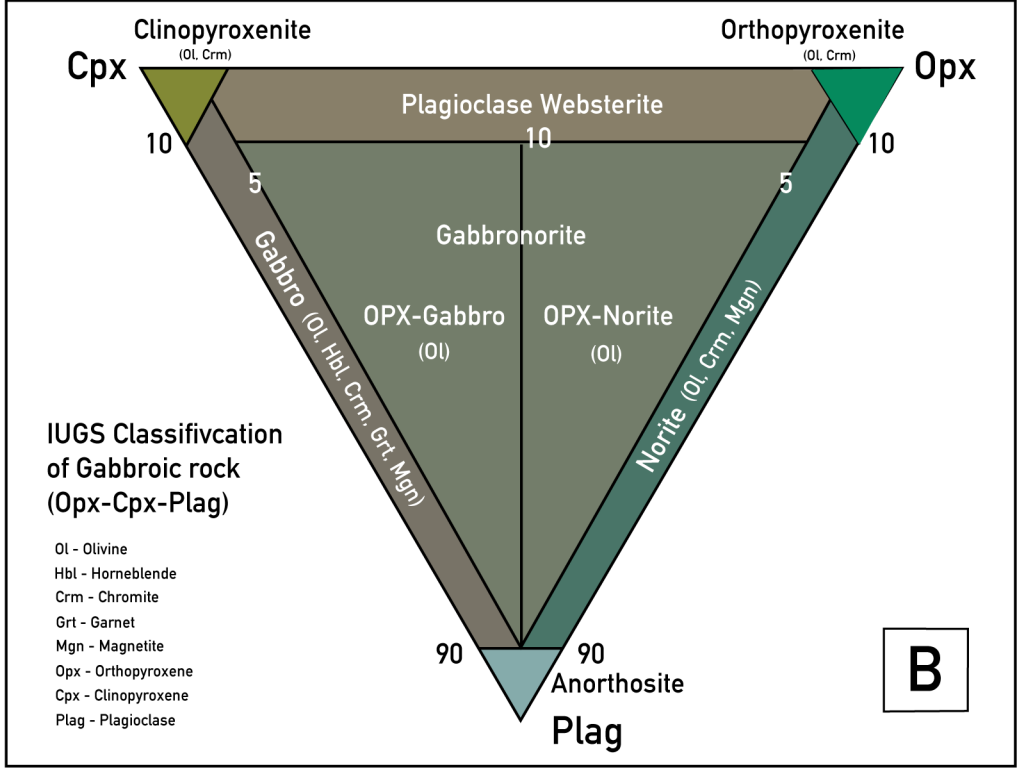
Model	**Model evaluation**								**Parameter (M)**
	Accuracy		**Precision**		**Recall**		**F1-Score**		
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	
(a) Training from scratch on the plutonic rock dataset									
PViT-B (Our study)	99.97	98.34	99.97	98.45	99.97	98.33	99.97	98.34	86
PViT-L (Our study)	99.96	97.52	99.97	97.61	99.97	97.52	99.97	97.51	307
PViT-H (Our study)	98.05	96.07	97.66	95.56	97.83	95.11	97.72	95.17	632
ViT-B/32 Dosovitskiy et al. (2020)	99.70	94.82	99.70	95.08	99.70	94.82	99.70	94.85	86
ViT-B/16 Dosovitskiy et al. (2020)	99.63	97.72	99.64	97.89	99.63	97.72	99.63	97.66	86
ViT-L/32 Dosovitskiy et al. (2020)	99.82	92.96	99.82	93.33	99.82	92.96	99.82	92.88	307
ViT-L/16 Dosovitskiy et al. (2020)	99.30	96.07	99.32	96.38	99.30	96.07	99.30	96.02	307
ViT-H/32 Dosovitskiy et al. (2020)	99.36	92.75	99.37	93.41	99.36	92.75	99.36	92.77	632
ViT-H/16 Dosovitskiy et al. (2020)	99.27	96.07	99.28	96.18	99.27	96.07	99.27	96.04	632
(b) ImageNet pre-trained models and fine-tuned with a linear head on the plutonic rock dataset									

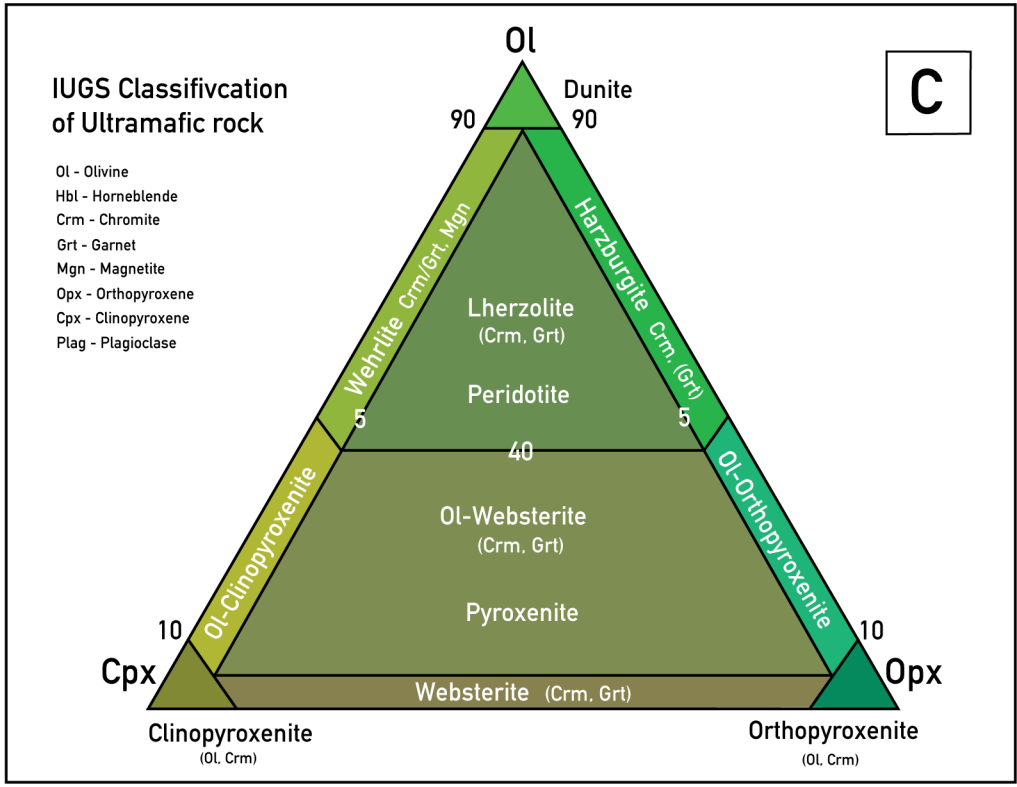
DeiT-T/16 (Distillation) Touvron et al. (2021)	99.03	97.10	99.04	97.16	99.03	97.10	99.03	97.09	6
DeiT-S/16 (Distillation) Touvron et al. (2021)	99.85	98.76	99.85	98.81	99.85	98.76	99.85	98.76	22
DeiT-B/16 (Distillation) Touvron et al. (2021)	99.79	99.38	99.79	99.39	99.79	99.38	99.79	99.38	87
DeiT-B/16 (Distillation-384) Touvron et al. (2021)	1.00	99.59	1.00	99.61	1.00	99.59	1.00	99.59	87
Swin-Ti (224 x 224-input) Liu et al. (2021)	1.00	99.38	1.00	99.38	1.00	99.38	1.00	99.38	28
Swin-S (224 x 224-input) Liu et al. (2021)	99.85	99.38	99.85	99.41	99.85	99.38	99.85	99.37	50
Swin-B (224 x 224-input) Liu et al. (2021)	99.94	98.96	99.94	99.03	99.94	98.96	99.94	98.96	88
Swin-L (224 x 224-input) Liu et al. (2021)	99.97	99.59	99.6	99.97	99.59	99.97	99.59	99.97	197
Swin-B (384 x 384-input) Liu et al. (2021)	99.94	99.17	99.94	99.23	99.94	99.17	99.94	99.17	88
Swin-L (384 x 384-input) Liu et al. (2021)	99.91	1.00	99.91	1.00	99.91	1.00	99.91	1.00	197
SwinV2-Ti (window8) Liu et al. (2021)	1.00	99.79	1.00	99.80	1.00	99.79	1.00	99.79	28
SwinV2-S (window8) Liu et al. (2021)	99.94	1.00	99.94	1.00	99.94	1.00	99.94	1.00	50
SwinV2-B (window8) Liu et al. (2021)	99.94	1.00	99.94	1.00	99.94	1.00	99.94	1.00	88
SwinV2-Ti (window16) Liu et al. (2021)	99.88	99.79	99.88	99.80	99.88	99.79	99.88	99.79	28
SwinV2-S (window16) Liu et al. (2021)	1.00	99.79	1.00	99.80	1.00	99.79	1.00	99.79	50
SwinV2-B (window16) Liu et al. (2021)	99.97	99.79	99.97	99.80	99.97	99.79	99.97	99.79	88
DaViT-Ti Ding et al. (2022)	1.00	99.79	1.00	99.80	1.00	99.79	1.00	99.79	28
DaViT-S Ding et al. (2022)	1.00	99.79	1.00	99.80	1.00	99.79	1.00	99.79	50

