Vol. 06, No. 02, pp. 384 –392 (2025)
ISSN: 2708-0757

JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS

www.jastt.org

Window-Based Vision Transformer Network Implementation For Multi-Label Image Classification In Remote Sensing

Emre Akkaş*, Selda Güney

Department of Electrical Electronics Engineering, Başkent University, Ankara, Türkiye, emreeakkass@gmail.com, seldakul@gmail.com

**Correspondence: emreeakkass@gmail.com*

Abstract

Swift process in technology and widespread availability of low-cost internet have led to a substantial rise in data volume in remote sensing, especially for high-resolution and very-high resolution images. Still, these images contain more complex information, and it is not appropriate to analyze the images using a solitary scene-level label while ignoring the distinct features provided by other labels in the images. In multi-label image classification applications, multiple labels are assigned to an image, reflecting various objects or features present in the scene. The classification of these images is critically important for monitoring environmental changes over large geographical areas, disaster management, urban planning, agriculture and forestry management, natural resource conservation, and military intelligence. Nowadays, many methods are used in such image classification problems, primarily deep learning algorithms. However, current deep learning approaches for multi-label remote sensing images often struggle to capture both local fine-grained details and global contextual relationships simultaneously, leaving a gap for models that can efficiently integrate these complementary representations. In this study, advanced neural networks are explored and evaluated for Multi-label AID dataset which contains 3000 images and 17 different labels; AlexNet, VGG16, DenseNet-201, Inception-v3 and ConvNeXt as the CNN models, ViT, SwinT as transformer models and MaxViT as the hybrid model that initially contains both CNN and transformer network. OneCycleLR as scheduler and AsymmetricLoss (ASL) as loss function are employed for each model to systematically evaluate their impact on model performance. MaxViT was chosen because its multi-scale window-based attention can jointly model local and global dependencies, making it particularly suitable for the complex spatial patterns in remote-sensing imagery compared with other hybrid architectures. The window-based MaxViT algorithm, which has not been previously applied to the Multi-label AID dataset in the current literature, has been evaluated. This constitutes the first application of MaxViT to this dataset and provides a novel benchmark for multi-label remote-sensing classification. This algorithm has demonstrated superior performance on this dataset, significantly outperforming existing models and setting a new benchmark with an mAP of 84.98%.

Keywords: *Remote Sensing, Scene Classification, MaxViT, Multi-label AID, OneCycleLR*
Received: August 11th, 2025 / Revised: October 12th, 2025 / Accepted: October 26th, 2025 / Online: November 06th, 2025

I. INTRODUCTION

The term "remote sensing" was initially originated by the United States Naval Research Officer, Ms. Evelyn Pruitt, during the 1950s [1]. In contemporary usage, it commonly refers to the scientific and artistic practice of identifying, observing, and quantifying an object without direct physical interaction. It is highly significant across multiple domains such as urban planning [2], forestry [3], geospatial analysis [4], ecological conservation of mountain grasslands [5] etc. to gather valuable information about the Earth's features, conditions, and changes over time.

Huge amount of imagery shall be analysed to extract meaningful information from the aforementioned domains; the well-known method for this is called as image classification. For the purpose of evaluating and analysing the remote sensing images, different datasets are needed. There are plenty of multi-class datasets available online such as EuroSAT [6], RSSCN7 [7], UC Merced (UCM) [8], AID [9] and so on. Convolutional Neural Networks (CNNs) have been dominant approach for image understanding tasks, based on their superior performance on classification problems and this success has extended to many other image understanding tasks. ImageNet [10] dataset played a vital role in their success due to availability of a large training set. The evolution of the cutting-edge on the ImageNet

doi: 10.38094/jastt62393

dataset demonstrates the advancements with CNN architectures and learning [11], [12]. A rising focus has emerged on architectures employing attention mechanisms with convolution networks [13]. Several attempts have been made to use transformers on image classification, but the performance was not as successful as convnets. Nonetheless, hybrid architectures which combine transformers and convnets, including the self-attention mechanism, exhibited notable results in image classification.

Vision Transformers (ViT) [14] have achieved SOTA (state-of-the-art) results on ImageNet without the use of convolution. After ViT [14] and Swin Transformer (SwinT) [15] are published and many studies have been performed with these vision transformers, Kaselimi et al. [16] implemented vision transformer to take advantage of the self-attention mechanism. Dynamically scalable vision transformer, DSViT was published by Wang et al. [17] to handle the limitations correlated with the global information extraction capabilities of single convolutional models and the computational overhead constrains by creating dynamically scalable attention model which integrates convolutional features with transformer features. Spatial-channel feature preserving ViT (SCViT) is developed by Lv et al. [18] which considers the contribution of distinct channels and considers geometric information in the classification token. This method generates tokens, introduces lightweight channel attention, models global interactions and uses a multilayer perceptron. DCNNs [19], [20] extract high level semantic features. However, these networks are mostly used for the single label remote sensing applications.

Multiple semantic labels are not being considered, nor are the dependencies between labels. This situation reflects a broader limitation of the prevailing single-label image classification approaches in remote sensing. Traditional convolutional networks and their numerous variants have achieved impressive accuracy when each image is assigned only one dominant land-cover category, yet they are intrinsically designed to predict a single class per scene. As a result, they cannot represent scenes containing multiple co-existing objects or land-cover types, and they ignore the semantic relationships among different classes. Even when such models are applied to complex high-resolution satellite imagery, they tend to force a single 'best' label, which oversimplifies heterogeneous landscapes such as urban areas where roads, vegetation, and water features appear together. This structural restriction of single-label methods motivates the development of models that can explicitly handle multiple labels and their interdependencies.

