LeMeViT: Efficient Vision Transformer with Learnable Meta Tokens for Remote Sensing Image Interpretation

Wentao Jiang[1] , **Jing Zhang**[2] , **Di Wang**[1] , **Qiming Zhang**[2] , **Zengmao Wang**[1*] and **Bo Du**[1]
[1]Wuhan University
[2]The University of Sydney
{jiang_wentao, wangzengmao, dubo, d_wang}@whu.edu.cn, jing.zhang1@sydney.edu.au, qzha2506@uni.sydney.edu.au

Abstract

Due to spatial redundancy in remote sensing images, sparse tokens containing rich information are usually involved in self-attention (SA) to reduce the overall token numbers within the calculation, avoiding the high computational cost issue in Vision Transformers. However, such methods usually obtain sparse tokens by hand-crafted or parallel-unfriendly designs, posing a challenge to reach a better balance between efficiency and performance. Different from them, this paper proposes to use learnable meta tokens to formulate sparse tokens, which effectively learn key information meanwhile improving the inference speed. Technically, the meta tokens are first initialized from image tokens via cross-attention. Then, we propose Dual Cross-Attention (DCA) to promote information exchange between image tokens and meta tokens, where they serve as query and key (value) tokens alternatively in a dual-branch structure, significantly reducing the computational complexity compared to self-attention. By employing DCA in the early stages with dense visual tokens, we obtain the hierarchical architecture LeMeViT with various sizes. Experimental results in classification and dense prediction tasks show that LeMeViT has a significant 1.7× speedup, fewer parameters, and competitive performance compared to the baseline models, and achieves a better trade-off between efficiency and performance. The code is released at https://github.com/ViTAE-Transformer/LeMeViT.

1 Introduction

Since the remarkable success of migrating Transformer [Vaswani *et al.*, 2017] from the field of natural language processing to the domain of computer vision, Vision Transformer (ViT) [Dosovitskiy *et al.*, 2020] has sparked significant interest in the community, highlighting great progress and advancements [Carion *et al.*, 2020, Touvron *et al.*, 2021]. Several works [Xu *et al.*, 2021, Zhang *et al.*, 2023] demonstrate

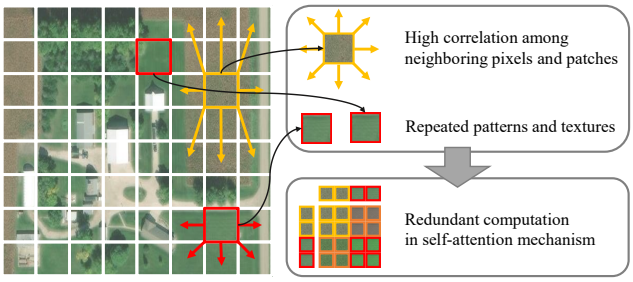


Figure 1: Due to the high correlation between neighboring pixels and image patches, as well as the repetitive nature of textures and patterns in remote sensing images, there is a significant amount of spatial redundancy. This results in redundant computation in self-attention mechanism.

ViT can model long-range dependency within visual information compared to traditional CNN networks with inherent inductive bias, unveiling its revolutionary potential in vision tasks including remote sensing image interpretation [Bazi *et al.*, 2021, Zhang *et al.*, 2021, Wang *et al.*, 2022b].

However, due to the significant spatial redundancy in remote sensing images, ViT suffers from redundant computational overhead, as illustrated in Fig. 1. The self-attention mechanism in ViT computes pairwise affinities between each two image patches regardless of how much useful information the tokens contain. Consequently, the 'background' homogeneous tokens may contribute marginally to the informative feature representations but consume much compute load, hindering the whole model's efficiency.

To address this issue, some works discover the sparse representation of redundant image tokens [Chen *et al.*, 2021c] in the natural image domain. These approaches use a shorter token sequence to represent the original image tokens and replace standard pairwise attention with cross-attention between image tokens and reduced tokens, decreasing the complexity of attention computation. For example, PVT [Wang *et al.*, 2021] obtains reduced tokens through convolutional downsampling, Paca-ViT [Grainger *et al.*, 2023] uses data-driven weight parameters to cluster image tokens, while BiFormer [Zhu *et al.*, 2023] selects a small subset of more informative tokens from coarse to fine level. These methods rely on strong priors that may overlook useful image

*Corresponding Author

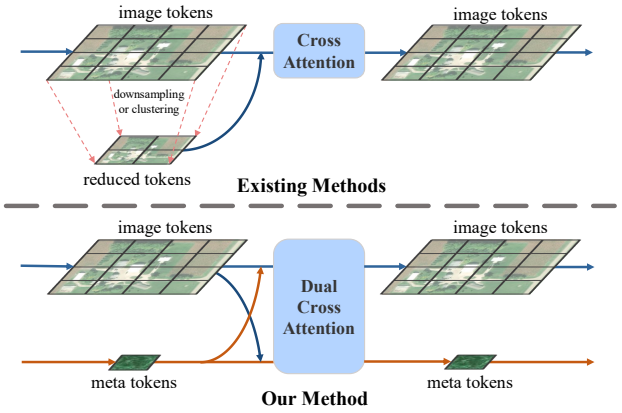


Figure 2: **Existing methods** commonly use downsampling or clustering to reduce the number of image tokens within the current block, which relies on strong priors or is parallel-computation unfriendly. **Our method** learns meta tokens to sparsely represent dense image tokens. Meta tokens exchange information with image tokens via the computationally efficient Dual Cross-Attention Block in an end-to-end way, promoting information flow stage-by-stage.

information, which leaves room for exploring more effective sparse representation. Besides, some methods employ parallel-unfriendly operators like clustering, slowing down computation and increasing memory access.