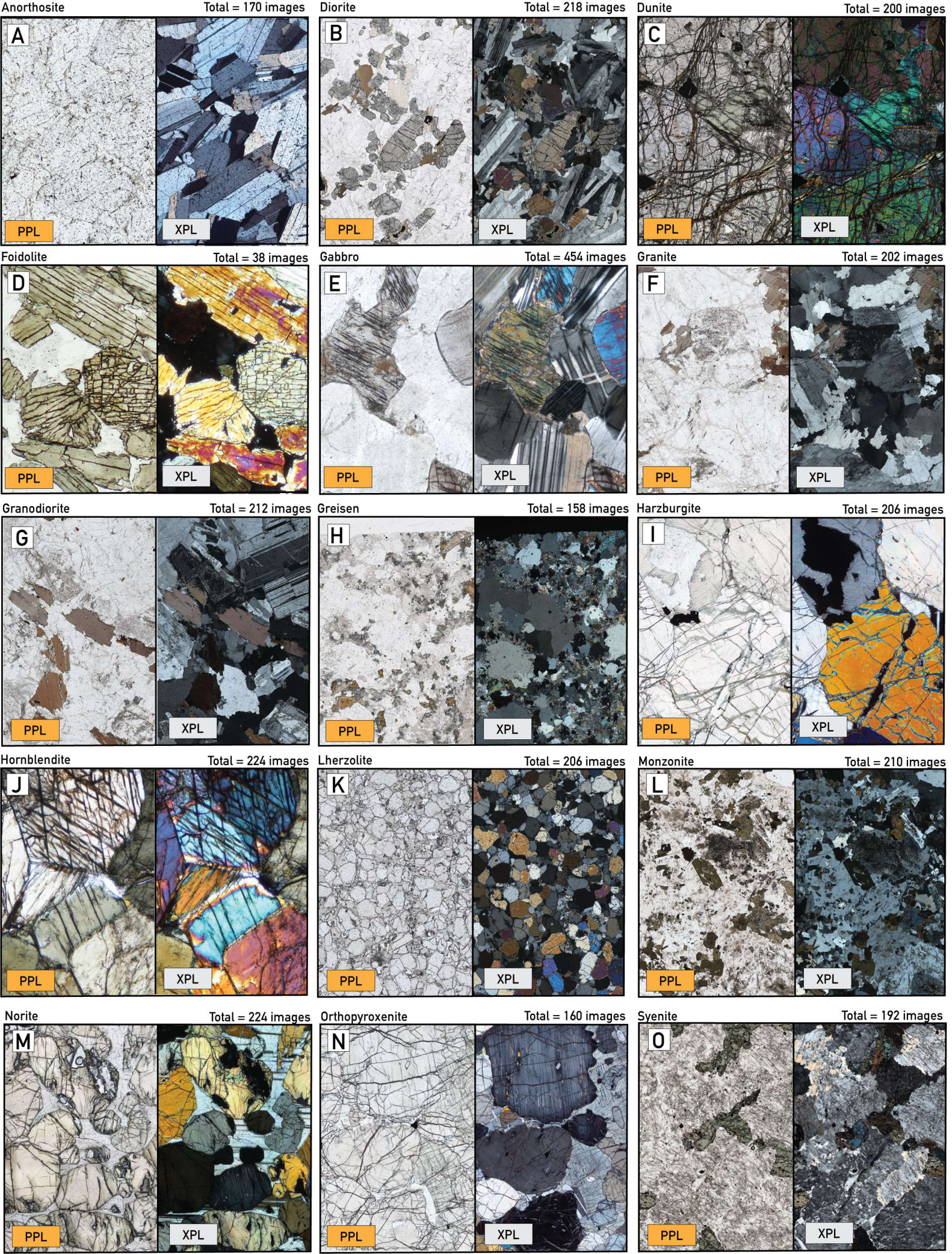
DaViT-B Ding et al. (2022)	99.91	1.00	99.91	1.00	99.91	1.00	99.91	1.00	88
CoCa-B/32 (Finetune) Yu et al. (2022)	94.39	86.54	95.51	90.04	94.39	86.54	93.89	85.82	86
CoCa-L/14 (Finetune) Yu et al. (2022)	99.73	98.96	99.74	99.13	99.73	98.96	99.73	98.97	303
EfficientNetV2-S Tan and Le (2021)	1.00	99.59	1.00	99.60	1.00	99.59	1.00	99.59	22
EfficientNetV2-M Tan and Le (2021)	1.00	99.17	1.00	99.23	1.00	99.17	1.00	99.18	54
EfficientNetV2-L Tan and Le (2021)	1.00	99.79	1.00	99.80	1.00	99.79	1.00	99.79	120
EfficientNetV2-XL Tan and Le (2021)	1.00	99.79	1.00	99.71	1.00	99.79	1.00	99.79	208