Hua et al. [21] utilized attention-based network to extract detailed semantic feature maps and used LSTM (Long-Short Term Memory) network to generate structured multiple object labels. CNN-RNN framework was proposed [22] regarding to multi-label image classification tasks. RNN is used following the CNN to capture a combined image-label representation and generated the label predictions. Contextual features can be extracted by vision transformer (ViT) [23], but it has high computational complexity and limited learning ability. Swin Transformer (SwinT) was published by Lie et al. [15] which can act as an adaptable framework for various tasks such as dense prediction and image classification. When compared to different attention-based transformers architectures, MaxViT has been observed as the best performer by Tu et al. [24]. It has outperformed models like Cross-ViT [25], DeepViT [26], DeiT [27], T2T [28] etc. with a top-1 accuracy of %85.2. It combines advantages of both enhanced CNNs and attention mechanisms within a novel "base-block". Base-block is composed of different blocks. "MBConv" block that incorporates SE (Squeeze-and-excitation) module. It is followed by a multi-axis attention block which is specifically crafted for the purpose of capturing local and global relationships between pixels.

The originators of Multi-label AID dataset, [29], created an attention-aware label relational reasoning network and achieved 88.72% CF1 score (per-category F1-Score) on their own dataset. Li et al. [30] utilized both visual and spatial information, combining CNN and GNN, and obtained 88.64% CF1 score. Tan et al. [31] proposed a network that contains two models: SSM (semantic sensitive module) and SRBM (semantic relation-building module). SSM captures the features using transformer to extract semantic attentional regions from visual attributes by DCNN (deep convolutional neural network), and SRBM uses the output of SSM to obtain the relation matrix for final classification, and this network achieved 89.97% CF1 score. Wu et al. [32] achieved 92.81% CF1 score by presenting a Transformer-based framework; SDM (Semantic Disentanglement Module), and MAT (Masked-Attention Transformer). Ma et al. [33] proposed LD-GCN (Label-Driven Graph Convolutional Network), inherent correlation of labels is learned by the label-correlation matrix and fed into LRGCN (Label Recognition Graph Convolutional Network) which harnesses the relationship between labels and images. This network not only achieved 92.81% CF1 score, but also 83.49% of mAP is obtained.

II. MATERIALS AND METHODS

A. Dataset

The original single-label AID dataset was developed by Xia et al. [9]. Google Earth imagery was the source of the high-resolution aerial images in the dataset with sizes of 3 x 600 x 600. A combined total of 10,000 images is grouped under 30 classes with spatial resolution ranging between 0.5 – 8m. The initial dataset was relabelled in 2020, and it has become a multi-label version of the original dataset with the same resolution of the images [29]. This new dataset, multi-label AID, consists of 3000 images in total with 17 labels. Every image receives manual annotation, assigning up to 11 labels. Fig. 1 illustrates several examples from the dataset with their associated labels.

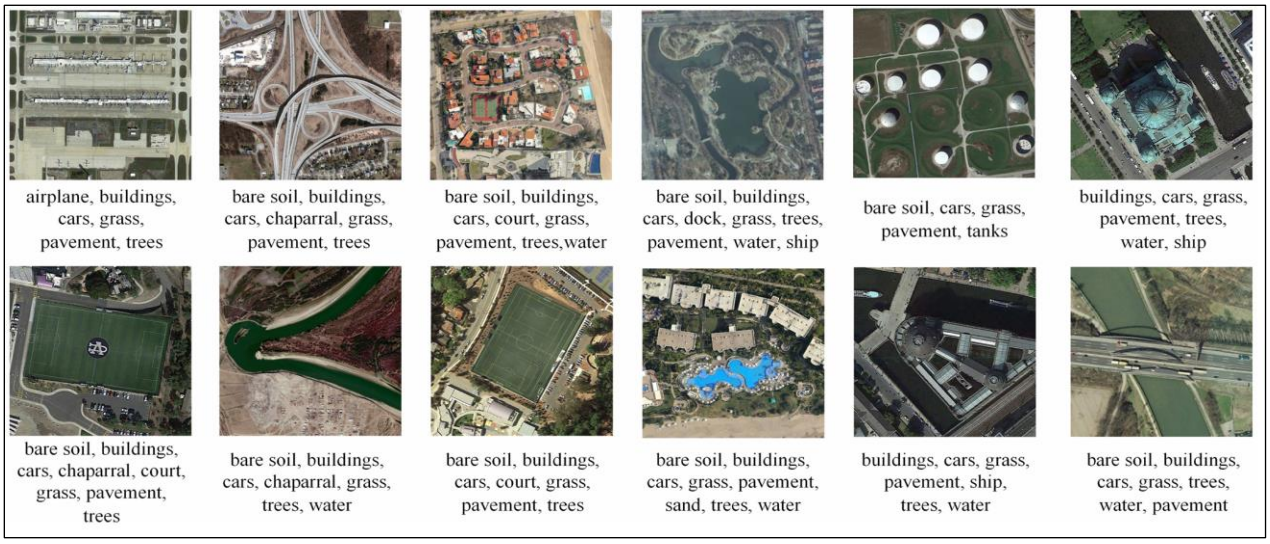

Fig. 1. Sample images from the Multi-label AID datase.

Table I shows the label distribution for training, validation and test sets.