In this paper, we propose a Vision Transformer with Learnable Meta Tokens called **LeMeViT**, aiming to leverage an extremely small number of learnable meta tokens to represent image tokens. Specifically, the meta tokens are first initialized from image tokens via cross-attention. Then, we propose Dual Cross-Attention (DCA) to promote information exchange between image tokens and meta tokens, where they serve as query and key (value) tokens alternatively in a dual-branch structure, significantly reducing the computational complexity from quadratic to linear compared to self-attention, as illustrated in Fig. 2. By employing DCA in the early stages with dense visual tokens, we obtain the hierarchical architecture LeMeViT with various sizes.

LeMeViT has been intentionally designed to be hardware-friendly. Since modern GPUs excel at parallel computing and coalesced matrix operations, our model exclusively employs simple and dense operators, such as matrix multiplication, standard convolutions, and activation functions. The effective designs of the model architecture and utilization of hardware-friendly operators achieve nearly 1.7× speedup compared to the state-of-the-art (SOTA) model, *e.g.*, ViTAE [Wang *et al.*, 2022a], while also improving performance. Additionally, the model can adapt to sequences of varying lengths, making it transferable to images of various resolutions, which is a common requirement in remote sensing tasks. Experimental results indicate that the model offers competitive performance while enjoying computational efficiency across multiple dense prediction tasks including semantic segmentation, object detection, and change detection.

Our contributions can be summarized as follows:

- We propose a novel Transformer architecture called LeMeViT, which addresses the spatial redundancy in images via efficient architecture designs, achieving a better trade-off between efficiency and performance.
- We propose to learn sparse meta tokens to represent the dense image tokens and promote the information change between meta tokens and image tokens via a novel and computationally efficient DCA module.
- Experiments on both natural images and remote sensing images demonstrate that LeMeViT achieves competitive performance compared to representative baseline models in both classification and dense prediction tasks.

2 Related Work

2.1 Vanilla Vision Transformer

Transformer [Vaswani *et al.*, 2017] quickly dominated the entire field of NLP since its inception and was later introduced into the realm of computer vision, which is known as Vision Transformer (ViT) [Dosovitskiy *et al.*, 2020]. DeiT [Touvron *et al.*, 2021] significantly alleviates the training difficulty of ViT, leading to a proliferation of ViT variants [Srinivas *et al.*, 2021, Tu *et al.*, 2022], establishing a burgeoning and popular domain. ViT brings new vitality to vision tasks by modeling long-range dependencies. However, the significant computational complexity of the vanilla Transformer has remained a substantial challenge in practical usage, stemming in part from the quadratic complexity of its self-attention mechanism. Addressing this challenge become a hot topic of research.

2.2 Efficient Vision Transformer

A considerable amount of work is currently dedicated to reducing the complexity of self-attention, with mainstream approaches including sparse attention and token sparse representation. Sparse attention aims to reduce the connections between tokens, while token sparse representation aims to represent the image using fewer tokens.

One pattern of sparse attention, specifically local attention, has garnered significant interest following the success of the Swin Transformer [Liu *et al.*, 2021], inspiring numerous works [Zhang *et al.*, 2022, Zhang *et al.*, 2024]. Additionally, some works [Liu *et al.*, 2022] explore other forms of sparse attention. For instance, KVT [Wang *et al.*, 2022c] selects only the top-k similar keys for every query in attention. QuadTree Attention [Tang *et al.*, 2021] draws on the method of quadtree segmentation to partition tokens into square blocks, facilitating attention at different granularities.

The earliest work utilizing token sparse representation may be PVT [Wang *et al.*, 2021], which reduces the number of keys and values through convolutional downsampling. CrossViT [Chen *et al.*, 2021a] utilizes image patches of larger size to reduce the number of tokens. Deformable Attention [Xia *et al.*, 2022] employs learnable sampling points to sample tokens from image features.

Some other methods, such as token pruning/merging [Kong *et al.*, 2022], remove or combine certain tokens using score functions. A typical token merging approach like ToMe [Bolya *et al.*, 2022] has shown excellent results when applied to Stable Diffusion [Bolya and Hoffman, 2023].