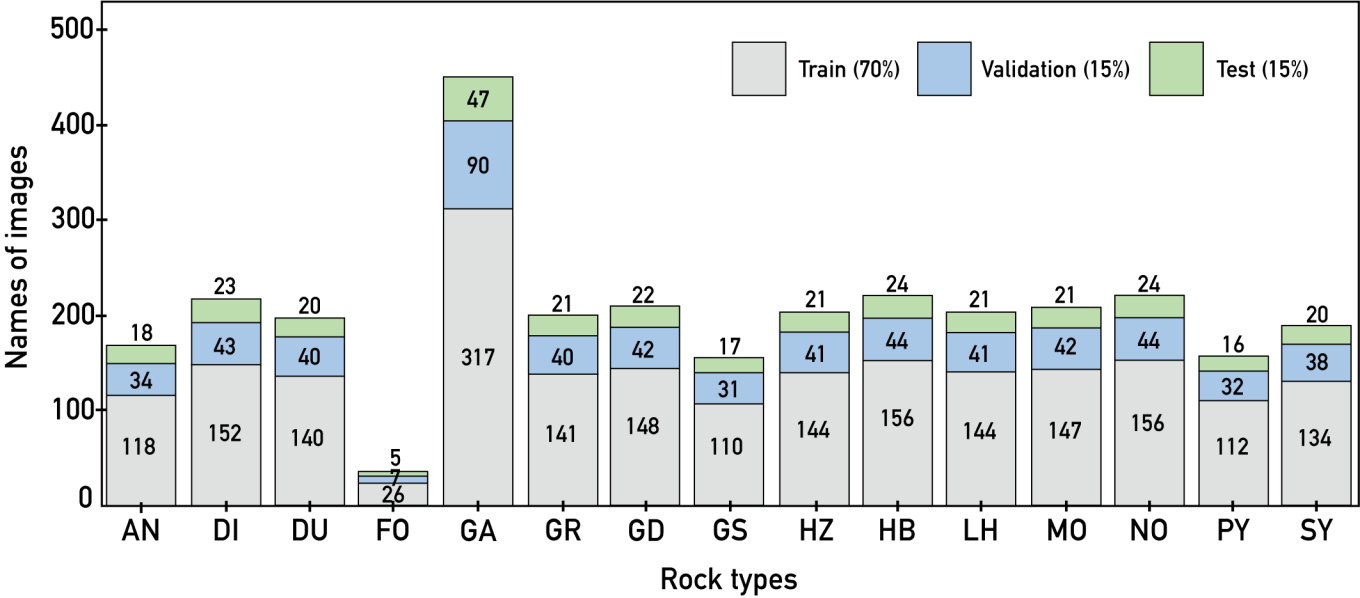

Myanmar
WEST BURMA
Mogok metamorphic belt
Mae Yuam Fault
98° E
102° E
CHIANG RAI
Fang
Wiang Pa Pao
Mae Yan
Mae Yan
Huai Pong
CHIANG MAI
Takhrai
Khuntan
Samoeng
Mae Sariang
Mae Khaem
Ban Hong
Baw Kaew
Mae Tun Li
Ban Huai
Sam Ngao
TAK
Laos
Mae Ping Fault
Three Pagodas Fault
Sagaing Fault
INDOCHINA BLOCK
16° N
Thong Lang
Chai Nat
Dluplex
Andaman Sea
A
Klaeng Fault
Tonlé Sap Basin
14° N
Ta Ko Pidthong
Pong Krating
Chon Buri
Rayong
Chanthaburi
C
Hub Krapong
B
North western
Tha Mai Fault
12° N
Huai Yang
Ranong Fault
Ban Yindang
Klong Marui Fault
Gulf of Thailand
western
0
100
200
Kilometers
Symbols
Major faults
Minor faults
Ductile shear sense
Brittile shear sense
Study area
Map legend
Eastern granite
Main Range granite
North Thai granite
Migmatite and metamorphic complex
Western granite
Siberia
Mongolia
Tien Shan
Pacfic Plate
South china
Tibet
Yunnan
Himalaya
ASRB
India
Indochina
South China Sea
Proto-South China Sea
Indian Ocean
500 km
Inset