TABLE I. LABEL DISTRIBUTION ACROSS TRAIN, VALIDATION AND TEST SETS

Label	Train	Validation	Test
airplane	62	17	20
bare-soil	934	237	304
buildings	1372	372	417
cars	1288	329	409
chaparral	56	19	37
court	212	57	75
dock	174	47	50
field	133	42	39
grass	1463	366	466
mobile-home	1	0	1
pavement	1488	382	458
sand	166	41	52
sea	143	34	44
ship	189	48	47
tanks	85	2	21
trees	1523	401	483
water	525	149	178

The Multi-label AID dataset exhibits a highly skewed label distribution, with some classes (e.g., mobile-home) represented by only a single image. We chose to keep this natural imbalance for two reasons: (i) the dataset is a widely used benchmark and altering its composition would compromise comparability with prior work, and (ii) generating synthetic samples for such rare classes risks unrealistic artifacts and would not materially improve model generalization. We therefore trained on the unaltered data and explicitly analyse the performance implications of this imbalance.

B. Data Augmentation

Because the number of images in the Multi-label AID dataset is limited, we applied data augmentation to mitigate overfitting and improve the model's ability to generalize. Each image was first resized to 256 x 256 pixels and then a random 224 x 224 patch was cropped for training. This strategy provides scale and translation invariance while matching the input size expected by our backbone network. Random horizontal and vertical flips were added to simulate viewpoint changes and increase orientation diversity, which are common in aerial imagery. We considered domain-specific spectral augmentations such as channel-wise intensity shifts, but refrained from applying them because the available images are RGB only and lack separate spectral bands, making such adjustments less meaningful. An example of the augmented is given in Fig. 2.

Fig. 2. Augmented versions of images from the Multi-label AID dataset

While we did not modify the dataset distribution, we applied random cropping and flipping to all images to increase the total training data and reduce overfitting. This augmentation improves generalization but does not change the relative class frequencies. Although random cropping and flipping helped improve generalization, the inherent class imbalance of the Multi-label AID dataset—particularly for rare categories such as mobile-home—remains a key challenge. Future work could investigate complementary strategies such as transfer learning from semantically related categories, class-balanced or focal re-sampling, and cost-sensitive loss weighting to provide additional support for underrepresented labels.

C. Methods

This study involves the implementation and optimization of CNN-based, vision transformer and hybrid networks. Transfer learning is employed using pre-trained weights for all the models. Models that are pre-trained consistently outperform those trained from scratch [34]. These weights were obtained from models pre-trained on ImageNet. Due to the scant quantity of the Multi-label AID images, it is more likely for the models to overfit. For the purpose of preventing overfitting, not only the data is augmented, but also patience parameter is selected as 5 to keep the models with the best validation accuracy under 20 epochs. In addition, one Dropout layer with 20% probability is added just before the very last layer, fully connected layer, for each model. OneCycleLR is used as scheduler which is provided by the torch.optim library. Maximum learning rate is set to $4 * 10^{-4}$. Adam [35] is used for the optimization where the learning rate is 0.0001 with a weight decay of 0.0001. These values were selected after a small pilot search informed by Tu et al. [24], which reports peak learning rates around $3\ x\ 10^{-3}$ and a weight decay rate of 0.05. Adapting these recommendations to the smaller Multi-label AID dataset, we found that lowering the learning rate and weight decay provided more stable training and the best validation mAP while remaining within the empirically validated range of the original architecture.

There are four cases which a classifying model can give as an output: true positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). These metrics can be used to assess a model's precision, recall and F1-score values. When a model gives positive outputs, the precision shows its reliability. When a model correctly classifies positive data points, that is the recall. Harmonic mean of both recall and precision is equal to F1-Score. Various evaluation metrics are used for the multi-label image classification as determination of this article's success. The classification report tool is used from sklearn.metrics library. The evaluation metrics are calculated in the Equations below.

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$F1-Score = \frac{2*Precision*Recall}{Precision+Recall} \tag{3}$$

Let $B(TP_j,\ FP_j,\ TN_j,\ FN_j)$ be a specific binary classification metric, where $B \in \{Precision, Recall, F1-Score\}$[36].

$$B_{macro} = \frac{1}{q}\sum_{j=1}^{q} B\left(TP_j, FP_j, TN_j, FN_j\right) \tag{4}$$

$$B_{micro} = B\left(\sum_{j=1}^{q} TP_j, \sum_{j=1}^{q} FP_j, \sum_{j=1}^{q} TN_j, \sum_{j=1}^{q} FN_j\right) \tag{5}$$

$$B_{weighted} = B(\sum_{j=1}^{q} TP_j.\frac{support_j}{p}, \sum_{j=1}^{q} FP_j.\frac{support_j}{p}, \sum_{j=1}^{q} TN_j.\frac{support_j}{p}, \sum_{j=1}^{q} FN_j.\frac{support_j}{p}) \tag{6}$$

Mean Average Precision (mAP) measures the overall performance across all classes. In other terms, it is the mean of the average precisions (AP) of all classes where AP is the area under the PR (precision-recall) curve. mAP can be calculated as shown in (7) where L denotes the number of classes.

$$mAP = \frac{1}{L}\sum_{i=1}^{L} P(i).R(i) \tag{7}$$

PR curve is used in this study due to the reason that ROC-curve's estimation might be inadequate as long as the positive class is substantially smaller [37]. Sigmoid function is used as an activation function. The reason for it to be used is not limited with its output values which are in the range [0, 1], but also it provides smooth gradients that helps with backpropagation during training phase. This ensures that the learning algorithm can effectively update the weights. Sigmoid outputs allow high probabilities for all labels. The sigmoid function is calculated as in (8).

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{8}$$

Two functions are applied as loss functions to the different networks, Binary Cross-Entropy With Logits Loss (BCEWithLogitsLoss) [38] and Asymmetric Loss (ASL)[39].