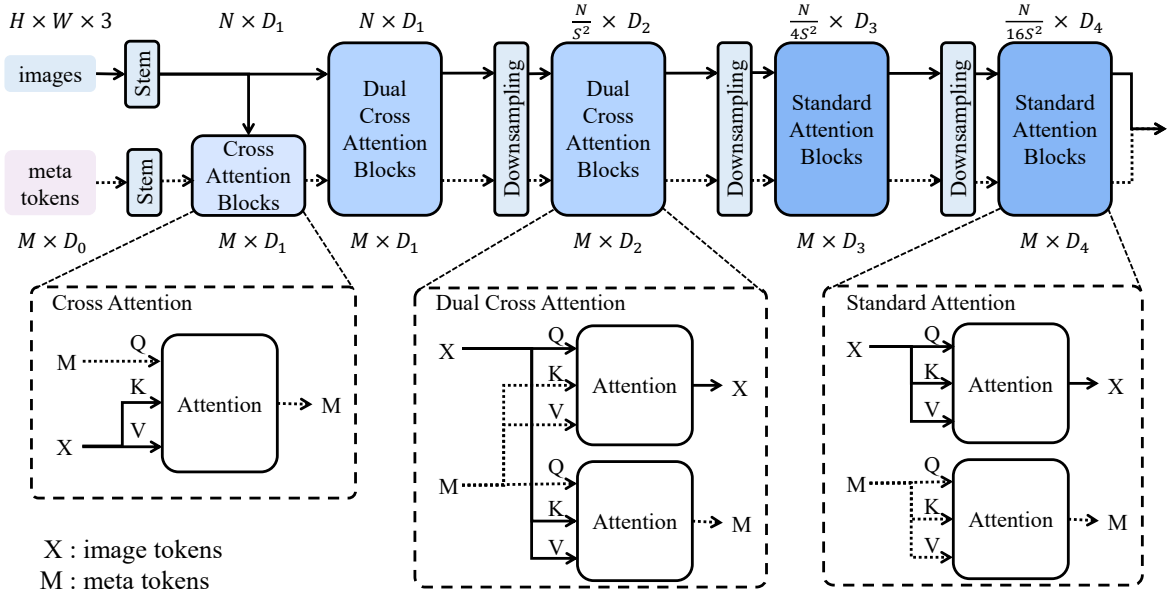


Figure 3: **The Overall Architecture of LeMeViT.** LeMeViT consists of three different attention blocks, arranged from left to right as Cross Attention Block, Dual Cross Attention Block, and Standard Attention Block. Specific details of attention computation method are provided.

2.3 ViT for Remote Sensing

ViT's excellent modeling capability has found wide applications in the field of remote sensing. Numerous endeavors have attempted to incorporate specific characteristics of remote sensing images into ViT, making it more applicable in this domain [Zhang *et al.*, 2021, Wang *et al.*, 2022d, Deng *et al.*, 2021]. These efforts are confined to specific tasks. Recently, RSP [Wang *et al.*, 2022a] adopted a remote sensing pre-training approach to train a foundational model for the remote sensing domain. It achieves state-of-the-art performance across various downstream tasks. However, this ViT-based model also suffers from a serious computational burden, making it an obstacle to practical deployment. Therefore, addressing this issue has become a pressing priority.

3 Method

3.1 Overview and Preliminaries

The overview of the proposed LeMeViT architecture is illustrated in Fig. 3. As depicted, LeMeViT follows a typical hierarchical ViT structure, with four stages connected by downsampling layers. As the stages get deeper, the spatial size of image features gradually reduces while feature dimensionality expands.The core components of the model are Meta Tokens and the DCA block. The meta tokens are first initialized from image tokens via cross-attention. Then, DCA is employed in the early stages to promote information exchange between image tokens and meta tokens, where they serve as query and key (value) tokens alternatively in a dual-branch structure. In the later stages, standard attention blocks based on self-attention are used.

In this paper, image tokens are represented as $\mathcal{X} \in \mathbb{R}^{N \times C}$ and meta tokens are represented as $\mathcal{M} \in \mathbb{R}^{M \times C}$, where C represents token dimension, N and M represents the number of image tokens and meta tokens, respectively. N varies across different stages, with $N \gg M$ in the early stages. Additionally, we employ D_1, D_2, D_3, and D_4 to signify the token dimensions across different stages.

Due to the extensive use of scaled dot-product attention in our model, we provide its formal definition here:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \qquad (1)$$

where queries $Q \in \mathbb{R}^{N_1 \times C}$, keys and values $K, V \in \mathbb{R}^{N_2 \times C}$. Scalar factor $\sqrt{d_k}$ is introduced to avoid gradient vanishing, where d_k is normally assigned token dimension C. For maintaining entropy-invariance, we use $\frac{\log N_1}{\log N_2}\sqrt{C}$ as scalar factor in cross-attention instead [Chiang and Cholak, 2022].

3.2 Key Components in LeMeViT

Fig. 3 shows all the components of LeMeViT, including learnable meta tokens, stem blocks, downsampling layers, and three types of attention blocks. Subsequently, we will introduce the detailed structures.

Learnable Meta Tokens. Meta tokens are a set of learnable tensors updated via information exchange with image tokens. They can be analogized to learnable queries in DETR [Carion *et al.*, 2020], although their learning method and function differ. After training, initial meta tokens are fixed, but they continue to update by interacting with image tokens. They serve as the model input alongside the image tokens. Initially, the shape of meta tokens is $M \times D_0$. Their dimensions expand as image tokens, but their length remains M. Based on empirical analysis, we set M to 16 in our experiments, as validated in the ablation studies (Sec. 4.4).