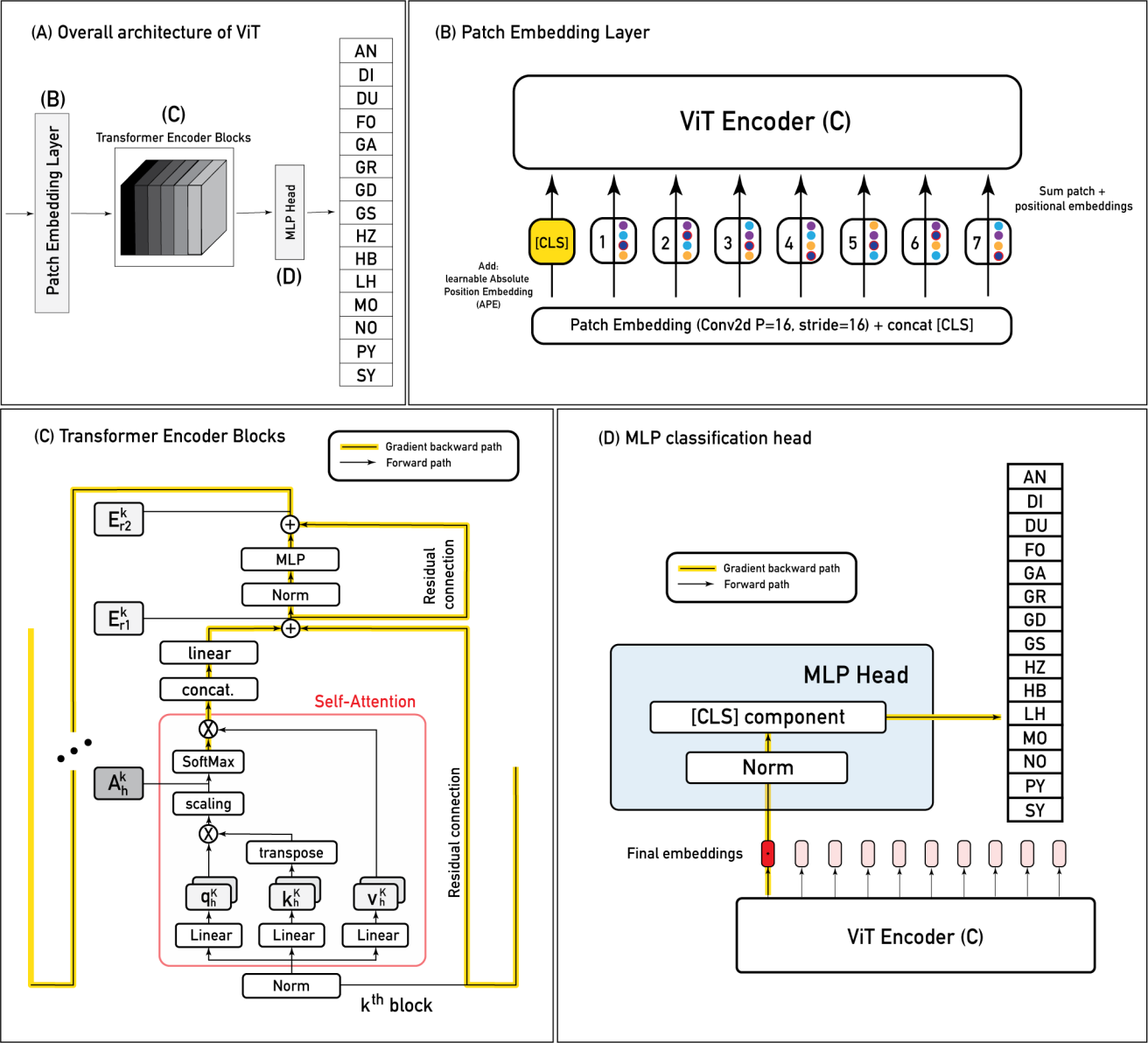

IUGS Classification
of Plutonic /Phaneritic Rocks
A
Q - Quartz
A - Alkali feldspar
P - Plagioclase (Ab-An)
F - Feldspathoid (Foid)
Q
A
P
F
Quartzolite
90
90
Greisen
Ms ±Trm ±Tps
60
60
Kfeldspar Granite Na-Amph, Na-Px, Bt
Syenogranite
Bi ±Ms
Monzogranite
Bi ±Ms
Granodiorite
Bi, Hbl, (Ms)
Tonalite Bi, Hbl
20
20
10
35
65
90
Kfeldspar Quartz syenite Na-Amph, Na-Px, Bi, Hbl
Quartz syenite
Bi, Hbl, Px
Quartz monzonite
Bi, Hbl, Px (Ol)
Quartz monzodiorite monzogabbro
Bi, Hbl, Px (Ol)
Quartz Diorite Gabbro
Py, Hbl, Bt, (Ol)
5
5
Kfeldspar syenite
Syenite Bi, Hbl, Px
Monzonite Bi, Hbl, Px (Ol)
Monzodiorite Monzogabbro Bi, Hbl, Px (Ol)
Diorite Gabbro
Foid-bearing Kfeldspar syenite
Foid-bearing syenite
Bi, Hbl, Px
Foid-bearing monzonite
Bi, Hbl, Px (Ol)
Foid-bearing monzodiorite
Bi, Hbl, Px (Ol)
Foid-bearing diorite
Py, Hbl, (Bt)
10
10
10
50
90
Foid syenite Na-px, Na-Hbl, (Bt, Grt, Ol)
Foid monzosyenite
Na-Px, Na-Hbl, (Bt, Ol)
Foid monzodiorite
Na-Px, Cpx, Bt, (Na-Hbl, Ol)
Foid diorite, babbro
Cpx, Na-Px, Bt, Ol, Mel(Na-Hbl)
60
60
Foidolite
Cpx, Na-Px, Ol, Mel
Ab - Albite
An - Anorthosite
Ms - Muscovite
Bt - Biotite
Hbl - Horneblende
Grt - Garnet
Px - Pyroxene (Opx, Cpx)
Opx - Orthopyroxene (Enstatite, Ferrosilite)
Cpx - Clinopyroxene (Diopside, Hedenbergite, Pigeonite)
Ol - Olivine
Na-Px - Sodic-Calcic Clinopyroxene (Aegirine, Augite)
Na-Amph - Sodic Amphibole (Arfvedsonite, Riebeckite)
Na-Hbl - Sodic Horneblende (Pargasite)
Ca-Hbl - Calcic Horneblende (Tschermakite)
Trm - Tourmaline
Tpz - Topas
Mel - Melilite
Mgn - Magnetite
Crm - Chromite

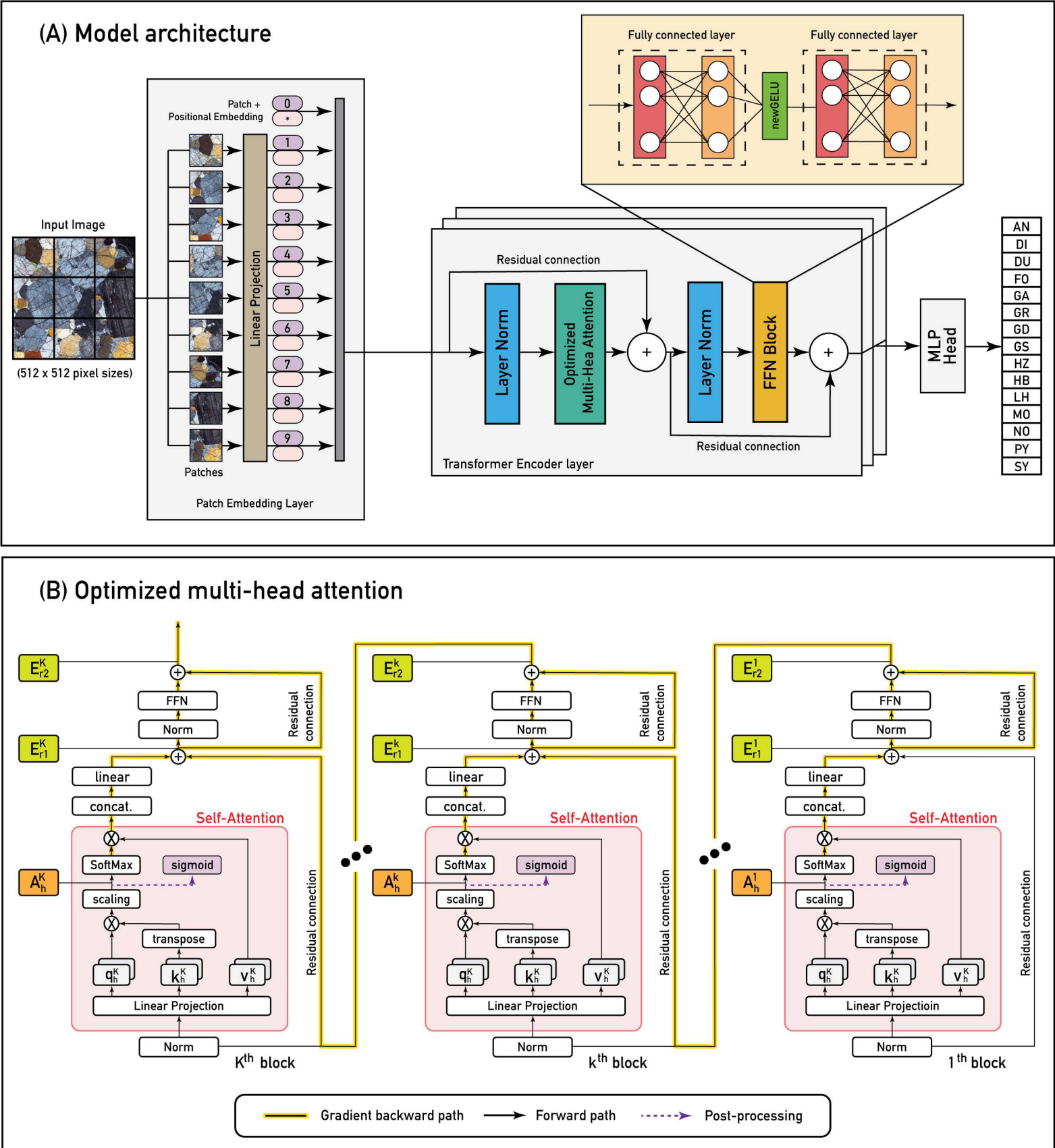

Clinopyroxenite
(Ol, Crm)
Orthopyroxenite
(Ol, Crm)
Cpx
Opx
Plagioclase Websterite
10
10
5
5
Gabbronorite
Gabbro (Ol, Hbl, Crm, Grt, Mgn)
Norite (Ol, Crm, Mgn)
OPX-Gabbro
(Ol)
OPX-Norite
(Ol)
90
90
Anorthosite
Plag
IUGS Classifivcation
of Gabbroic rock
(Opx-Cpx-Plag)
Ol - Olivine
Hbl - Horneblende
Crm - Chromite
Grt - Garnet
Mgn - Magnetite
Opx - Orthopyroxene
Cpx - Clinopyroxene
Plag - Plagioclase
B