A given problem with multiple labels is divided into distinct binary problems for each label by BCEWithLogitsLoss which makes it possible to assign multiple labels to each item. The sigmoid function produces separate real-valued outputs for every input. Thus, the actual output is estimated. Sigmoid allows high probabilities for all classes, unlike Softmax which assigns high probability to high value. The logits represent the raw outputs before Sigmoid is applied. In multi-label image classification each class must be predicted independently, so the activation function must allow multiple outputs to be simultaneously high. Softmax enforces a probability distribution that sums to one and therefore assumes mutual exclusivity among classes, making it unsuitable when an image may belong to several categories at once. In contrast, the sigmoid function produces an independent probability for each class in the range [0, 1], enabling the network to assign high confidence to any subset of labels. Coupled with BCEWithLogitsLoss, this treats each label as a separate binary classification problem and provides smooth gradients for stable back-propagation. Formula for BCEWithLogitsLoss is given in (9) where p_i' is fully connected layer's logits (Sigmoid function is applied) and p_i is the true label.

$$BCEWithLogitsLoss = -\frac{1}{n}\sum_{i=1}^{n} p_i \log(p_i') + (1-\ p_i)\log(1-(p_i'))] \tag{9}$$

ASL is one of the variations of binary cross-entropy loss. They are generally combined by a sigmoid function (8) for the purpose of converting the model outputs to probabilities. ASL consists of two complementary asymmetric mechanisms, that operate in distinct ways on positive and negative samples,

enabling direct control. ASL enables for the selective increase in weight of minority negatives and meanwhile maintains the original weighting for common positives, biasing frequent classes is dismissed by this. Re-balancing is also enabled by ASL, unlike batch-dependent schemes such as distribution-balanced loss. The ASL formula is expressed as in (10). If γ^+ and γ^- are both set to be 0 in (10), BCEWithLogitsLoss can be calculated as well (see (9)).

$$ASL = -\sum_{n=1}^{N}(p_i(1-p_i')^{\gamma^+}\log(p_i') + (1-p_i)(p_i'^{\gamma^-})(\log(1-p_i'))) \quad (10)$$

MaxViT is a hybrid vision transformer that combines the local feature-extraction ability of convolutions with the long-range dependency modeling of attention. Each MaxViT block contains two main components: an MBConv (Mobile Inverted Bottleneck) layer, which expands and contracts the channel dimension to capture rich local representations with fewer parameters, and a Multi-Axis Attention module that first applies Block Attention to model short-range spatial relationships and then uses Grid Attention to capture global context across the entire image. This sequence of local and global attention provides strong contextual modeling while remaining computationally efficient. A standard MaxViT network consists of an initial stem convolution, a stack of MaxViT blocks, global average pooling, and a final fully connected layer for classification (Fig. 3).

For the experiments: Pytorch is used as the primary deep learning framework on a personal computer with NVIDIA Geforce RTX 3070 Ti graphic card with 16 GB of memory.

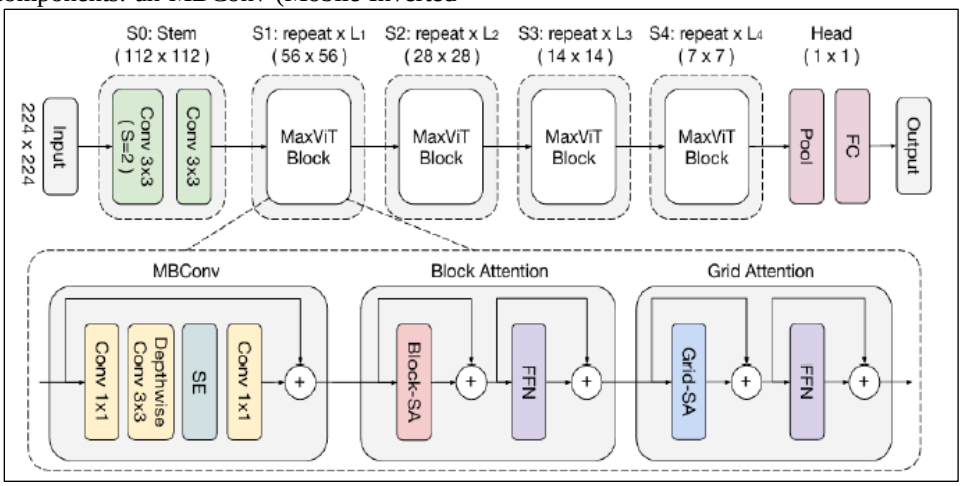

Fig. 3. MaxViT Architecture [24]

III. Results And Discussion

MaxViT is a hybrid model that initially consists of CNN and transformer network, combining global and local features with a simple and scalable design while maintaining computational efficiency. The comparison of MaxViT with the other well-known models is presented in Table II.

TABLE II. BENCHMARKING MAXVIT AGAINST DIFFERENT MODELS

Model	Type	Key Features	MaxViT Comparison
AlexNet	CNN	Simple architecture, foundational deep learning model	MaxViT provides advanced features, better scalability
VGG16	CNN	Deep architecture with 16 layers	MaxViT offers better efficiency, global-local interactions
DenseNet-201	CNN	Dense connections, efficient parameter usage	MaxViT is more scalable, efficient with global interactions
InceptionV3	CNN	Factorized convolutions, multi-scale processing	MaxViT simplifies design, maintains high performance
ConvNeXt	Modern CNN	Transformer-inspired designs, competitive performance	MaxViT integrates convolution and attention mechanisms
ViT	Transformer	Image patches as sequences, transformer-based	MaxViT combines convolution and attention for scalability
SwinT	Transformer	Shifted window attention, hierarchical architecture	MaxViT offers linear complexity global-local interactions

The MaxViT-T model demonstrates a compelling advantage by combining CNN and transformers with 31M parameters and 5.6G FLOPs. MaxViT-T achieves 83.6% accuracy on the ImageNet-1K dataset, surpassing AlexNet and VGG16 with 60M and 138M parameters as well as 63.3% and 71.5% accuracy respectively. Even more advanced models like DenseNet-201 and Inceptionv3 do not match MaxViT's performance with accuracies of 77.2% and 77.9%. Transformers such as ViT and SwinT require careful optimization and longer training to reach competitive accuracy.