Stem and Downsampling Layers. The Stem block divides the input image into patches and embeds them into tokens. We employ an overlapping patch embedding technique. Specifically, we implement this block using two 3×3 convolutions with a stride of 2 and padding of 1. The convolutional windows slide in an overlapping manner, and after two layers, the image is precisely divided into tokens, each of which corresponds to a patch of size 4×4. To align with the dimensions of image tokens, we introduce an extra Stem block for meta tokens, which consists of two MLP layers. After passing through the Stem block, the image is transformed into N tokens, and we have $N = \frac{H}{4} \times \frac{W}{4}$. Both the image tokens and meta tokens share the feature dimension D_1. The downsampling layer similarly adopts an overlapping patch embedding approach but employs only one convolution to achieve a downsampling ratio of 2.

Then, three distinct attention blocks are used in LeMeViT, including the Cross Attention (CA) block, Dual Cross Attention (DCA) block, and Standard Attention (SA) block. They share similar structures, involving Conditional Positional Encodings (CPE), LayerNorm (LN), Attention, Feed Forward Network (FFN), and residual connections. Their only difference lies in the Attention layer. Notably, meta tokens and image tokens share the same FFN for parameter efficiency.

Cross Attention Block. CA block is employed to learn meta tokens from image tokens. Due to the considerable gap between the initial meta tokens and image tokens, directly using meta tokens as keys and values to update the image tokens might lead to the collapse of image features and information loss. Therefore, CA is designed to only update meta tokens. It employs a cross-attention mechanism, where query is the projection of meta tokens, and key and value are projections of image tokens. Its formulation can be described as follows:

$$\mathcal{M} \Leftarrow \text{Attention}(\mathcal{M}_Q, \mathcal{X}_K, \mathcal{X}_V), \tag{2}$$

where $\mathcal{M}_Q$ denotes query projection of meta tokens and $\mathcal{X}_K, \mathcal{X}_V$ denote key and value projections of image tokens.

Dual Cross-Attention Block. DCA block is the core component for enhancing computational efficiency. It replaces the pairwise self-attention among image tokens with two cross-attention between image tokens and meta tokens, reducing the computational complexity from quadratic $\mathcal{O}(N^2)$ to linear $\mathcal{O}(2MN)$. Considering that $M \ll N$, the efficiency improvement is notably evident. Meanwhile, it retains strong representation capabilities. Unlike other cross-attention strategies [Wang *et al.*, 2022e, Zhu *et al.*, 2023] that reduce image tokens explicitly, DCA implicitly preserves most of the image information from all stages through meta tokens. Within the DCA block, image tokens fuse global information held by meta tokens via cross-attention while aggregating local information of each patch into meta tokens via another cross-attention. Specifically, image tokens and meta tokens serve as each other's query and key/value, which can be formulated as:

$$\mathcal{X} \Leftarrow \text{Attention}(\mathcal{X}_Q, \mathcal{M}_K, \mathcal{M}_V). \tag{3}$$

$$\mathcal{M} \Leftarrow \text{Attention}(\mathcal{M}_Q, \mathcal{X}_K, \mathcal{X}_V), \tag{4}$$

Standard Attention Block. In the last two stages, we adopted the standard attention mechanism for the trade-off between efficiency and performance. In hierarchical ViTs, the number of image tokens decreases as the stages deepen. Consequently, the assumption of $M \ll N$ may no longer hold in the final two stages. Additionally, due to the increased dimensions, computational and parameter overhead caused by projection layers becomes substantial. Therefore, DCA may not be as efficient as in the first two stages, making standard self-attention a preferable choice. Specifically, image tokens and meta tokens perform self-attention individually.

Finally, image tokens and meta tokens are processed by global average pooling separately and then being added together for the classification prediction. Additionally, only image tokens from each stage, which have different scales, are used to perform dense prediction tasks.

3.3 Details of Architecture Design

Based on the overall architecture, we devise three model variants of different sizes, *i.e.*, Tiny, Small, and Base. We tailored these sizes by adjusting the number of blocks and dimensions of features in each stage, as listed in Table 1. Other configurations are shared between all variants. We set each head dimension of attention to 32, the MLP expansion rate to 4, and the conditional positional encoding kernel size to 3. The length of meta tokens is set to 16.

Version	Blocks					Dims			
	S_0	S_1	S_2	S_3	S_4	D_1	D_2	D_3	D_4
LeMeViT-Tiny	1	2	2	8	2	64	128	192	320
LeMeViT-Small	1	2	2	6	2	96	192	320	384
LeMeViT-Base	2	4	4	18	4	96	192	384	512

Table 1: **Architecture details of different LeMeViT variants.** S_0 to S_4 represent the number of blocks in the CA stage (S_0), two DCA stages (S_1 and S_2), and two SA stages (S_3 and S_4). D_1 to D_4 signify the dimensions of features in each stage, as shown in Fig. 3.

3.4 Computational Complexity Analysis

We primarily analyze the computational complexity of DCA, the core module responsible for LeMeViT's efficiency improvement. We compute the complexity across the projection layer, attention layer, and FFN layer, forming an entire DCA block. We assume that the shape of image tokens is $N \times D$, the shape of meta tokens is $M \times D$, and the MLP expansion rate in FFN is E. The results are shown in Table 2.