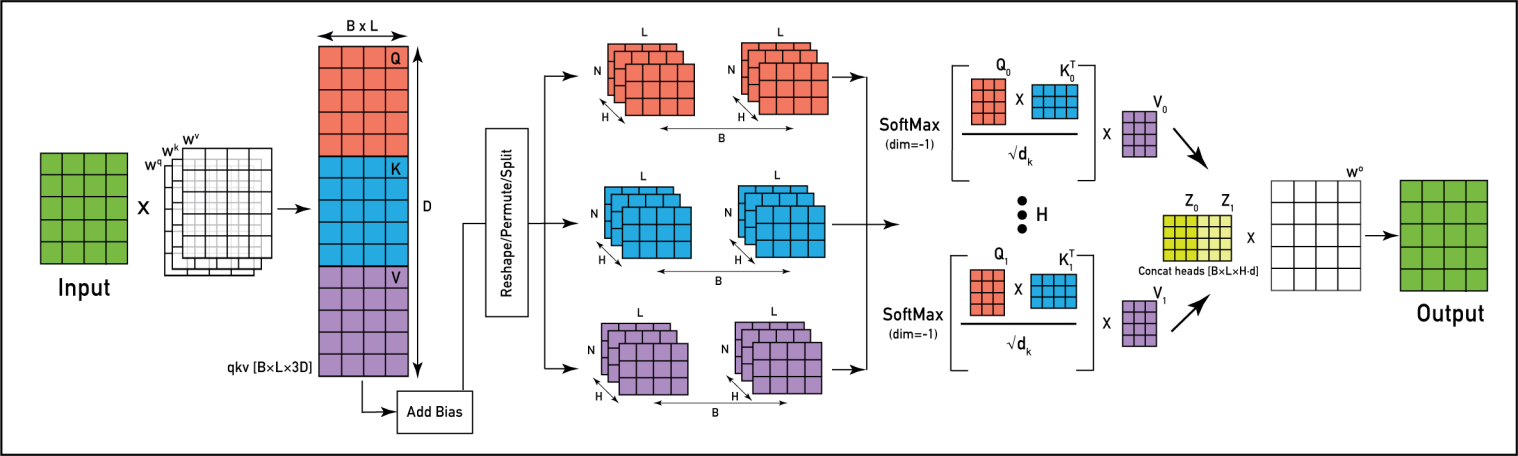

IUGS Classifivcation
of Ultramafic rock
C
Ol - Olivine
Hbl - Horneblende
Crm - Chromite
Grt - Garnet
Mgn - Magnetite
Opx - Orthopyroxene
Cpx - Clinopyroxene
Plag - Plagioclase
Ol
Dunite
90
90
Wehrlite Crm/(Grt, Mgn
Harzburgite Crm, (Grt)
Lherzolite
(Crm, Grt)
Peridotite
5
5
40
Ol-Clinopyroxenite
Ol-Orthopyroxenite
Ol-Websterite
(Crm, Grt)
Pyroxenite
10
10
Cpx
Opx
Websterite (Crm, Grt)
Clinopyroxenite
(Ol, Crm)
Orthopyroxenite
(Ol, Crm)
C

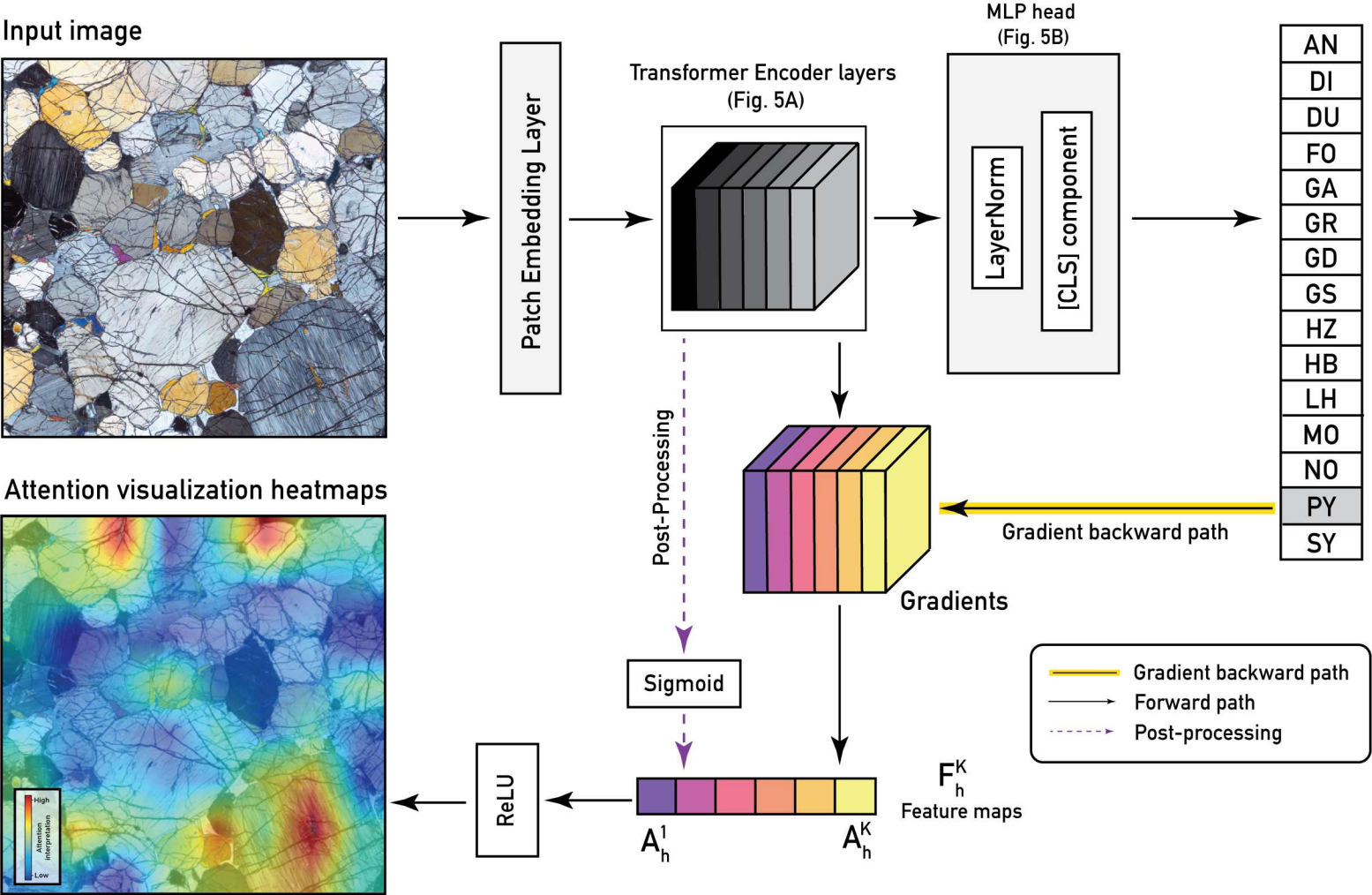

Anorthosite
Total = 170 images
A
PPL
XPL
Diorite
Total = 218 images
B
PPL
XPL
Dunite
Total = 200 images
C
PPL
XPL
Foidolite
Total = 38 images
D
PPL
XPL
Gabbro
Total = 454 images
E
PPL
XPL
Granite
Total = 202 images
F
PPL
XPL
Granodiorite
Total = 212 images
G
PPL
XPL
Greisen
Total = 158 images
H
PPL
XPL
Harzburgite
Total = 206 images
I
PPL
XPL
Hornblendite
Total = 224 images
J
PPL
XPL
Lherzolite
Total = 206 images
K
PPL
XPL
Monzonite
Total = 210 images
L
PPL
XPL
Norite
Total = 224 images
M
PPL
XPL
Orthopyroxenite
Total = 160 images
N
PPL
XPL
Syenite
Total = 192 images
O
PPL
XPL