According to the original ViT paper, ViT trained directly on ImageNet-1K achieves about 77.9 % top-1 accuracy with roughly 86 M parameters and about 55 G FLOPs, while SwinT reports about 28 M parameters and 4.5 G FLOPs with 81.3 % accuracy. Thus, MaxViT-T delivers accuracy that is competitive with or better than these transformer baselines while keeping computational requirements close to the most efficient SwinT configuration. MaxViT-L achieves superior results with 85.2% accuracy with 212M parameters and 43.9G FLOPs [24]. Due to the computational complexity of MaxViT-L, MaxViT-T is chosen as a main network in this study. Technically, MaxViT-T has a far smaller footprint—31 M parameters and 5.6G FLOPs versus 212 M and 43.9G FLOPs for MaxViT-L—keeping per-GPU memory below 8 GB for 224×224 inputs and enabling a batch size of 128 without gradient checkpointing. This

configuration trains with mixed-precision (FP16) in roughly half the wall-clock time of MaxViT-L and consumes about 40 % less energy, while the larger model offers only a ~1.6 percentage-point gain in top-1 ImageNet accuracy. It is claimed that symmetric loss functions such as focal loss or cross entropy loss is sub-optimal for learning positive samples' features which can be clearly seen on the evaluation metrics of the labels that support results [39].

Drawing inspiration from this study, ASL is applied as a loss function for the MaxViT model. Adjusting γ^+and γ^- (see (10)) helps giving appropriate importance to the minority class, meanwhile the model does not become biased towards the majority class. Therefore, different values are applied on MaxViT model where mAP is determined as the success criteria. Following the search range proposed by Gao et al. [40], we applied the same γ^+ and γ^- intervals to our dataset. Because the datasets differ, our validation experiments identified a different optimum - γ^+ = 1 and γ^- = 2 – which produced the highest mAP. The results can be seen in Table III.

TABLE III. IMPACT OF DIFFERENT Γ COMBINATIONS OF MAP (%) ON MAXVIT

γ- \ γ+	**0**	**1**	**2**	**3**	**4**
2	84.2	84.98	83.82	83.46	84.41
4	83.58	84.03	83.83	82.95	83.92

In deep learning applications, a single hyperparameter does not always play the main role. Scheduler is also one of the factors that plays a vital role in the success of the model for classifying. OneCycleLR allows the learning rate to rise during training by following a cyclical pattern where the learning rate increases to a maximum value and decreases. The performance of OneCycleLR on the same models with the aforementioned two loss functions is compared. Comparison showed that BCEWithLogitsLoss is slightly better than ASL in CNN models except Inceptionv3. It aligns with binary classification objectives and is robust in handling data imbalances by providing stable probability outputs. These factors together lead to improved precision and recall, and results in higher mAP scores.

Conversely, ASL performs better than BCEWithLogitsLoss for the vision transformer models, except a small difference in ViT. ASL handles the class imbalances more effectively by weighting positive and negative examples differently that deals with imbalanced datasets, such as Multi-label AID, where one class is underrepresented. For example, airplane and chaparral have less instances when compared to the other labels, 20 and 37 respectively. This contrast likely arises from architectural differences: CNNs benefit from the independent, stable gradients of BCEWithLogitsLoss, which match their localized feature learning and reduce overfitting, whereas transformers' global attention amplifies negative-sample dominance. ASL's asymmetric focusing (γ^+, γ^-) counteracts this effect, enabling transformers to better learn rare labels despite strong inter-class correlations. The success of combining ASL and OneCycleLR on vision transformer models is undeniable, as they achieved the highest F1-Score in 16 out of a total of 17 classes where F1-Score is more comprehensive and critical, considering recall and precision metrics. The comparison of MaxViT model with two loss functions, ASL weights the positive and negative samples much better than BCEWithLogitsLoss.

Table IV highlights the overall performance landscape of all evaluated networks, showing clear differences in how each architecture handles the diverse set of object categories. Models with transformer components generally maintain stronger and more balanced accuracy across classes, while purely convolutional approaches display wider variation between frequent and infrequent labels.

ANOVA was applied to the results presented in Table IV. Since the p-value was found to be less than 0.05, the null hypothesis (H_0) was rejected for all cases, indicating that at least one group mean significantly differed from the others. Subsequently, pairwise comparisons of the algorithm performances were conducted using the Student's t-test to assess whether the observed differences were statistically significant. Examining the results presented in Table V, it appears that the proposed MaxViT method performed better.

MaxViT performs both grid and global attention mechanisms that allows the model to capture multi-scale features more effectively than a simple fusion of CNN and transformer models. As far as is known, mAP of 84.98% has been achieved for the first time on Multi-label AID dataset. A comparison of performance metrics between earlier research and MaxViT is given in Table VI.