	Dual Cross-Attention	Standard Attention
Projection	$4ND^2 + 4MD^2$	$4ND^2$
Attention	$2NMD$	$2N^2D$
FFN	$2E(N+M)D^2$	$2E(N+M)D^2$

Standard Attention	$(2E+4)ND^2 + 2N^2D$
Dual Cross-Attention	$(2E+4)(N+M)D^2 + 2NMD$

Table 2: **Above:** Computation complexity of specific layers in DCA and standard attention. **Below:** Total computational complexity of the two attention blocks.

Model	Infer ↑ (img/sec)	Params ↓ (M)	MACs ↓ (G)	Acc@1 ↑ (%)
PVTv2-b1	4897.43	14.01	2.03	78.70
MobileViTv2	5162.10	4.90	1.41	78.10
Efficientformerv2	1617.41	6.19	0.63	79.00
LeMeViT-tiny	**5316.58**	8.64	1.78	**79.07**
Swin-tiny	2872.48	28.29	4.35	81.30
PVTv2-b2	2866.77	25.36	3.88	82.00
FLatten-Swin-T	1918.84	28.50	4.39	**82.10**
FLatten-PVT-S	1863.83	24.72	3.70	81.70
PacaViT-tiny	2157.28	12.20	3.10	80.63
BiFormer-tiny	2889.70	13.14	2.20	81.40
LeMeViT-small	**3608.12**	16.40	3.74	81.88
Swin-small	1717.31	49.61	8.51	83.00
Swin-base	1215.39	87.77	15.13	83.30
PVTv2-b4	1494.79	62.56	9.79	83.60
CrossViT-base	**1911.69**	105.03	20.10	82.20
PacaViT-base	927.16	46.91	9.26	83.96
BiFormer-base	799.07	56.80	9.32	84.30
LeMeViT-base	1482.70	53.10	11.06	**84.35**

Table 3: Comparison of different models on ImageNet-1K.

Based on the results in Table 2, we can draw the following conclusion. The computational complexity of DCA is linear regarding N, which is significantly lower than the quadratic complexity of the standard attention. Specifically, for our three model variants, the computational complexity is reduced by about $10\times$ compared to using standard attention, *e.g.*, given an image size of 224, the first DCA block in the tiny/small/base model has 0.16/0.36/0.36 GFLOPs while the standard attention has 1.41/2.24/2.24 GFLOPs. Experiments in Supplementary Material demonstrate that DCA achieve notably higher inference speed than SA.

4 Experiments

To assess the efficiency and performance of our model, we first evaluate it on the ImageNet-1K dataset for image classification, comparing it against other efficient ViTs in Sec. 4.1. Then, we conduct a series of experiments in remote sensing tasks compared to the representative Swin Transformer [Liu *et al.*, 2021] and SOTA ViTAE [Xu *et al.*, 2021]. Specifically, we pre-train the model on the MillionAID [Long *et al.*, 2021] dataset (Sec. 4.2) and then transfer it to downstream tasks including object detection, semantic segmentation, and change detection (Sec. 4.3). Additionally, we conduct an ablation study to validate the setting of the length of meta tokens in Sec. 4.4. Finally, we visualize and analyze the attention map in Sec. 4.5.

4.1 Image Classification on ImageNet-1K

We first conduct image classification experiments on the ImageNet-1K benchmark. We train the three variants of our model and compare them with other efficient ViTs of different sizes. These representative methods include Swin Transformer [Liu *et al.*, 2021], PVTv2 [Wang *et al.*, 2022e], CrossViT [Chen *et al.*, 2021a], MobileViTv2 [Mehta and Rastegari, 2021], Efficientformerv2 [Li *et al.*, 2023], Flatten Attention [Han *et al.*, 2023], PacaViT [Grainger *et al.*, 2023], and BiFormer [Zhu *et al.*, 2023]. For a fair comparison, all our three models are trained using the same settings following DeiT, as mentioned below.

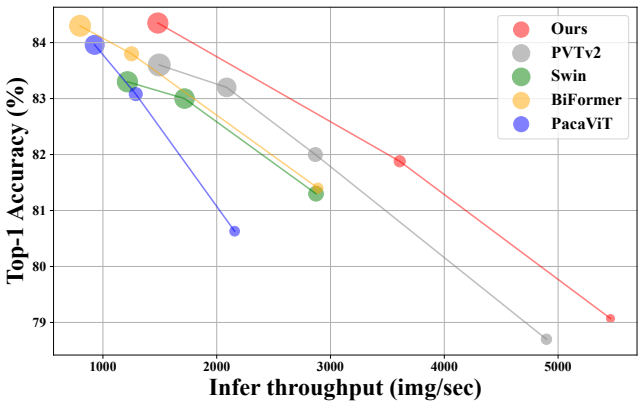


Figure 4: Visualization of comparison between different models. The size of scatter represents the parameter count of the model.

Metrics. We employ efficiency and classification performance metrics for evaluation, where efficiency metrics include throughput, parameter count, and MACs, while top-1 accuracy serves as the classification performance metric. However, MACs may have a low correlation with actual inference latency, due to the fact that MACs can not reflect memory access efficiency. Therefore, we primarily use throughput as the efficiency metric.