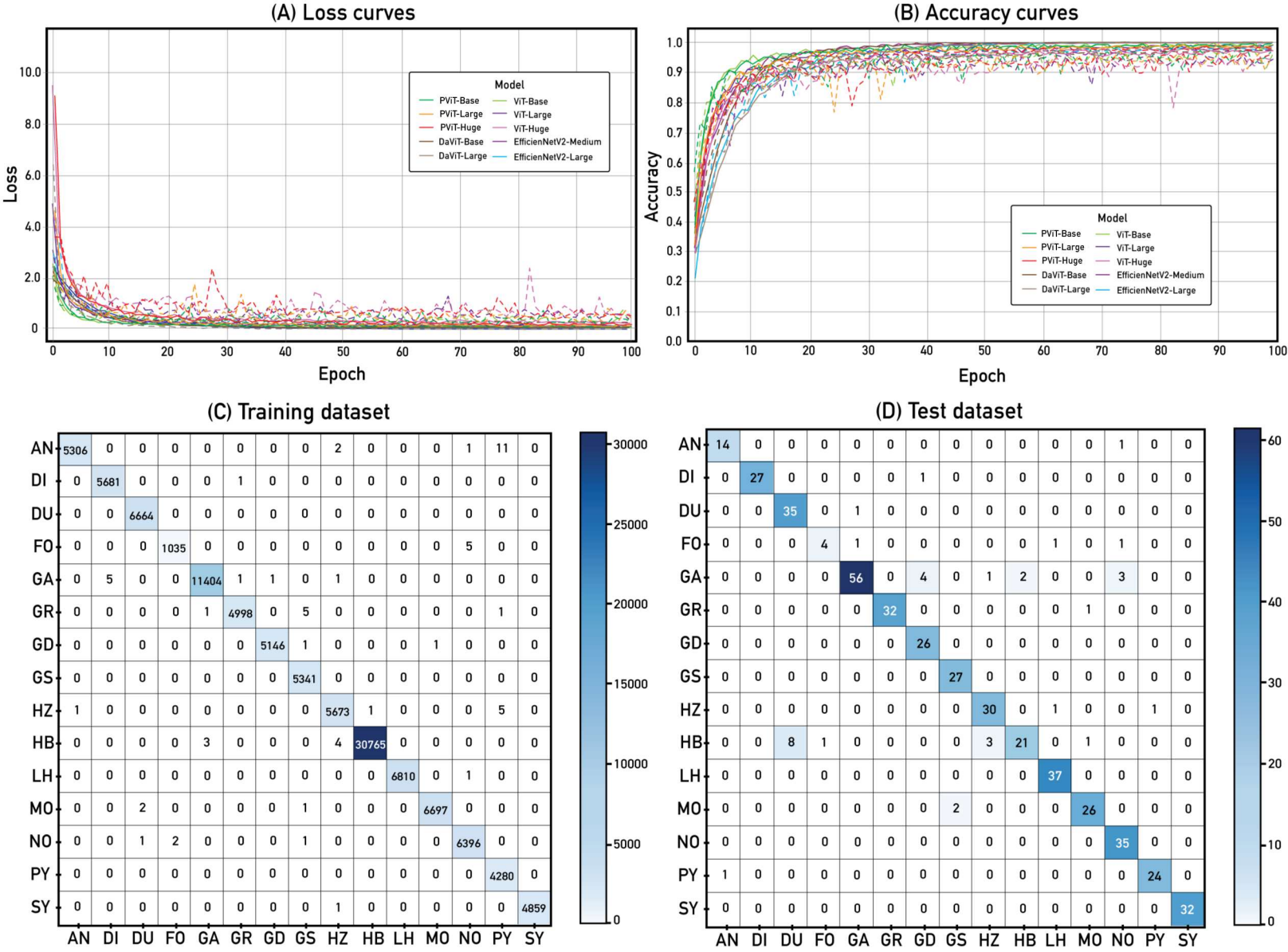

Train (70%)
Validation (15%)
Test (15%)
Names of images
Rock types
0
100
200
300
400
500
AN
DI
DU
FO
GA
GR
GD
GS
HZ
HB
LH
MO
NO
PY
SY
18
34
118
23
43
152
20
40
140
5
7
26
47
90
317
21
40
141
22
42
148
17
31
110
21
41
144
24
44
156
21
41
144
21
42
147
24
44
156
16
32
112
20
38
134