TABLE IV. AP VALUES OF EACH CLASS WITH THE MODELS

Object Labels	AlexNet	VGG16	DenseNet-201	Inceptionv3	ConvNeXt	ViT	SwinT	MaxViT
airplane	0,817	0,372	0,769	0,915	0,731	0,673	0,787	1,000
bare-soil	0,805	0,796	0,853	0,828	0,838	0,859	0,815	0,883
buildings	0,975	0,985	0,984	0,990	0,994	0,992	0,991	0,995
cars	0,975	0,964	0,979	0,970	0,982	0,979	0,983	0,988
chaparral	0,267	0,354	0,476	0,406	0,375	0,367	0,381	0,481
court	0,539	0,663	0,627	0,692	0,676	0,665	0,768	0,812
dock	0,625	0,683	0,695	0,805	0,773	0,807	0,814	0,819
field	0,609	0,631	0,753	0,749	0,768	0,782	0,686	0,804
grass	0,960	0,976	0,980	0,981	0,985	0,985	0,979	0,981
mobile-home	0,002	0,002	0,006	0,062	0,004	0,002	0,091	0,004
pavement	0,986	0,992	0,992	0,990	0,993	0,997	0,993	0,995
sand	0,928	0,882	0,957	0,969	0,958	0,956	0,929	0,961
sea	0,915	0,926	0,956	0,965	0,975	0,960	0,945	0,979
ship	0,549	0,671	0,761	0,843	0,720	0,725	0,737	0,881
tanks	0,940	0,894	0,946	0,946	0,949	0,973	0,909	1,000
trees	0,961	0,985	0,981	0,990	0,988	0,973	0,980	0,987
water	0,715	0,823	0,826	0,811	0,788	0,818	0,802	0,875
mAP	0,739	0,741	0,797	0,818	0,794	0,795	0,799	0,849

TABLE V. STATISTICAL COMPARISON OF THE PROPOSED MAXVIT METHOD WITH THE OTHERS.

Compared Methods	P(T<=t) Single ended	P(T<=t) two ended
MaxViT- Alexnet	1,46307 E-16	1,78094E-16
MaxViT- VGG16	8,103714E-13	1,61385E-12
MaxViT- DenseNet-201	1,35845E-06	2,7169E-06
MaxViT- Inceptionv3	0,0000327510	0,000075422
MaxViT- ConvNeXt	1,379103E-07	2,756E-07
MaxViT- ViT	5,91E-07	1,18E-06
MaxViT- SwinT	3,30504E-06	6,0829E-06

TABLE VI. A COMPARISON OF PERFORMANCE METRICS BETWEEN EARLIER RESEARCH AND MAXVIT, UTILIZING THE MULTI-LABEL AID DATASET

Method	Score (CF1/CP/CR)	mAP (%)
AL-RN-ResNet50 [29]	88.72 / 91.00 / 88.95	-
MLRSSC-CNN-GNN [30]	88.64 / 89.83 / 90.20	-
ResNet50-SR-Net [31]	89.97 / 89.42 / 90.52	-
S-MAT-ResNet50 [32]	90.90 / 92.17 / 89.69	-
LD-GCN [33]	90.93 / 92.81 / 89.06	83.49
MaxViT	91 / 92 / 91	84.98

The success of MaxViT can be explained by its architecture since both convolutional operations and multi-head self-attention are leveraged. This dual capability enables both local and global features to be captured effectively and particularly beneficial for multi-label classification tasks where different labels might correspond to distinct features at various scales and regions within an image. Spatial relationships and local patterns within the images of Multi-label AID are captured by MaxViT which facilitates hierarchical feature extraction. The features are extracted at different levels of abstraction, from edges and textures in the initial layers to more complex shapes and objects in the deeper layers. Extracted initial feature maps, which contains localized and spatially aware representations of the image, serve as input for the subsequent transformer blocks. These input feature maps processed by transformer blocks using self-attention and FFN.

Global features are captured by the self-attention mechanism in each MaxViT block, with the importance of each feature being computed relative to all others in the feature map. FFN processes the combined local and global features, adds non-linearity. As a result, the extraction of local features through convolutional layers and global features via the multi-head self-attention mechanism creates a powerful and flexible architecture, enhancing the model's capability to identify and distinguish multiple labels within a single image. In addition, the integration of ASL and OneCycleLR not only improved mAP value but also showed robustness in handling the complexities of multi-label image classification. PR curve for all the labels can be seen in Fig. 4, using MaxViT where γ^+ and γ^- are set as 1 and 2, respectively.