Experiment Details. The models are implemented by PyTorch and trained 300 epochs from scratch. We employ AdamW as the optimizer, and apply a cosine decay learning rate schedule. The complete hyper-parameters for learning strategy and data augmentation techniques are provided in the Supplementary Material. We apply stochastic depth with 0.15 probability for our models. We set different batch sizes, *i.e.*, 400, 256, and 64, for Tiny, Small, and Base model variants, respectively. We train the models on four Nvidia RTX 4090 GPUs and test on one. We report the top-1 accuracy on the ImageNet-1K validation set. For a fair comparison, we use the same TIMM benchmark tool to test the throughput, parameter count, and MACs of our models and different models in the same environment.

Results. Quantitative results are listed in Table 3, and plotted in Fig. 4. The results show that our LeMeViT achieves the best trade-off in efficiency and performance among all the comparison methods, which can be easily observed from Fig. 4. For quantitative analysis, compared to the most competitive PVTv2 [Wang *et al.*, 2022e], LeMeViT-Base obtains 0.75% better accuracy than PVTv2-b4, with similar throughput and fewer parameters. Compared to BiFormer-Base [Zhu *et al.*, 2023], LeMeViT-Base achieves $1.85\times$ speedup and slight accuracy increase. Compared to Swin Transformer and Paca-ViT, our model outperforms them both in throughput and accuracy. Taking into account the trade-off between accuracy, throughput, and parameter count, our model gener-

Model	Throughputs (img/sec) ↑		Memory Usage (GB) ↓		Params ↓	MACs ↓	Acc@1 ↑	Acc@5 ↑
	Train	Infer	Train	Infer	(M)	(G)	(%)	(%)
Swin-tiny	872.85	2874.62	6.14	1.87	27.56	4.35	98.42 (98.59)	99.87 (99.88)
ViTAEv2-small	624.09	1847.65	8.44	1.70	18.87	5.48	98.97	99.88
LeMeViT-tiny	1389.33	5327.47	3.91	1.36	8.33	1.78	98.80	99.82
LeMeViT-small	968.59	3612.68	5.33	1.69	16.04	3.74	99.00	**99.90**
LeMeViT-base	408.92	1484.09	11.08	1.91	52.61	11.06	**99.17**	99.88

Table 4: Results of Scene Recognition on MillionAID. The results in the parentheses denote the results of Swin-tiny trained for 300 epochs, while the others are based on models trained for 100 epochs.

ally outperforms other efficient ViT models.

4.2 Remote Sensing Scene Recognition

After evaluating LeMeViT on natural image classification, we apply the model to the remote sensing domain. Following the remote sensing model RSP [Wang *et al.*, 2022a], we first pre-train our model on the MillionAID dataset for aerial scene recognition. Most experimental details remain consistent with the classification task, with the only difference being the epoch setting, which is adjusted to 100 to ensure a fair comparison with RSP. We evaluate the model in the same setting, which is detailed in the Supplementary Material.

Datasets. MillionAID is a large-scale dataset in the remote sensing (RS) domain, comprising 1,000,848 non-overlapping scenes. Notably, MillionAID is RGB-based, making it more compatible with existing deep models developed in the natural image domain. The MillionAID dataset comprises 51 categories organized in a hierarchical tree structure, with 51 leaves distributed across 28 parent nodes at the second level. The images vary in size from 110 $\times$ 110 to 31,672 $\times$ 31,672. We utilize the same dataset split as RSP, which randomly chooses 1,000 images in each category to form the validation set of 51,000 images, with the remaining 949,848 images used for training.

Results. Results for MillionAID aerial scene recognition is summarized in Table 4. We mainly compare LeMeViT-Small with the representative Swin Transformer and the SOTA ViTAE. Compared to Swin-Tiny, LeMeViT-Small achieves 1.25 $\times$ faster inference speed, and its accuracy trained for 100 epochs surpasses Swin-Tiny trained for 300 epochs by 0.41%. In comparison to ViTAE, our model achieves nearly a 1.96$\times$ speedup in inference, a 1.47 $\times$ speedup in training, and even a slight improvement in accuracy by 0.03%. Other LeMeViT variants also show competitive efficiency and performance.

4.3 Remote Sensing Downstream Tasks

To validate the transferability of LeMeViT to dense prediction tasks, we apply the pre-trained models from Sec. 4.2 to various remote sensing downstream tasks, including object detection, semantic segmentation, and change detection. Following RSP, we fine-tune the models using appropriate methods and settings for each task. Specific training recipes are provided in Supplementary Material.

Object Detection. Aerial object detection involves detecting oriented bounding boxes (OBB) instead of the typical horizontal bounding boxes (HBB) used in conventional natural image tasks. We conduct aerial object detection experiments using the DOTA dataset [Xia *et al.*, 2018], which is the most famous large-scale dataset for OBB detection. It contains 2,806 images with sizes ranging from 800 $\times$ 800 to 4,000 $\times$ 4,000, encompassing 188,282 instances across 15 categories. We use Oriented-RCNN [Xie *et al.*, 2021] as the OBB detection head. We train models on the merged training set and validation set of DOTA datasets and evaluate them on the testing set, which is only accessible on the evaluation server. We use the mean average precision (mAP) of all categories as the metric.