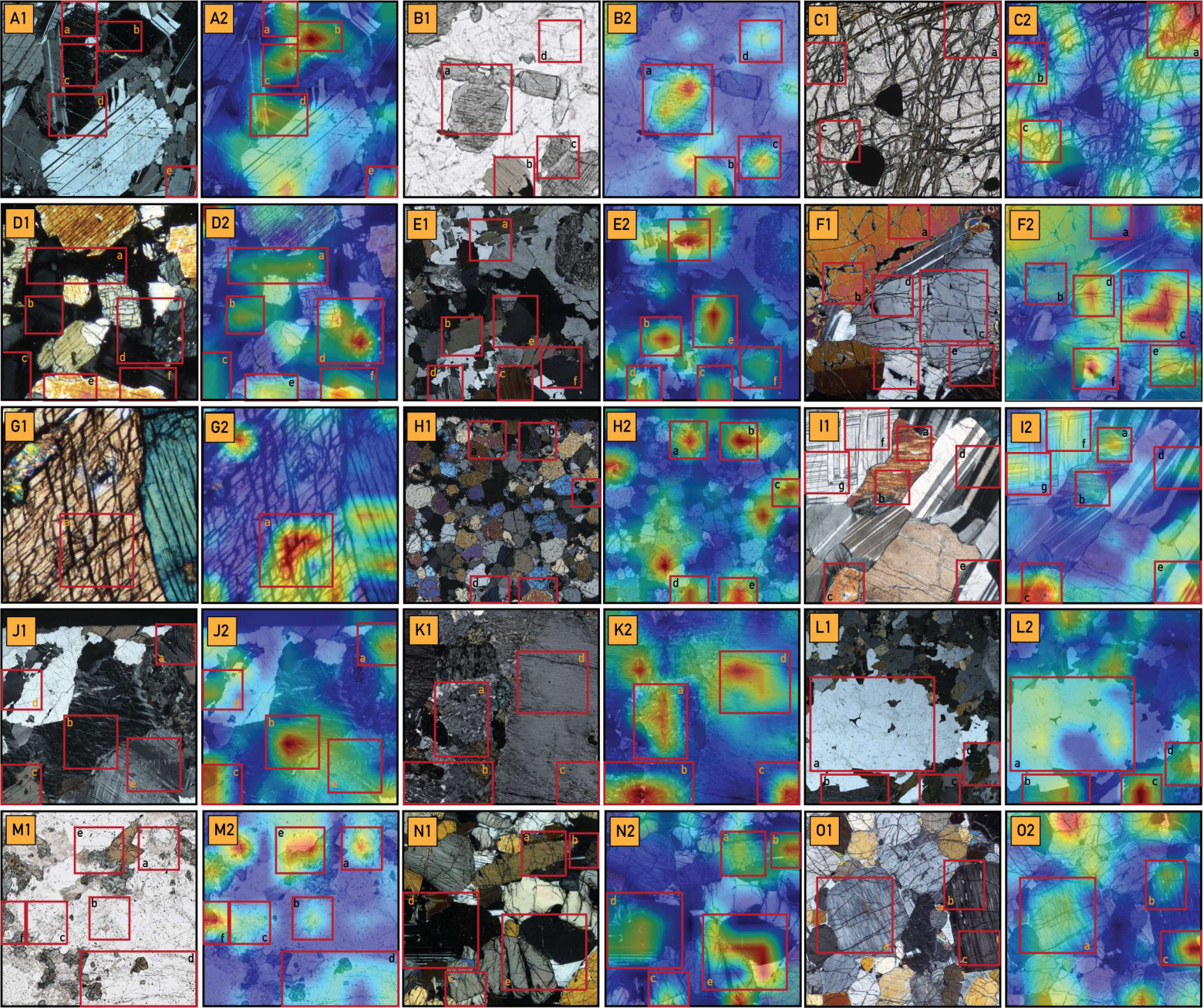

(A) Overall architecture of ViT
(B)
Patch Embedding Layer
(C)
Transformer Encoder Blocks
MLP Head
(D)
AN
DI
DU
FO
GA
GR
GD
GS
HZ
HB
LH
MO
NO
PY
SY
(B) Patch Embedding Layer
ViT Encoder (C)
Sum patch + positional embeddings
[CLS]
1
2
3
4
5
6
7
Add: learnable Absolute Position Embedding (APE)
Patch Embedding (Conv2d P=16, stride=16) + concat [CLS]
(C) Transformer Encoder Blocks
Gradient backward path
Forward path
E^k_{r2}
MLP
Norm
Residual connection
E^k_{r1}
linear
concat.
Self-Attention
SoftMax
A^k_h
scaling
transpose
q^K_h
k^K_h
v^K_h
Linear
Linear
Linear
Residual connection
Norm
k^{th} block
(D) MLP classification head
Gradient backward path
Forward path
MLP Head
[CLS] component
Norm
Final embeddings
ViT Encoder (C)

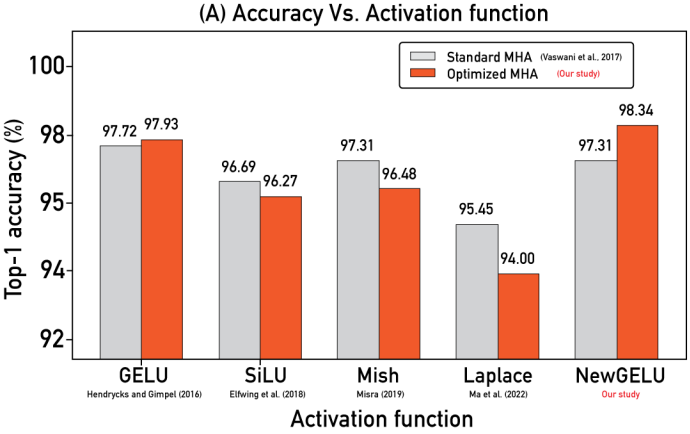

(A) Model architecture
Fully connected layer
newGELU
Fully connected layer
Input Image
(512 x 512 pixel sizes)
Patch +
Positional Embedding
Linear Projection
Patches
Patch Embedding Layer
Residual connection
Layer Norm
Optimized
Multi-Hea Attention
Layer Norm
FFN Block
Residual connection
Transformer Encoder layer
MLP
Head
AN
DI
DU
FO
GA
GR
GD
GS
HZ
HB
LH
MO
NO
PY
SY
(B) Optimized multi-head attention
FFN
Norm
Residual
connection
linear
concat.
Self-Attention
SoftMax
sigmoid
scaling
transpose
Linear Projection
Linear Projectioin
Residual connection
K th block
k th block
1 th block
Gradient backward path
Forward path
Post-processing

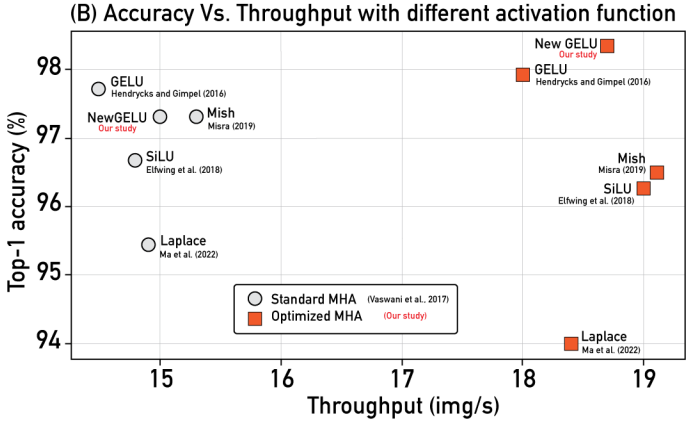

Input
X
W^q
W^k
W^v
B x L
Q
K
V
D
qkv [B×L×3D]
Add Bias
Reshape/Permute/Split
L
N
H
B
SoftMax
(dim=-1)
Q_0
K_0^T
$\sqrt{d_k}$
V_0
Q_1
K_1^T
V_1
H
Z_0
Z_1
Concat heads [B×L×H·d]
W^o
Output

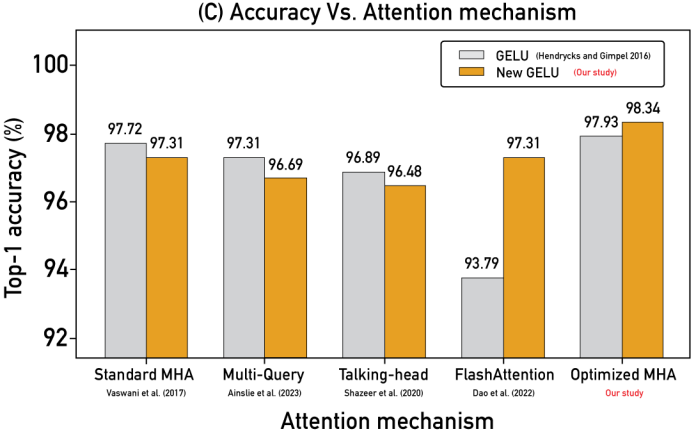

Input image
Patch Embedding Layer
Transformer Encoder layers
(Fig. 5A)
MLP head
(Fig. 5B)
LayerNorm
[CLS] component
AN
DI
DU
FO
GA
GR
GD
GS
HZ
HB
LH
MO
NO
PY
SY
Gradient backward path
Gradients
Post-Processing
Sigmoid
ReLU
A_h^1
A_h^K
F_h^K
Feature maps
Attention visualization heatmaps
High
Attention interpretation
Low
Gradient backward path
Forward path
Post-processing