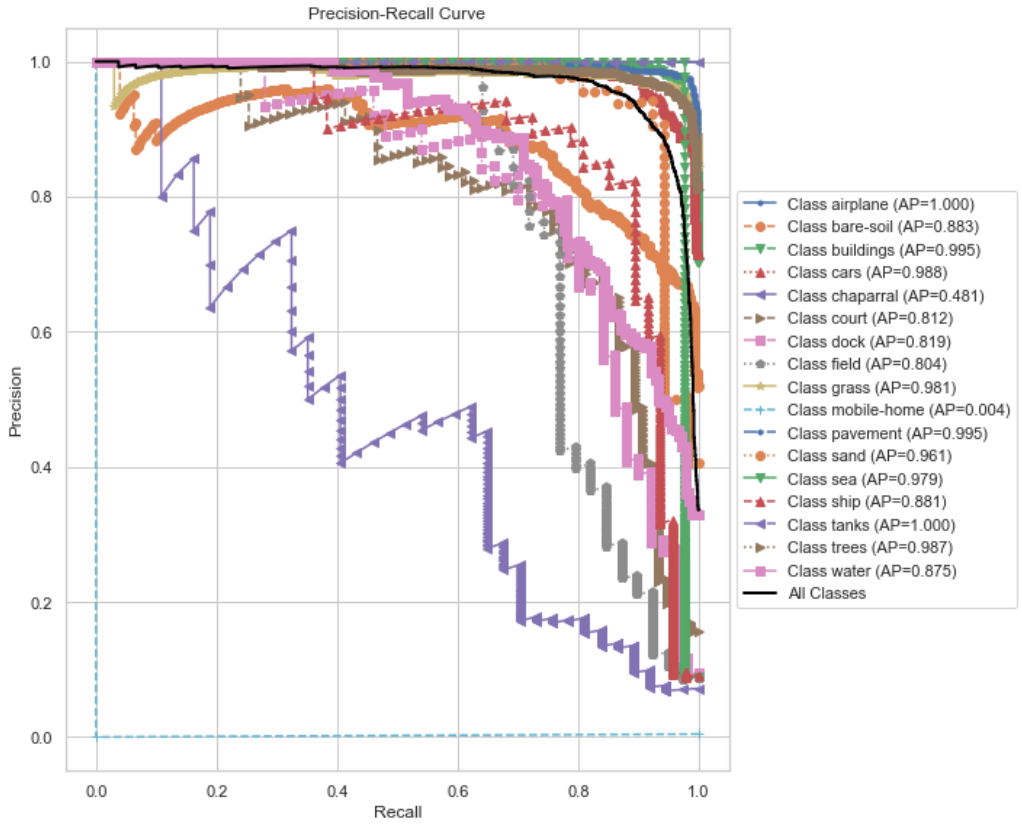

Fig. 4. PR curve of the labels for MaxViT where γ^{+}= 1, γ^{-}=2, mAP = 84.98%

IV. Conclusion

The multi-label AID dataset has some shortcomings. Table I clearly reveals a severe class imbalance across the Multi-label AID dataset. While common land-cover types such as trees, pavement and grass dominate the distribution, several categories are extremely under-represented. Notably, mobile-home appears only once in the entire dataset (1 training image and 1 test image, no validation samples), airplane has only 62 training and 20 test images, and chaparral has just 56 training and 37 test images. Other classes such as tanks, court, dock and field also have far fewer examples than the dominant categories. Lower precision and recall are inevitable for rare labels because categories with only a handful of images provide insufficient variability for the networks to learn discriminative features.

In our experiments, labels such as chaparral and airplane consistently exhibit low recall, while mobile-home cannot be classified at all by any model. This extreme scarcity also imposes a ceiling on the mAP, since mobile-home label contributes only one positive instance, the theoretical upper bound of the dataset-level mAP is 94.12%, even if all other labels were predicted perfectly. In addition, the large imbalance biases the training process toward majority classes, abundant categories such as trees, pavement and grass dominate the loss and encourage the models to favor these labels, which can mask poor performance on minority classes when only overall accuracy is reported. Although we did not perform an exhaustive comparison with all learning-rate schedulers and loss functions, prior studies [39] report that symmetric losses such as focal loss or cross-entropy are sub-optimal for learning discriminative features of positive samples in highly imbalanced multi-label settings.

Drawing on this evidence, we adopted Asymmetric Loss (ASL) for the MaxViT model, as it is specifically designed to handle class imbalance through asymmetric treatment of positive and negative samples. Coupling ASL with the OneCycleLR scheduler, which provides dynamic learning-rate adjustments and has been shown to accelerate convergence and improve generalization, improved overall model accuracy. OneCycleLR proved especially effective for the MaxViT-T architecture on the MultiLabel-AID dataset because its dynamic, non-monotonic schedule matches both the model's depth and the dataset's class imbalance. MaxViT's hybrid CNN–transformer blocks require an initial period of rapid feature exploration to stabilize attention weights, followed by a slower refinement stage.

The single–cycle policy—starting with a low learning rate to avoid divergence, rising to a high peak to escape sharp local minima, and then annealing—encourages wide-basin convergence that improves generalization. MultiLabel-AID further benefits from this approach: its heterogeneous aerial scenes contain overlapping labels and rare classes, so a temporary high learning rate early in training helps the optimizer traverse saddle points and learn minority-class features, while the final decay reduces overfitting to dominant categories. Among the different configurations we tested, this combination achieved the best validation performance on the MaxViT-T model, supporting our choice while leaving a broader comparison with other schedulers and loss functions as future work. The findings demonstrate that the choice of loss function and learning rate scheduler in an appropriate network model impacts the performance of the model significantly.

In the next studies, classification performance can be increased by focusing on achieving high classification rates on small instances and finding the best combination of the loss function, scheduler and window-based vision transformer. Due to the limited computational resources, the MaxViT-L algorithm could not be implemented in this study.

In future applications, this algorithm can be adopted for better classification results. Training MaxViT-L on the available workstation was infeasible because the model's 212 M parameters and roughly 44 GFLOPs per 224×224 image exceeded the GPU's memory budget during forward and backward passes. A practical path to overcome this limitation would be to distribute the model across multiple GPUs using model or pipeline parallelism, which allows parameters and activations to be split. Alternatively, future experiments could migrate to cloud or high-performance computing resources equipped with GPUs offering far larger memory capacities—such as NVIDIA A100 or H100 cards with 40 GB or more—so that MaxViT-L can be trained end-to-end without architectural changes.

References

[1] D. K. Sinha, P. P. Chakraborty, A. Rahman, A. Kumar Saha, and A. R. Siddiqui, "Remote Sensing and GIS Module: Basic Physics of Remote Sensing."

[2] G. Koukiou, "SAR Features and Techniques for Urban Planning—A Review," Jun. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/rs16111923.

[3] J. S. Estrada, A. Fuentes, P. Reszka, and F. Auat Cheein, "Machine learning assisted remote forestry health assessment: a comprehensive state of the art review," 2023, *Frontiers Media S.A.* doi: 10.3389/fpls.2023.1139232.

[4] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sens Environ*, vol. 202, pp. 18–27, Dec. 2017, doi: 10.1016/j.rse.2017.06.031.