Semantic Segmentation. Aerial semantic segmentation refers to pixel-level classification of the aerial scene. We use the ISPRS Potsdam[1] dataset as the benchmark for this task. This dataset has 38 images, each with an average size of 6,000 $\times$ 6,000 pixels. These images are cropped into 512 $\times$ 512 patches with a stride of 384, and they contain six categories: impervious surface, building, low vegetation, tree, car, and clutter. The dataset is divided into training and testing sets, with 24 and 14 images respectively. We use UperNet [Xiao *et al.*, 2018] as the segmentation framework. The overall accuracy (OA) and mean F1 score (mF1) are used for evaluation.

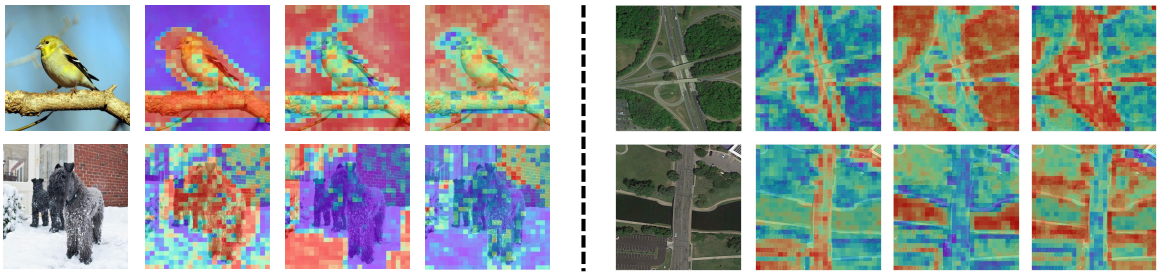
Change Detection. Aerial change detection involves binary classification to pixel-wise label the dissimilarities between two images captured in the same scene at different timestamps. We use the pre-processed CDD dataset [Lebedev *et al.*, 2018] to evaluate models on this task. The image pairs are cropped into a sequence of 256 $\times$ 256 patches, and the sizes of the training, validation, and testing sets are 10,000/3,000/3,000, respectively. We adopt the BIT [Chen *et al.*, 2021b] framework for change detection. We report the mean F1 score (mF1) on the testing set.

Results. The results of LeMeViT-Small and other models are presented in Table 5. Since the inference speed depends on many factors, such as the number of instances in an image for object detection, we mainly report MACs as the efficiency metric. Further details and results of other models can be found in the Supplementary Material. In the object detection task, LeMeViT-Small exhibits only a 0.14% mAP loss but has 20% less computations compared to ViTAEv2-Small, and outperforms Swin-Tiny in both detection accuracy and computational efficiency. For semantic segmentation, there

[1]https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

Backbone	Object Detection		Semantic Segmentation			Change Detection	
	mAP ↑	MACs ↓	OA ↑	mF1 ↑	MACs ↓	mF1 ↑	MACs ↓
Swin-tiny	76.50	215.68	90.78	90.03	234.79	95.21	15.63
ViTAE-v2-small	77.72	234.82	91.21	90.64	238.28	96.81	15.94
LeMeViT-tiny	76.63	**154.12**	91.03	90.55	**217.88**	95.56	**5.75**
LeMeViT-small	77.58	193.91	91.23	90.62	228.16	96.64	10.71
LeMeViT-base	**78.00**	335.53	**91.35**	**90.85**	263.75	**97.32**	28.47

Table 5: Comparison on three downstream tasks. More details are provided in Supplementary Material.



(a) Natural images (b) Remote sensing images

Figure 5: Visualization of the attention maps between three meta tokens in the last layer and image tokens. **(a)** illustrates the attention maps on natural images, while **(b)** illustrates attention maps on remote sensing images.

is almost no difference in OA and mF1 between LeMeViT and ViTAE models. In change detection, the mF1 score of our model is inferior to ViTAEv2-Small but comparable with Swin-Tiny, but they require almost 50% more computations.

4.4 Ablation Study

In the ablation studies, we primarily investigate the impact of the length of meta tokens. The results are shown in the Table 6. An interesting conclusion can be drawn that the length of meta tokens has a marginal impact on the performance. With lengths of 64, 32, 16, and 8, the accuracy is almost the same. This further confirms the redundancy in images and the vanilla attention calculation, suggesting the motivation of using a smaller number of meta tokens to represent the dense image tokens. Finally, considering both efficiency and accuracy, we choose 16 as the default setting of meta token length.

More ablation studies are conducted to validate the effectiveness of components, *i.e.*, cross attention block, meta token stem and token fusion method in the final layer. Results provided in Supplementary Material demonstrate these designs increase the accuracy.

Token Length	**Inference Throughputs ↑**	**MACs ↓**	**MillionAID Acc@1 ↑**
64	3268.48	4.39	98.96
32	3481.31	3.95	98.97
16	3609.64	3.74	**99.00**
8	3639.84	3.63	98.96

Table 6: Results of different settings of meta token length in LeMeViT-small.

4.5 Visualization

To gain a deeper understanding of how meta tokens work, we visualize the cross-attention maps between meta tokens and image tokens in the last block of the DCA, as illustrated in Fig. 5. We visualize both natural images and remote sensing images. The cross-attention in natural images (Fig. 5(a)) reveals that the learned meta tokens can well attend to semantic parts of images, *i.e.*, foreground objects, leading to a better object representation and effective information exchange with image tokens, which contribute to the improved classification accuracy. Fig. 5(b) shows the cross-attention for different meta tokens, providing a clear indication that different meta tokens are responsible for different semantic parts of images, *e.g.*, roads, grasslands, and forests. The results imply that the meta tokens can learn effective representations by aggregating important semantic regions in images. The visualization offers a clear way to explain how meta tokens function, enhancing the interpretability of LeMeViT.