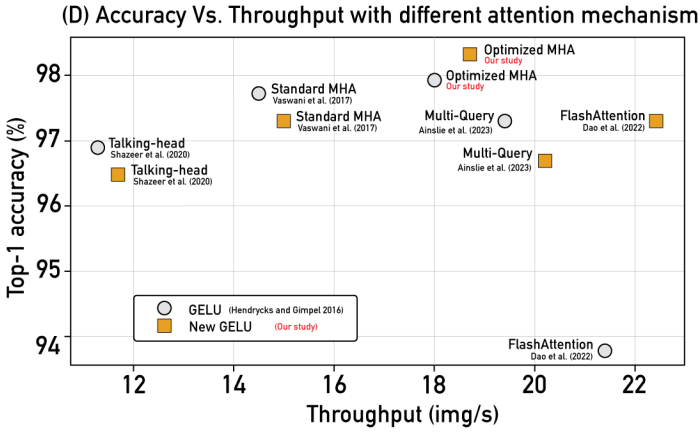

(A) Loss curves
Loss
Epoch
Model
PViT-Base
PViT-Large
PViT-Huge
DaViT-Base
DaViT-Large
ViT-Base
ViT-Large
ViT-Huge
EfficienNetV2-Medium
EfficienNetV2-Large
(B) Accuracy curves
Accuracy
Epoch
(C) Training dataset
AN DI DU FO GA GR GD GS HZ HB LH MO NO PY SY
(D) Test dataset
AN DI DU FO GA GR GD GS HZ HB LH MO NO PY SY

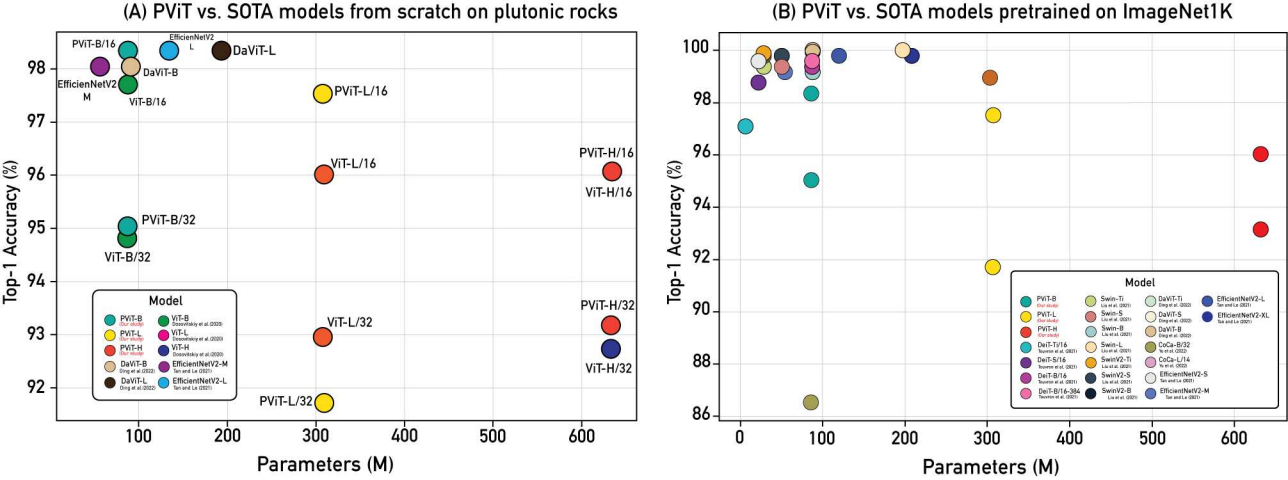

A1
A2
B1
B2
C1
C2
D1
D2
E1
E2
F1
F2
G1
G2
H1
H2
I1
I2
J1
J2
K1
K2
L1
L2
M1
M2
N1
N2
O1
O2


(A) Accuracy Vs. Activation function
Standard MHA (Vaswani et al., 2017)
Optimized MHA (Our study)
Top-1 accuracy (%)
100
98
96
94
92
97.72
97.93
96.69
96.27
97.31
96.48
95.45
94.00
97.31
98.34
GELU
Hendrycks and Gimpel (2016)
SiLU
Elfwing et al. (2018)
Mish
Misra (2019)
Laplace
Ma et al. (2022)
NewGELU
Our study
Activation function


(B) Accuracy Vs. Throughput with different activation function
Top-1 accuracy (%)
98
97
96
95
94
New GELU
Our study
GELU
Hendrycks and Gimpel (2016)
GELU
Hendrycks and Gimpel (2016)
NewGELU
Our study
Mish
Misra (2019)
SiLU
Elfwing et al. (2018)
Mish
Misra (2019)
SiLU
Elfwing et al. (2018)
Laplace
Ma et al. (2022)
Standard MHA (Vaswani et al., 2017)
Optimized MHA (Our study)
Laplace
Ma et al. (2022)
15
16
17
18
19
Throughput (img/s)


(C) Accuracy Vs. Attention mechanism
GELU (Hendrycks and Gimpel 2016)
New GELU (Our study)
Top-1 accuracy (%)
100
98
96
94
92
97.72
97.31
97.31
96.69
96.89
96.48
93.79
97.31
97.93
98.34
Standard MHA
Vaswani et al. (2017)
Multi-Query
Ainslie et al. (2023)
Talking-head
Shazeer et al. (2020)
FlashAttention
Dao et al. (2022)
Optimized MHA
Our study
Attention mechanism


(D) Accuracy Vs. Throughput with different attention mechanism
Top-1 accuracy (%)
98
97
96
95
94
Optimized MHA
Our study
Optimized MHA
Our study
Standard MHA
Vaswani et al. (2017)
Standard MHA
Vaswani et al. (2017)
Multi-Query
Ainslie et al. (2023)
FlashAttention
Dao et al. (2022)
Talking-head
Shazeer et al. (2020)
Talking-head
Shazeer et al. (2020)
Multi-Query
Ainslie et al. (2023)
GELU (Hendrycks and Gimpel 2016)
New GELU (Our study)
FlashAttention
Dao et al. (2022)
12
14
16
18
20
22
Throughput (img/s)


(A) PViT vs. SOTA models from scratch on plutonic rocks
Top-1 Accuracy (%)
Parameters (M)
PViT-B/16
EfficienNetV2 M
DaViT-B
ViT-B/16
EfficienNetV2 L
DaViT-L
PViT-L/16
ViT-L/16
PViT-H/16
ViT-H/16
PViT-B/32
ViT-B/32
ViT-L/32
PViT-H/32
ViT-H/32
PViT-L/32
Model
PViT-B
PViT-L
PViT-H
DaViT-B
DaViT-L
ViT-B
ViT-L
ViT-H
EfficientNetV2-M
EfficientNetV2-L
(B) PViT vs. SOTA models pretrained on ImageNet1K
Top-1 Accuracy (%)
Parameters (M)
Model
PViT-B
PViT-L
PViT-H
DeiT-T/16
DeiT-S/16
DeiT-B/16
DeiT-B/16-384
Swin-Ti
Swin-S
Swin-B
Swin-L
SwinV2-Ti
SwinV2-S
SwinV2-B
DaViT-Ti
DaViT-S
DaViT-B
CoCa-B/32
CoCa-L/14
EfficientNetV2-S
EfficientNetV2-M
EfficientNetV2-L
EfficientNetV2-XL