[5] D. V. Malakhov and O. V. Dolbnya, "Remote sensing as a tool of biological conservation and grassland monitoring in mountain areas of Southeastern Kazakhstan," Journal of Applied Science and Technology Trends, vol. 4, no. 1, pp. 72–79, Jan. 2023, doi: 10.38094/jastt401175.

[6] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," Aug. 2017, [Online]. Available: http://arxiv.org/abs/1709.00029

[7] N. Ali and B. Zafar, "RSSCN7 Image dataset," Sep. 2018, *figshare*. doi: 10.6084/m9.figshare.7006946.v1.

[8] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010, pp. 270–279.

[9] G. S. Xia *et al*., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017, doi: 10.1109/TGRS.2017.2685945.

[10] O. Russakovsky *et al*., "ImageNet Large Scale Visual Recognition Challenge," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.0575

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks." [Online]. Available: http://code.google.com/p/cuda-convnet/

[12] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with Noisy Student improves ImageNet classification," Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.04252

[13] H. Zhang *et al.*, "ResNeSt: Split-Attention Networks," Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.08955

[14] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[15] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.14030

[16] M. Kaselimi, A. Voulodimos, I. Daskalopoulos, N. Doulamis, and A. Doulamis, "A Vision Transformer Model for Convolution-Free Multilabel Classification of Satellite Imagery in Deforestation Monitoring," *IEEE Trans Neural Netw Learn Syst*, vol. 34, no. 7, pp. 3299–3307, Jul. 2023, doi: 10.1109/TNNLS.2022.3144791.

[17] F. Wang, J. Ji, and Y. Wang, "DSViT: Dynamically Scalable Vision Transformer for Remote Sensing Image Segmentation and Classification," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 16, pp. 5441–5452, 2023, doi: 10.1109/JSTARS.2023.3285259.

[18] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "SCViT: A Spatial-Channel Feature Preserving Vision Transformer for Remote Sensing Image Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022, doi: 10.1109/TGRS.2022.3157671.

[19] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[20] C. Szegedy *et al.*, "Going Deeper with Convolutions," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.4842

[21] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently Exploring Class-wise Attention in A Hybrid Convolutional and Bidirectional LSTM Network for Multi-label Aerial Image Classification," Jul. 2018, doi: 10.1016/j.isprsjprs.2019.01.015.

[22] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A Unified Framework for Multi-label Image Classification."

[23] K. Xu, P. Deng, and H. Huang, "Vision Transformer: An Excellent Teacher for Guiding Small Networks in Remote Sensing Image Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022, doi: 10.1109/TGRS.2022.3152566.

[24] Z. Tu *et al.*, "MaxViT: Multi-Axis Vision Transformer," Apr. 2022, [Online]. Available: http://arxiv.org/abs/2204.01697

[25] C.-F. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.14899

[26] D. Zhou *et al.*, "DeepViT: Towards Deeper Vision Transformer," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.11886

[27] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," Dec. 2020, [Online]. Available: http://arxiv.org/abs/2012.12877

[28] L. Yuan *et al.*, "Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet," Jan. 2021, [Online]. Available: http://arxiv.org/abs/2101.11986

[29] Y. Hua, L. Mou, and X. X. Zhu, "Relation Network for Multi-label Aerial Image Classification," Jul. 2019, doi: 10.1109/TGRS.2019.2963364.

[30] Y. Li, R. Chen, Y. Zhang, M. Zhang, and L. Chen, "Multi-label remote sensing image scene classification by combining a convolutional neural network and a graph neural network," *Remote Sens (Basel)*, vol. 12, no. 23, pp. 1–17, Dec. 2020, doi: 10.3390/rs12234003.

[31] X. Tan, Z. Xiao, J. Zhu, Q. Wan, K. Wang, and D. Li, "Transformer-Driven Semantic Relation Inference for Multilabel Classification of High-Resolution Remote Sensing Images," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 15, pp. 1884–1901, 2022, doi: 10.1109/JSTARS.2022.3145042.

[32] H. Wu, C. Xu, and H. Liu, "S-MAT: Semantic-Driven Masked Attention Transformer for Multi-Label Aerial Image Classification," *Sensors*, vol. 22, no. 14, Jul. 2022, doi: 10.3390/s22145433.

[33] B. Ma *et al.*, "Label-Driven Graph Convolutional Network for Multilabel Remote Sensing Image Classification," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 17, pp. 2245–2255, 2024, doi: 10.1109/JSTARS.2023.3344106.

[34] A. Rangel, J. Terven, D. M. Cordova-Esparza, and E. A. Chavez-Urbiola, "Land Cover Image Classification," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2401.09607

[35] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Dec. 2014, [Online]. Available: http://arxiv.org/abs/1412.6980

[36] X. Wang, L. Duan, and C. Ning, "Global context-based multilevel feature fusion networks for multilabel remote sensing image scene classification," *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 14, pp. 11179–11196, 2021, doi: 10.1109/JSTARS.2021.3122464.

[37] I. Puleko, O. Svintsytska, V. Chumakevych, V. Ptashnyk, and Y. Polishchuk, "The Scalar Metric of Classification Algorithm Choice in Machine Learning Problems Based on the Scheme of Nonlinear Compromises," 2022.

[38] The PyTorch Foundation, "BCEWITHLOGITSLOSS," 2022.

[39] T. Ridnik *et al.*, "Asymmetric Loss For Multi-Label Classification." [Online]. Available: https://github.com/Alibaba-MIIL/ASL.

[40] Q. Gao, T. Long, and Z. Zhou, "Mineral identification based on natural feature-oriented image processing and multi-label image classification," Expert Syst Appl, vol. 238, Mar. 2024, doi: 10.1016/j.eswa.2023.122111.