5 Conclusion

This paper introduces LeMeViT, a novel Vision Transformer architecture designed to efficiently address computational bottlenecks in traditional attention layers. Inspired by the spatial redundancy in images, particularly in remote sensing images, we suggest learning meta tokens to represent dense image tokens. To enhance computational efficiency, we replace the original self-attention mechanism with a Dual Cross-Attention, promoting information exchange between meta tokens and image tokens. LeMeViT is versatile, supporting image classification, scene recognition and diverse dense prediction tasks. Experimental results on different public benchmarks show LeMeViT's superior balance between efficiency and performance.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62225113, 62271357, the National Key Research and Development Program of China 2023YFC2705700, the Natural Science Foundation of Hubei Province under Grants 2023BAB072, and the Fundamental Research Funds for the Central Universities under Grants 2042023kf0134.

References

[Bazi *et al.*, 2021] Yakoub Bazi, Laila Bashmal, Mohamad M Al Rahhal, Reham Al Dayil, and Naif Al Ajlan. Vision transformers for remote sensing image classification. *Remote Sensing*, 13(3):516, 2021.

[Bolya and Hoffman, 2023] Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4598–4602, 2023.

[Bolya *et al.*, 2022] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations*, 2022.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[Chen *et al.*, 2021a] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

[Chen *et al.*, 2021b] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021.

[Chen *et al.*, 2021c] Tianlong Chen, Yu Cheng, Zhe Gan, Lu Yuan, Lei Zhang, and Zhangyang Wang. Chasing sparsity in vision transformers: An end-to-end exploration. *Advances in Neural Information Processing Systems*, 34:19974–19988, 2021.

[Chiang and Cholak, 2022] David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7654–7664, 2022.

[Deng *et al.*, 2021] Peifang Deng, Kejie Xu, and Hong Huang. When cnns meet vision transformer: A joint framework for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[Grainger *et al.*, 2023] Ryan Grainger, Thomas Paniagua, Xi Song, Naresh Cuntoor, Mun Wai Lee, and Tianfu Wu. Paca-vit: learning patch-to-cluster attention in vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18568–18578, 2023.

[Han *et al.*, 2023] Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5961–5971, 2023.

[Kong *et al.*, 2022] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, et al. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *European conference on computer vision*, pages 620–640. Springer, 2022.

[Lebedev *et al.*, 2018] MA Lebedev, Yu V Vizilter, OV Vygolov, Vladimir A Knyaz, and A Yu Rubis. Change detection in remote sensing images using conditional adversarial networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:565–571, 2018.

[Li *et al.*, 2023] Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Rethinking vision transformers for mobilenet size and speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16889–16900, 2023.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Liu *et al.*, 2022] Liu Liu, Zheng Qu, Zhaodong Chen, Fengbin Tu, Yufei Ding, and Yuan Xie. Dynamic sparse attention for scalable transformer acceleration. *IEEE Transactions on Computers*, 71(12):3165–3178, 2022.

[Long *et al.*, 2021] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 14:4205–4230, 2021.

[Mehta and Rastegari, 2021] Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2021.

[Srinivas *et al.*, 2021] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish

Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16519–16529, 2021.

[Tang *et al.*, 2021] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. In *International Conference on Learning Representations*, 2021.

[Touvron *et al.*, 2021] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[Tu *et al.*, 2022] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2021] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.

[Wang *et al.*, 2022a] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[Wang *et al.*, 2022b] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2022.

[Wang *et al.*, 2022c] Pichao Wang, Xue Wang, Fan Wang, Ming Lin, Shuning Chang, Hao Li, and Rong Jin. Kvt: k-nn attention for boosting vision transformers. In *European conference on computer vision*, pages 285–302. Springer, 2022.

[Wang *et al.*, 2022d] Wei Wang, Chen Tang, Xin Wang, and Bin Zheng. A vit-based multiscale feature fusion approach for remote sensing image segmentation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[Wang *et al.*, 2022e] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[Xia *et al.*, 2018] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[Xia *et al.*, 2022] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4794–4803, 2022.

[Xiao *et al.*, 2018] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[Xie *et al.*, 2021] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3520–3529, 2021.

[Xu *et al.*, 2021] Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Advances in neural information processing systems*, 34:28522–28535, 2021.

[Zhang *et al.*, 2021] Jianrong Zhang, Hongwei Zhao, and Jiao Li. Trs: Transformers for remote sensing scene classification. *Remote Sensing*, 13(20):4143, 2021.

[Zhang *et al.*, 2022] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vsa: Learning varied-size window attention in vision transformers. In *European conference on computer vision*, pages 466–483. Springer, 2022.

[Zhang *et al.*, 2023] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *International Journal of Computer Vision*, 131(5):1141–1162, 2023.

[Zhang *et al.*, 2024] Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. Vision transformer with quadrangle attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Zhu *et al.*, 2023] Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson WH Lau. Biformer: Vision transformer with bi-level routing attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10323–10333, 2023.